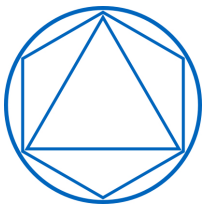


Technische Universität München

Department of Mathematics

# Statistical analysis of energy appliance data

Master's Thesis

by

Susan Kollosche

| | |
|---|---|
| Supervisor: | Prof. Ph.D. Claudia Czado |
| Advisor: | Prof. Ph.D. Claudia Czado |
| | Alexander Kreuzer |
| Submission date: | March 05, 2020 |

I hereby declare that this thesis is my own work and that no other sources have been used except those clearly indicated and referenced.

Munich, March 05, 2020

# Acknowledgments

First of all, I would like to express my sincere gratitude to Prof. Claudia Czado for giving me the opportunity to work on this topic under her supervision. I would like to thank her for all the guidance and the helpful explanations and hints in the meetings that enriched the progress of my work.

I would also like to thank my advisor Alexander Kreuzer for his support and for all the useful advice during the last months.

The discussions with Prof. Claudia Czado and Alexander Kreuzer have helped me to learn a lot and to evolve for future projects, for which I am very grateful.

Last but not least, I want to express my deep gratitude to my family and friends for all their unconditional support and love, providing an indispensable source of energy and strength.

# Abstract

In this master thesis we develop statistical models for energy use of appliances. The factors influencing energy consumption becomes more important every day and making good predictions for an intelligent energy distribution is increasingly being discussed.

This thesis presents four final statistical models to quantify the effect of temperature, humidity and very deciding time effect covariates on the appliances energy consumption. To create models for predictions, the multiple linear regression (LM) and the generalized additive model (GAM) are used. These models are improved and adapted by allowing for linear and non-linear structure on a logarithmic transformed response variable the appliances energy use. Additionally, we consider not only the main effect, but also interactions with a time effect, namely the hour effect. The evaluation and comparison of model fits are based on the adjusted coefficient of determination ($R^2_{adj}$) and the Akaike's information criterium (AIC). Better results are seen for GAM's, but due to the integration of non-linear components and interactions, we also reach satisfying results for the LM's.

# Zusammenfassung

Im Rahmen dieser Masterarbeit entwickeln wir statistische Modelle für die Energienutzung von Geräten. Die Einflussfaktoren auf den Energieverbrauch werden von Tag zu Tag wichtiger und das Treffen von gute Vorhersagen für eine intelligente Energieverteilung werden zunehmend diskutiert.

Diese Arbeit präsentiert vier finale statistische Modelle zur Quantifizierung des Einflusses von Temperatur, Luftfeuchtigkeit und sehr entscheidende Zeiteffekt-Kovariaten auf den Energieverbrauch der Geräte. Um Modelle für die Vorhersagen zu erstellen, werden die multiple lineare Regression (LM) und das verallgemeinerte additive Modell (GAM) verwendet. Diese Modelle werden verbessert und angepasst, indem sie eine lineare und nichtlineare Struktur auf einer logarithmisch transformierten Zielgröße, dem Energieverbrauch der Geräte, ermöglichen. Zusätzlich berücksichtigen wir nicht nur den Haupteffekt, sondern auch Interaktionen mit einem Zeiteffekt, und zwar dem Effekt der Stunden. Die Bewertung und der Vergleich der Modellanpassungen basieren auf dem adjustierten Bestimmtheitsmaß ($R^2_{adj}$) und dem Akaike Informationskriterium (AIC). Bessere Ergebnisse werden für GAM's erzielt, aber aufgrund der Integration von nichtlinearen Komponenten und Interaktionen erreichen wir auch zufriedenstellende Ergebnisse für die LM's.

# Contents

# Chapter 1

# Introduction

The advantage of statistical methods is the ability to create a mathematical equation that reflects the complexity of the relation between a response variable and a set of explanatory variables or covariates. In addition, reasonable arguments for using statistical methods include the availability of simple tools to interpret the results, for checking the significant predictors, for assessing their relative importance and for graphical representation.

In the last decades, there is an increasing interest for using non-parametric techniques, e.g. generalized additive models (GAM), in fields such as ecology, finance or medicine. However, most researchers are still loyal to the application of parametric techniques, e.g. the linear models (LM), due to their robustness and lower computational cost. Since these two methods, LM and GAM, are very popular in the application, which can be explained by their ability of catching real existing dynamics, we use these methods to explain thermodynamic systems and processes, that is describing the energy use with help of temperature and humidity. In the paper of Candanedo et al. (2017), four statistical models have been used to explain the dynamic. They found that kitchen, laundry, living room ranked as the most important predictors for the energy use, also that the atmospheric pressure may be relevant. The worst model was the multiple linear regression and the best model the gradient boosting machine, when focusing on given weather data by their prediction importance. The method GAM was not applied. In particular, the importance of time effects or the inclusion of interaction terms was not further emphasized. This thesis will consider these points to develop a satisfying linear model (LM) with a data set, containing recordings like appliances energy consumption, temperature, humidity and other climate factors. In summary, this work examines the modeling of the impact of climate factors in and outside a house on appliances energy consumption with additional time specification, mixing parametric and non-parametric approaches and distinguish between the main and interaction effect in the model.

Appliances consume a high proportion of the whole electricity demand, i.e. $20 - 30\%$ (c.f. Kavousian et al. (2015) and Cetin et al. (2014)). Even in standby mode the electricity consumption is existent. That is the reason for the high interest on that subject and study works dealing with the question on what impacts the energy use. Because of the electricity demand, different regression with different variables have been made to predict electrical energy consumption in buildings to improve the energy distribution. Applications such as

model predictive control on where the loads are needed (c.f. Candanedo et al. (2013)) or electricity demand behavior on large residential appliances (c.f. Cetin (2016)) have been investigated. Another approach is presented in a work from Fumo et al. (2010), where the application of a series of predetermined coefficients from electrical and fuel utility bills estimate hourly energy consumption. There has been considerable amount of research devoted to this topic.

Of course, the number and type of electrical devices and the use of these appliances by occupants are big factors when analyzing the electricity consumption in buildings, since the level of device use varies from area to area, as discussed in Firth et al. (2008). For interpretation and prediction purposes, the questions might be, which of the appliances are having the biggest positive effects on electricity consumption and their locations, so that the pattern of the occupants and their location during the day is detected. Also the usage of these devices by the occupants obviously leave its marks that can be measured in vicinity or area of these devices. The measurements are for example temperature, humidity, light, noise and much more. Studies show an impact of thermal conditions on electrical appliances in thermally well-insulated buildings. This was shown by applying dynamic thermal model of electrical appliances, see Ruellan et al. (2016). Instead of predicting the energy loads, we also deal with modeling the aggregate appliances energy use.

This thesis is organized as follows: Chapter 2 gives an overview on relevant statistical distribution and exploration for finding suitable models. In Chapter 3 and Chapter 4 a theoretical foundation of statistical models, these are multiple linear regression and generalized additive models, and the way to use them is discussed with some explanatory examples. After the theory, the thesis continues with a description of the observed house and the data set, in Chapter 5. In the next step, Chapter 6, we investigate a variety of exploration tools to identify relevant and significant factors that influence the energy consumption. First, we look at the marginal exploration including the time series and perform a logarithmic transformation on the response variable, followed by the pairwise investigation and then we evaluate the pattern over time. Chapter 7 deals with the application of multiple linear regression models (LM) to the energy data, where we also introduce the time effect variable, i.e. weekday and hours. The appliances obviously depend on hour in a non-linear way. To get the form of dependence on the hour, we use a functional form of the hour. We also conducting a hypothesis test on the interaction time effects, i.e. hour and weekday, to decide on the non-linear polynomial degree of hour. The resulting polynomial degree of hour is then entered as a covariate in the linear model. We set main effect models followed by the interaction models. To verify the explained variation, we provide the adjusted coefficient of determination ($R^2_{adj}$) and significance of covariates for each model and decide which model and covariate should be used for further model fits. After finding the final linear models with help of $R^2_{adj}$ and Akaike's information criterium (AIC), a prediction is presented. In Chapter 8 we create the flexible models where the hour is modeled non-parametrically together with the temperature and humidity covariates. Again, we first consider all covariates for the main effect GAM and then the interaction effect GAM using the tensor product. A first comparison of the four final models, the main and interaction effect GAM and the main and interaction effect LM,

is made. Chapter 9 gives a final model comparison based on $R^2_{adj}$ and AIC, which is the focus of this thesis. Furthermore, possible promising comparison methods are discussed. Chapter 10 concludes.

The model fits and figures were created using the program *R*. The book of Crawley (2012) was used to help perform the computations and visualizations.

# Chapter 2

# Univariate and multivariate distributions and their exploration

## 2.1 Univariate and multivariate distributions

In the following Table 2.1 we will present all statistical distributions we use in this thesis.

To better understand the transformations and applications of statements in the theory part of studentized residuals and F-tests, we will provide the full definitions of the t- and F-Distribution here.

**Definition 2.1** *(t-Distribution) A continuous variable $X$ has a t-distribution with $n$ degrees of freedom if it has p.d.f.*

$$f(x) = \frac{\gamma(n+1)/2}{\sqrt{n\pi}(n/2)(1+x^2/n)^{(n+2)/2}}.$$

*The mean and variance are given by*

$$E(X) = 0, \;\; n > 1, \;\; Var(X) = n/(n-2), \;\; n > 2$$

*We write $x \sim t_n$. The $t_1$-distribution is also called Cauchy distribution.*

*If $X \sim N(0,1)$ and $Y \sim \chi^2_n$ are independent. then $T = \frac{X}{\sqrt{\frac{Y}{n}}} \sim t_n$.*

*If $X_1, \ldots, X_n$ are i.i.d. $N(\mu, \sigma^2)$ random variables, then $\frac{\bar{X}-\mu}{S}\sqrt{n} \sim t_{n-1}$, with $S = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2$ and $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$.*

**Definition 2.2** *(F-Distribution). A continuous random variable $X$ has a F-distribution with $n$ and $m$ degree of freedom (df), if is has p.d.f.*

$$f(x) = n^{n/2}m^{m/2}\frac{\Gamma(n/2+m/2)}{\Gamma(n/2)\Gamma(m/2)}\frac{x^{n/2-1}}{(nx+m)^{(n+m)/2}}, \quad x \geq 0$$

*This can be written as $F \sim F_{n,m}$.*

*If $X_1 \sim \chi^2_n$ and $X_2 \sim \chi^2_m$ are independent, then*

$$X = \frac{X_1/n}{X_2/m}$$

*has a F-distribution with $n$ and $m$ df.*

*If $Y$ is t-distributed with $m$ df, then $X = Y^2 \sim F_{1,m}$*

| Distribution family | | Density (probability) function | | Parameter |
|---|---|---|---|---|
| Discrete uniform | $X \sim \mathcal{U}\{1,\dots,n\}$ | $P(X=k) = n^{-1}$ | $k = 1,\dots,n$ | $n \in \mathbb{N}$ |
| Uniform (continuous) | $X \sim \mathcal{U}(a,b)$ | $f(x) = (b-a)^{-1}$ | $x \in [a,b]$ | $a < b \in \mathbb{R}$ |
| univariate normal | $X \sim N(\mu,\sigma^2)$ | $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{(x-\mu)^2}{2\sigma^2})$ | $x \in \mathbb{R}$ | $\mu \in \mathbb{R}, \sigma > 0$ |
| multivariate normal | $\mathbf{X} \sim N_n(\boldsymbol{\mu}, \Sigma)$ | $f(\mathbf{x}) = \frac{1}{\sqrt{2\pi\lvert\Sigma\rvert}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)$ | $\mathbf{x} \in \mathbb{R}^n$ | $\boldsymbol{\mu} \in \mathbb{R}^n$, $\Sigma \in \mathbb{R}^{n\times n} p.d$ |
| Chi-squared | $X \sim \chi^2_n$ | $f(x;n) = \frac{1}{2^{n/2}\Gamma(\frac{n}{2})} x^{\frac{n}{2}-1} \exp\left(-\frac{x}{2}\right)$ | $x > 0$ | $n \in \mathbb{N}$ |
| Student's t | $X \sim t_n$ | $f(x) = \frac{\Gamma(\frac{n+1}{2})}{\Gamma(\frac{n}{2})\Gamma(\frac{1}{2})\sqrt{n}}(1+\frac{x^2}{n})^{-\frac{n+1}{2}}$ | $x \in \mathbb{R}$ | $n \in \mathbb{N}$ |
| | | with gamma function for $n > 0$<br>$\Gamma(n) = (n-1)!$ | <br>$n \in \mathbb{N}$ | |
| F | $X \sim F_{n,m}$ | $f(x;n,m) = \frac{n^{n/2}m^{m/2}}{B(n/2,m/2)} \frac{x^{\frac{n}{2}-1}}{(m+nx)^{n+m/2}}$, | $x > 0$ | $n, m \in \mathbb{N}$ |
| | | with Beta function<br>$B(x,y) = \int_0^1 t^{x-1}(1-t)^{y-1}dt$ | <br>$Re(x) > 0$,<br>$Re(y) > 0$ | |

Table 2.1: A list of statistical distributions being used. Note that $\mathbb{N} = \{1, 2, \dots\}$ and *p.d.* is an abbreviation for positive definite, i.e. $\boldsymbol{a}\Sigma\boldsymbol{a} > 0 \ \forall \boldsymbol{a} \in \mathbb{R}^n$.

## 2.2 Univariate and multivariate descriptions and data exploration

### 2.2.1 Histogram

In this paragraph we take a look at some statistical methods to describe univariate data for large data sets. The most relevant tools such as histograms and measures of location and variance, and some simple techniques of explorative data analysis such as box-plots and scatter-plots are illustrated. Moreover it is a basis for multivariate statistical problems. More relevant theory to this subject is discussed in Fahrmeir et al. (2016, Chapter 2 and 3).

Representing the distribution of the raw observations by histograms gives a well-arranged visualization of the grouped frequency table. Supposing an ordinal variable, the data can be grouped in classes by neighboring intervals

$$[c_0, c_1), [c_1, c_2), \ldots, [c_{n-1}, c_n).$$

With the class width of $d_i = c_{i-1} - c_i$ and a resulting height of $f_i/d_i$ which is total or proportional to the absolute or relative frequencies, where $f_i$ is the total or proportional area of the constructed rectangle, the histogram will be built.

Especially for large data sets, we have to choose the number of classes and the width of the classes. We use equal class widths, since individual rectangles do not show much explanatory power about the data. For the number of classes $k$, one often chooses $k = [\sqrt{n}]$, $k = 2[\sqrt{n}]$ or $k = [10log_{10}n]$.

**Example 1** *Given a data set of energy measurements in a house with an appliances variable* ***app****.*

| | $(c_{i-1}, c_i]$ | $f_i$ | | $(c_{i-1}, c_i]$ | frequency |
|---|---|---|---|---|---|
| 1 | [0, 25] | 352 | 12 | (525, 575] | 58 |
| 2 | (25, 75] | 11952 | 13 | (575, 625] | 73 |
| 3 | (75, 125] | 4436 | 14 | (625, 675] | 40 |
| 4 | (125, 175] | 857 | 15 | (675, 725] | 39 |
| 5 | (175, 225] | 364 | 16 | (725, 775] | 21 |
| 6 | (225, 275] | 419 | 17 | (775, 825] | 13 |
| 7 | (275, 325] | 354 | 18 | (825, 875] | 6 |
| 8 | (325, 375] | 310 | 19 | (875, 925] | 4 |
| 9 | (375, 425] | 224 | 20 | (925, 1075] | 1 |
| 10 | (425, 475] | 127 | 21 | (1075, 1125] | 1 |
| 11 | (475, 525] | 84 | | | |

Table 2.2: Frequencies of variable app with the class width $d_i(\texttt{app}) = 50\ \forall i = 1, ..., n$, $n = 19735$.

*The frequency distribution of the characteristic* ***app*** *in watt-hours is shown in the Table 2.2. This forms the basis for the histogram. There are 21 classes for the variable* ***app***

*which have been selected here, i.e. with appropriate class width of* $d_i(\texttt{appliances}) = 50$*, to give an overview and idea. With the rule for the number of classes, we would have* $k = \sqrt{19735} = 140.48 \approx 141$ *or* $k = [10log_{10}(19735)] = 42.95237 \approx 43$*. The latter class* $k \approx 43$ *is sufficient, which we use in the histogram given in Figure 2.1.*

Having a histogram built, a distribution is visible. Is it uni-modal or multi-modal? How many peaks do we have? In case of one peak, so the uni-modal case, it is clear where we have the highest density. In the case multi-modal the interpretation have to be done carefully. Multi-modal distributions can occur if the data is composed from different sub-population or sub-units.

Next step is to look at the symmetry and skewness. The distribution is called symmetric, if there exists a center line, so that the right and left sides of the distribution are approximately reflected to each other. Otherwise the distribution is called skewed. In that case either it is right-skewed or left-skewed.

**Example 2** *Now it gets interesting by representing the frequency distribution of* `app` *through a histogram, we introduced in Table 2.2.*

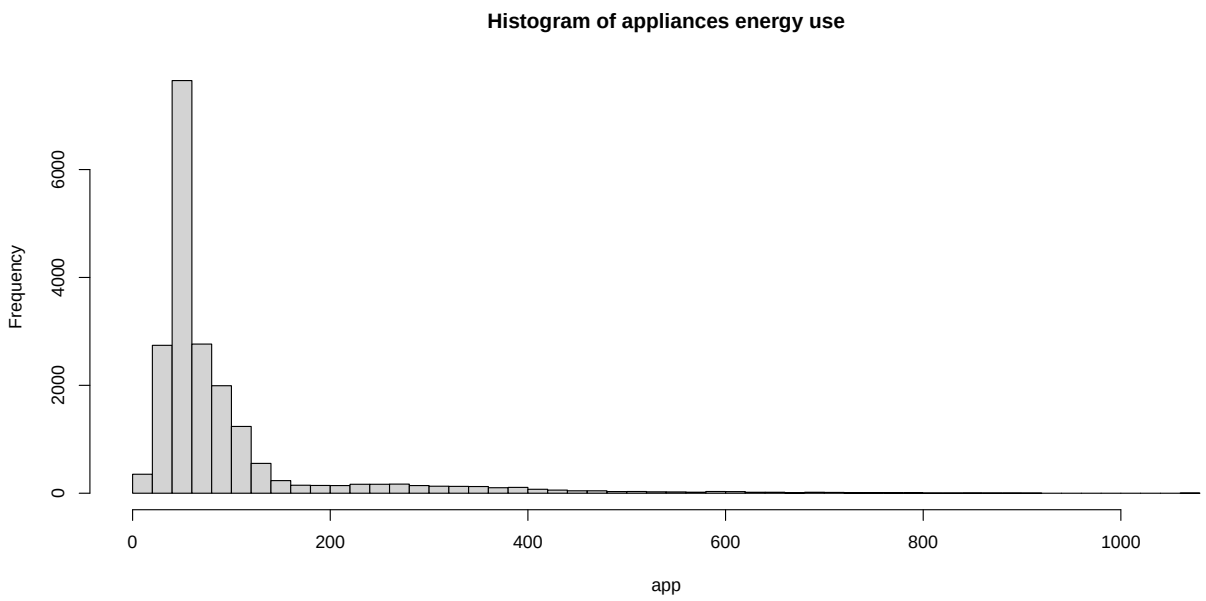



Figure 2.1: Histogram with all data points $n = 19735$ of variable `app` in Wh. The number of classes $k = 43$ is chosen.

*The Figure 2.1 is showing a histogram, where the energy use in Wh are measured at* $n = 19735$ *equidistant time points. The measured energy use of appliances shows the highest energy consumption between 50 and 70 watt-hour over all the time point. We can conclude, the occupants of the house are not using appliances excessively, since the highest frequencies are in the lower range.*

### 2.2.2 Description of distributions

With a first look at graphical visualizations some questions arises. The questions focus on the expected value, statistical spread, symmetry and skewness, outliers and further.

#### Position and measure of central tendencies

A formal quantification for numerical data values of a given distribution is summarized in the following table.

| | | |
|---|---|---|
| Empirical mean: | $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ with n values $x_1, \ldots, x_n$ | sensitive to outliers |
| Empirical median: | $x_{med} = \left\{ \begin{array}{ll} x_{(\frac{n+1}{2})} & n \text{ odd} \\ \frac{1}{2}(x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}) & n \text{ even} \end{array} \right\}$ orders the original list of n values $x_{(1)} \leq \cdots \leq x_{(n)}$ | robust to outliers |
| Empirical mode: | $x_{mod}$: character or value with highest frequency | robust to outliers |

The *rule of position* allows us now to detect symmetry and skewness

(i) symmetric distribution: $\bar{x} \approx x_{med} \approx x_{mod}$

(ii) right-skewed distribution: $\bar{x} > x_{med} > x_{mod}$

(iii) left-skewed distribution: $\bar{x} < x_{med} < x_{mod}$

**Example 3** *Again, take a look at the histogram in Figure 2.1. We already detected the highest peak between 50 and 70 Wh. To be more precise, 50 Wh has the highest frequency, i.e.* $\texttt{app}_{mode} = 50\,Wh$. *In both histograms we have an uni-modal distribution, where the frequencies decrease steeper to the left than to the right, i.e. the distribution is not symmetric, but skewed, more precisely right-skewed. From the definition of the empirical median* $\texttt{app}_{med} = \texttt{app}_{(\frac{19735+1}{2})} = \texttt{app}_{(9868)} = 60$ *and mean* $\overline{\texttt{app}} = \frac{1}{19735}\sum_{i=1}^{19735} \texttt{app}_i = 97.695$, *it is obvious that the these values are bigger than the empirical mode* $\texttt{app}_{mode}$.

### 2.2.3 Quantile and Box-plot

To complete the description of distributions, not only the measurements for position have to be made, also the dispersion of the data around its center should be supplemented. The quantile and the resulting box-plot provide a suitable way to characterize the variation of the data in a graphical summary.

**Definition 2.3** *(Quantile). Let $P$ be a probability distribution. Any value $x_p$ with $0 < p < 1$ is called a p-quantile, if the following holds:*

$$P(x \leq x_p) \geq p \qquad \text{and} \qquad P(x \geq x_p) \geq 1 - p, \quad x \in \mathbb{R}$$

For the metric characteristics, the quantile also provides direct information about the width of the distribution spread. The measure which derives this dispersion is called the interquartile range, i.e. $d_Q = x_{0.75} - x_{0.25}$. The interquartile range is robust against outliers, so to detect outlier candidates we can create an lower and upper border, i.e. $z_l = x_{0.25} - 1.5d_Q$ and $z_u = x_{0.75} + 1.5d_Q$ respectively, and inspect the candidates lying beyond that borders.

To get the full range of the data set, it is useful to examine all the following data quantile points to get all information about the distribution.

The summary of a distribution:

$$x_{min},\ x_{0.25},\ x_{0.5} = x_{med},\ x_{0.75},\ x_{max}$$

Based on these points, the visualization of the distribution through a box-plot is built. With a box-plot it is easy to conclude about the symmetry of observations and to detect outliers and so on.

**Box-plot:**

(i) Box boundaries

$x_{0.25}$ = beginning of the box

$x_{0.75}$ = end of the box

$d_Q$ = length of the box

(ii) The median is marked by the line in the box.

(iii) The whiskers outside the box run up to $z_l$ and $z_u$.

*1. Case* $x_{min}, x_{max} \in [z_l, z_u]$: Then we have $z_l = x_{min}$ and $z_u = x_{max}$.

*2. Case* $x_{min}, x_{max} \notin [z_l, z_u]$: Then the minimum and maximum, including other points that are not in the interval, are marked as circle beyond the whiskers.

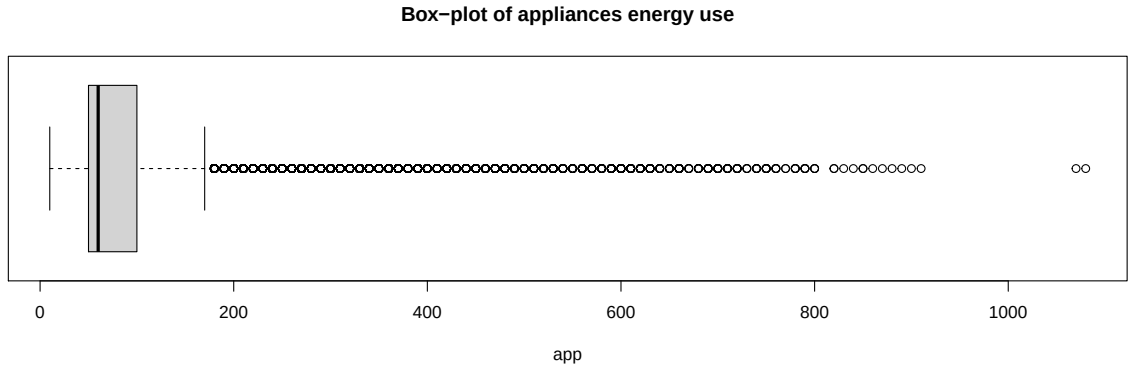



Figure 2.2: Box-plot of the variable `app` in watt-hours.

**Example 4** *The box-plot in Figure 2.2 shows the energy use of the appliances. We see that the median of the appliances energy consumption has a value of 60 Wh, the lower whisker has a value of 10 Wh and the upper whisker has a value of 170 Wh. The box-plots also fortifies in this case that the data above the median varies more, i.e. the empirical distribution is not symmetric, and that there are several outliers which are circled.*

### 2.2.4 Deviation and variance

The well-known measurements of these dispersion of a distribution around their mean $\bar{x}$ are called the deviation and variances.

In the following table, we give an overview of the empirical variance and deviation for the values $x_1, \ldots, x_n$.

| | | |
|---|---|---|
| Empirical variance: | $\bar{s}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$ | sensitive to outliers |
| Empirical standard deviation: | $\bar{s} = +\sqrt{\bar{s}^2}$ | sensitive to outliers |

**Example 5** *For the empirical variance and standard deviation of the variable* $\boldsymbol{app}$*, we use the already calculated mean* $\overline{\boldsymbol{app}} = 97.695$.

$\bar{s}^2 = \frac{1}{19735-1}\sum_{i=1}^{19735}(\boldsymbol{app}_i - \overline{\boldsymbol{app}})^2 = \frac{1}{19735-1}\sum_{i=1}^{19735}(\boldsymbol{app}_i - 97.695)^2 = 10511.35$,

$\bar{s} = +\sqrt{\bar{s}^2} = \sqrt{10511.35} = 102.525$

### 2.2.5 Scatter-plots - a multivariate description

The graphical representation for quantitative characteristics, especially for continuous variables, is the scatter-plot. Measured data $(\mathbf{x}, \mathbf{y}) = (x_1, y_1), \ldots, (x_n, y_n)$ are visualized in a (x, y)-coordinate system and show the relationship between them.

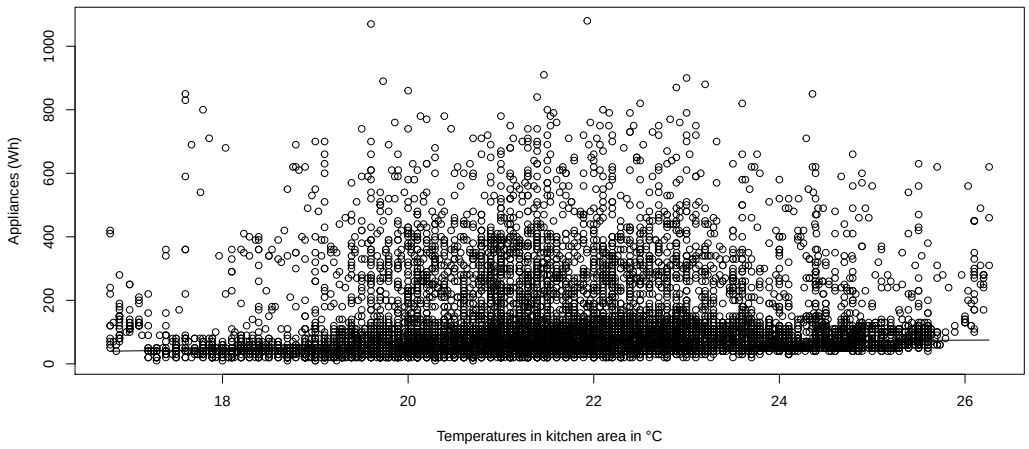



Figure 2.3: Scatter-plot of `app` versus `T1.kitchen`. Linear relationship between the energy consumption of appliances and temperatures in the kitchen.

**Example 6** *A first impression can be gained from scatter-plot in Figure 2.3, whether the two variables are related and the intensity of the relationship.*

*For Figure 2.3 we are taking another variable into account, the temperature measurements in the kitchen. Here, we have lot of continuous data points. One can see that a growing temperature is accompanied, as expected, by a slightly higher appliances energy consumption. As occupants staying in a room, who increasing the temperature in the room by their body temperature, also leads to higher appliances usage. There is a tendency towards larger appliances spread as temperatures grows. Further in the mid temperature range, we have a higher variability. This can be explained by a appliances usage during day and its indoor temperature range.*

### 2.2.6 Correlation coefficient

Scatter-plots and density assessment are graphical tools that can be used to determine the composition of the observation points. What kind of relationship is there between $\mathbf{x}$ and $\mathbf{y}$? A measure for scaling or quantifying the intensity of this correlation is the empirical correlation coefficient, also known as the *Bravais Pearson Correlation Coefficient*.

**Definition 2.4** *(Bravais Pearson correlation coefficient). Given a pair of random variables (X, Y), the correlation coefficient is defined as*

$$\rho = \rho_{X,Y} = Corr(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}. \tag{2.1}$$

*Applying this Pearson's correlation coefficient on a sample, the corresponding estimate of the correlation coefficient is given by*

$$\hat{\rho} = \hat{\rho}_{\mathbf{x},\mathbf{y}} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} = \frac{\tilde{s}_{\mathbf{x},\mathbf{y}}}{\tilde{s}_{\mathbf{x}}\tilde{s}_{\mathbf{y}}}, \tag{2.2}$$

*where* $\tilde{s}_{\mathbf{x},\mathbf{y}} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$ *stands for the empirical covariance and the standard deviation of the characteristics* $\mathbf{x}$ *and* $\mathbf{y}$ *described by* $\tilde{s}_{\mathbf{x}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$ *and* $\tilde{s}_{\mathbf{y}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2}$*, respectively, is for normalization or standardization of* $\hat{\rho}$*.*

*Further, it has the following properties:*

*(i) The range of the correlation coefficient is* $\rho \in [-1, 1]$*.*

*(ii) The direction and strength of the linear relationship for* $\rho$*:*

$\rho > 0$*: positive correlation* $\rho < 0$*: negative correlation*

$\rho = 0$*: no correlation*

*so therefore*

$|\rho| < 0.5$*: weak correlation*

$0.5 \leq |\rho| < 0.8$*: medium correlation*

$0.8 \leq |\rho|$*: strong correlation*

**Example 7** *As we already seen in Figure 2.3, there is a positive estimated correlation between the appliances and kitchen temperature variable with* $\hat{\rho}(\texttt{app}, \texttt{T1.kitchen}) = 0.06$*. Since* $\hat{\rho}(\texttt{app}, \texttt{T1.kitchen}) < 0.5$*, we have a very weak correlation between these characterizations. In Figure 2.4, where we included some more variables, we detect some strong estimated correlations for* $\hat{\rho}(\texttt{T1.kitchen}, \texttt{T2.living}) = 0.84$*,* $\hat{\rho}(\texttt{T1.kitchen}, \texttt{T3.laundry}) = \hat{\rho}(\texttt{T1.kitchen}, \texttt{T5.bath}) = 0.89$ *or* $\hat{\rho}(\texttt{T1.kitchen}, \texttt{T4.office}) = 0.88$*. The largest estimated correlation is determined for* $\hat{\rho}(\texttt{RH3.laundry}, \texttt{RH4.office}) = 0.9$*. Whereas we have no linear dependency between temperature in the kitchen and humidity in the living room, since* $\hat{\rho}(\texttt{T1.kitchen}, \texttt{RH2.living}) = 0.00$*.*

*Overall we can say with respect to* `app` *that the higher the temperatures, the higher the energy consumption by appliances. For the humidity we observe very small and even negative correlations. Furthermore, we have high linear dependencies only between the room temperatures and a higher linear dependencies between the room humidities than between temperatures and humidities.*

For more data exploration tools, we recommend the book of Fahrmeir et al. (2016, Chapter 2 and 3).

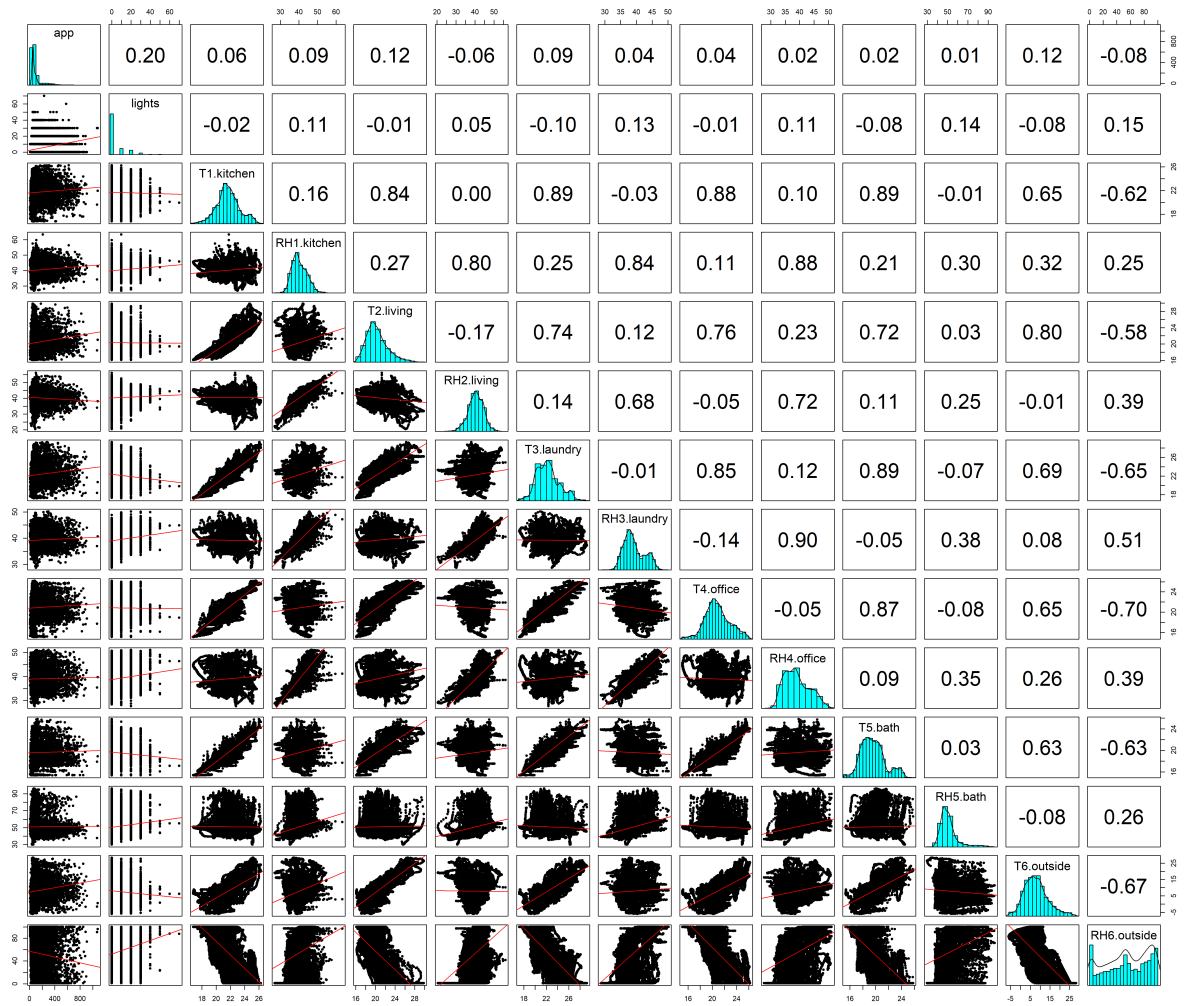



Figure 2.4: Pairs-plot. Relationship between the energy consumption of appliances `app` with `lights`, `T1.kitchen`, `RH1.kitchen`, `T2.living`, `RH2.living`, `T3.laundry`, `RH3.laundry`, `T4.office`, `RH4.office`, `T5.bath`, `RH5.bath`, `T6.outside`, `RH6.outside`. `T` denotes the temperatures, `RH` the humidities. The figure shows bivariate scatter-plots with red linear regression lines below the diagonal, histograms along the diagonal, and the estimated Pearson correlation, measured of linear dependence between two variables, above it.

# Chapter 3

# Multiple linear regression model

In this section we introduce linear models together with their assumptions, estimation procedures and predictions. To get further details and more model settings, we refer to the book of Fahrmeir et al. (2013) and Czado and Schmidt (2011).

## 3.1 Model formulation

In a regression we want to analyze the relationship between the variable of interest the response or dependent variable and other given variables which represent the covariates or independent variables.

Let $\mathbf{Y}$ be our continuous response variable, and let $(X_1, X_2, \ldots, X_p)$ denote the $p$ continuous or categorical/factorial random regressors or predictor variables.

For a given data set of $n$ data points, $x_{1j}, x_{2j}, \ldots, x_{nj}$, for each of the $j = 1, \ldots, p$ predictor, and $y_1, y_2, \ldots, y_n$ associated response values with expectation $\mu_i \equiv E[Y_i]$, $i = 1, \ldots, n$, the main goal is to analyze the influence of the covariates on the mean value of the response variable. Formally our regression model is built as follows:

$$Y_i = E[Y_i|x_1, \ldots, x_p] + \varepsilon_i = f(x_1, \ldots, x_p) + \varepsilon_i, \quad i = 1, \ldots, n, \tag{3.1}$$

For simplification we use the abbreviation of the expression: $E[Y_i] = E[Y_i|x_1, \ldots, x_p]$

The linear regression is a special case of (3.1). In this case the function $f$ is linear, which means that the conditional mean of $Y_i$ is a linear combination of the covariates. Moreover this model is applicable for continuous and approximately independent normal distributed response variables $Y_i \sim N(\mu_i, \sigma_i^2)$ with $\mu_i \equiv E[Y_i]$.

Finally we specify our *multiple linear regression model* using (3.1) with a linear function $f$.

**Definition 3.1** *(Multiple linear regression model). The multiple linear model in terms of their components can be written as:*

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i, \quad i = 1, \ldots, n, \tag{3.2}$$

*where $\beta_0$ is the intercept and $\beta_j$ the unknown regression parameters of the corresponding covariate $x_{ij}$ with $j = 1, \ldots, p$. And $\varepsilon_i$ is the error term.*

**Example 8** *We illustrate a simple multiple linear regression using the data we already introduced in the previous chapter. In Figure 2.4 we have seen scatter-plots between some variables which displayed an approximate linear relationship. Now we build a model in which we are interested in, the linear relationship of response variable appliances and the covariates temperatures in the kitchen and living room.*

$$\texttt{app}_i = \beta_0 + \beta_1 \texttt{T1.kitchen}_i + \beta_2 \texttt{T2.living}_i + \varepsilon_i, \quad \text{for } i = 1, \ldots, 19735. \tag{3.3}$$

*The errors $\varepsilon_i$ are random deviations from the regression line.*

### 3.1.1 Matrix notion in regression

To simplify the model formulation and calculation, we introduce the matrix-vector notation.

In order to rewrite the multiple linear regression model of (3.2) in the matrix-vector notation, we have to define the four different model components.

(i) Vector of the response variables: $\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} \in \mathbb{R}^n$

(ii) Design matrix $\mathbf{X}$, which contains $p$ predictors with their $n$ data points or observations in its rows. The first column equals 1 which corresponds to the intercept $\beta_0$ of the model. Note that there is a column for each covariate, including any added interaction, transformation, indicators, and so on.

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix} \in \mathbb{R}^{n \times (p+1)}$$

Denote by $\boldsymbol{x}_i \in \mathbb{R}^{p+1}$ the $i$-th row of the design matrix.

(iii) Vector of the regression coefficients: $\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} \in \mathbb{R}^{p+1}$

(iv) Vector of random error variables: $\boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix} \in \mathbb{R}^n$

**Definition 3.2** *(Linear model in matrix-vector notation). The multiple linear regression (3.2) can now be formulated as*

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{3.4}$$

**Example 9** *Let's rewrite our example of a multiple linear regression (3.3) in matrix-vector notation.*

$$\texttt{app} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

*with* $\mathbf{Y} := \texttt{app} \in \mathbb{R}^{19735}$, $\mathbf{X} := (\mathbf{1}, \texttt{T1.kitchen}, \texttt{T2.living}) \in \mathbb{R}^{19735\times 3}$, $\boldsymbol{\beta} \in \mathbb{R}^3$ *and* $\boldsymbol{\varepsilon} \in \mathbb{R}^{19735}$

## 3.2 The error term

The additive error terms $\varepsilon_i$ defined as in (3.2) are assumed to be independent and identical distributed *(i.i.d.)* with expectation zero, $E[\varepsilon_i] = 0$, and constant variance to ensure for homoscedasticity across all errors, $Var[\varepsilon_i] = \sigma^2$.

Since we also want to construct confidence intervals and conduct statistical tests, it is reasonable to assume a Gaussian error $\varepsilon_i \sim N(0, \sigma^2)$.

**Definition 3.3** *(Distribution of the error term). The distribution of the error term is assumed to be*

$$\boldsymbol{\varepsilon} \sim N_n(0, \sigma^2 \mathbf{I}_n),$$

*where* $\boldsymbol{I}_n$ *denotes the* $n$*-dimensional identity matrix and* $N_d(\boldsymbol{\mu}, \Sigma)$ *denotes the* $d$*-dimensional multivariate normal distribution with mean vector* $\boldsymbol{\mu}$ *and covariance matrix* $\Sigma$*.*

*This implies the independence of* $\varepsilon_i$ *for* $i = 1, \ldots, n$*.*

These results can be summarized to the following definition of model assumptions.

**Definition 3.4** *(Model assumptions).*

*(i) Linearity of covariate effects: As we introduced in (3.2), the relationship between the covariate vector* $\boldsymbol{x}_i$ *and the random response* $Y_i$ *has the form*

$$Y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \varepsilon_i,$$

*with random error variable* $\varepsilon_i$ *satisfying* $E[\varepsilon_i] = 0$*, so that*

$$E[Y_i] = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij},$$

$i = 1, \ldots, n$*.*

*In matrix notation:* $E[\mathbf{Y}] = \mathbf{X}\boldsymbol{\beta}$*.*

*(ii) Homoscedastic error variances: The error variables* $\varepsilon_i$ *have constant variance*

$$Var[Y_i] = Var[\varepsilon_i] = \sigma^2, \quad i = 1, \ldots, n.$$

*(iii) Uncorrelated error: The random variables* $\varepsilon_i$ *are independent, i.e.*

$$Cov(Y_j, Y_j) = Cov(\varepsilon_j, \varepsilon_j) = 0$$

*(iv) Normality: The random error variables $\varepsilon_i$ are jointly normally distributed. Thus we have*

$$\mathbf{Y} \sim N_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n)$$

Further, the first three assumptions show the additivity of error variables.

**Example 10** *In the case of a multiplicative error structure, we have an exponential relationship between the response and independent variable, where the errors are proportional to the mean value of* $\mathbf{Y}$*. The data then are generated from an exponential model* $Y_i = \exp(\beta_0 + \beta_1 x_{i1}, \ldots, \beta_p x_{ip} + \varepsilon_i)$ *Since it is difficult to interpret these models, we transform models with multiplicative errors. A logarithmic transformation results in a linear model with additive errors*

$$\log(Y_i) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i, \quad i = 1, \ldots, n.$$

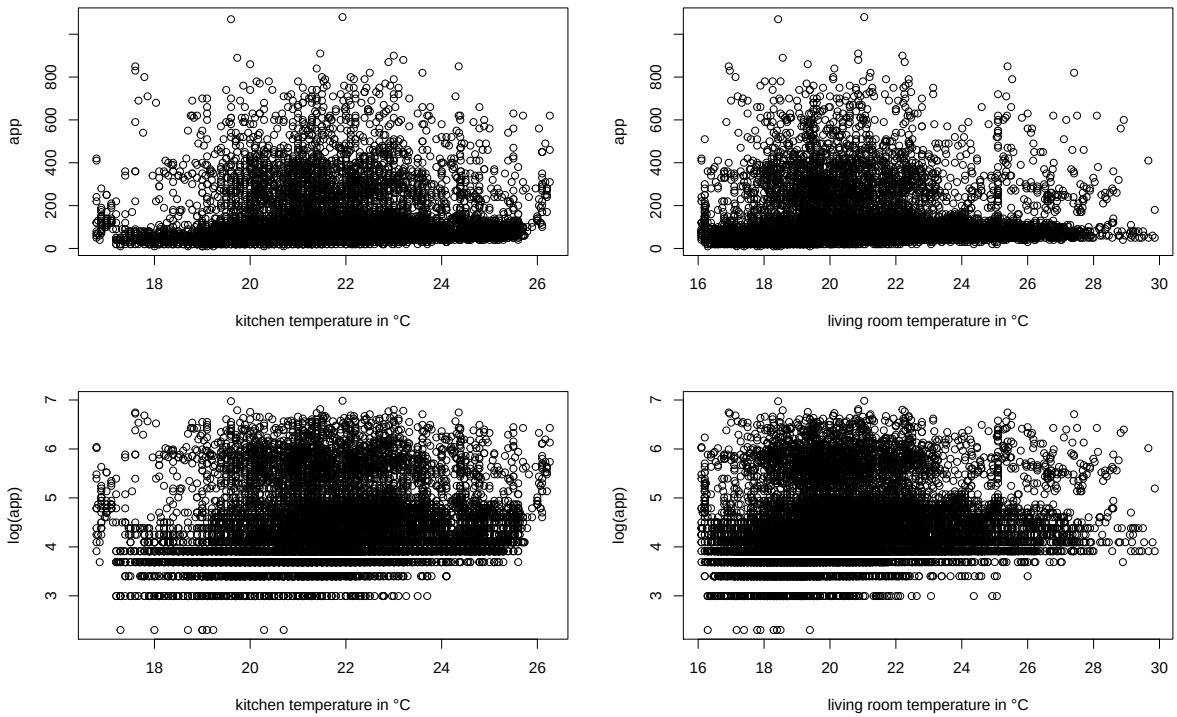



Figure 3.1: Scatter-plots. Relationship between the appliances energy consumption and two continuous covariates `T1.kitchen` and `T2.living`. The top panel of the figure shows the relation of the original response variable with the two covariates and the two graphs in the lower panel show the improved relation with the transformed response variable, i.e. $\log(\texttt{app}_i)$ , $i = 1, \ldots, 19735$, we created in (3.5).

**Example 11** *Example 10 is applicable to our multiple regression (3.3), due to the right-skewed of the appliances variable (c.f. Figure 2.1) and the measurement at equidistant*

*time points which makes the errors not additive (c.f. scatter-plots in Figure 2.3 and 2.4). So transferred to our example regression (3.3), we get*

$$\log(\texttt{app}_i) = \beta_0 + \beta_1 \texttt{T1.kitchen}_i + \beta_2 \texttt{T2.living}_i + \varepsilon_i, \quad i = 1, \ldots, 19735. \tag{3.5}$$

*In Figure 3.1 we are doing scatter-plots so see the relationships between the response variable and the covariates. Furthermore, we compare the relations with the original response variable and the transformed one from (3.5).*

## 3.3 Modeling the effects of covariates

As already hinted in the model equation settings, not only linear relationships are possible, we can also fit non-linear relationships within the class of linear models. Dealing with continuous explanatory variables, it is often necessary to take non-linear methods into account. The two most established alternatives are the *variable transformations* and *polynomial regression*.

**Example 12** *A simple transformation which can be used to model a non-linear relationship is given by $f(x) = \frac{1}{x}$. Setting $p = 1$ and assuming the regression model (3.2), the model results in $Y_i = \beta_0 + \beta_1 f(x_i) + \varepsilon_i = \beta_0 + \beta_1 \frac{1}{x_i} + \varepsilon_i$. Other transformation are, of course, possible, e.g. $f(x_i) = \log(x_i)$.*

### 3.3.1 Polynomial regression

Here we focus on the polynomial regression since we are only dealing with this method for our *energy use data set* (c.f. Table 5.2), we introduce later in Chapter 6 and 7. In this case non-linear covariate effects are fitted through polynomials.

**Theorem 3.5** *(Polynomial regression). If the continuous covariate $Z_i$ has an approximately polynomial effect of an degree d, then the generated model $Y_i = \beta_0 + \beta_1 Z_i + \beta_2 Z_i^2 + \cdots + \beta_d Z_i^d + \cdots + \varepsilon_i$ can be transformed into our well-known linear regression model*

*$Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_d x_{id} + \cdots + \varepsilon_i$,*

*where the polynomial of degree d is substituted by $x_{i1} = Z_i^1$, $x_{i2} = Z_i^2$ and $x_{id} = Z_i^d$.*

*The centering of the vectors $\mathbf{X}^j = (x_{1j}, \ldots, x_{nj})'$ for all $j = 1, \ldots, d$, to $\mathbf{X}^1 - \bar{\mathbf{X}}_d, \ldots, \mathbf{X}^d - \bar{\mathbf{X}}_d$, with mean vector $\bar{\mathbf{X}}_j = (\bar{X}_j, \ldots, \bar{X}_j)'$ helps with the interpretation of the estimated effects. Further orthogonalization improves numerical instability of the estimation procedure.*

**Definition 3.6** *(Gram-Schmidt Orthogonalization, detailed process in Leon et al. (2013)). The vectors $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n$ of the orthogonal system is calculated by the algorithm as follows:*

$$\boldsymbol{v}_1 = \boldsymbol{w}_1, \qquad \boldsymbol{v}_n = \boldsymbol{w}_n - \sum_{i=1}^{n-1} \frac{\langle \boldsymbol{v}_i, \boldsymbol{w}_n \rangle}{\langle \boldsymbol{v}_i, \boldsymbol{v}_i \rangle} \boldsymbol{v}_i, \tag{3.6}$$

*where $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_n$ are linear independent vectors, i.e. equation $\sum_{i=1}^n \lambda_i \boldsymbol{w}_i = 0$ can only be satisfied by $\lambda_i = 0$ for $i = 1, \ldots, n$. Furthermore, $\langle \boldsymbol{a}, \boldsymbol{b} \rangle$ denotes the scalar product, i.e. $\langle \boldsymbol{a}, \boldsymbol{b} \rangle = \sum_{i=1}^n a_i b_i$ with vectors $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^n$.*

| hour | $poly(\texttt{hour}_i, 1)$ | $poly(\texttt{hour}_i, 2)$ | $poly(\texttt{hour}_i, 3)$ |
|---|---|---|---|
| 1 | -0.012 | 0.014 | -0.015 |
| 2 | -0.011 | 0.010 | -0.007 |
| 3 | -0.0098 | 0.0071 | -0.0011 |
| 4 | -0.0087 | 0.0041 | 0.0032 |
| 5 | -0.0077 | 0.0014 | 0.0062 |
| 6 | -0.0067 | -0.00094 | 0.0079 |
| 7 | -0.0057 | -0.0029 | 0.0085 |
| 8 | -0.0046 | -0.0046 | 0.0082 |
| 9 | -0.0036 | -0.0059 | 0.0071 |
| 10 | -0.0026 | -0.0069 | 0.0055 |
| 11 | -0.0016 | -0.0076 | 0.0035 |
| 12 | -0.00052 | -0.0079 | 0.0012 |
| 13 | 0.00051 | -0.0079 | -0.0012 |
| 14 | 0.0015 | -0.0076 | -0.0035 |
| 15 | 0.0026 | -0.0069 | -0.0055 |
| 16 | 0.0036 | -0.0059 | -0.0071 |
| 17 | 0.0046 | -0.0046 | -0.0082 |
| 18 | 0.0056 | -0.0029 | -0.0085 |
| 19 | 0.0067 | -0.00095 | -0.0078 |
| 20 | 0.0077 | 0.0014 | -0.0062 |
| 21 | 0.0087 | 0.0041 | -0.0032 |
| 22 | 0.0098 | 0.007 | 0.0011 |
| 23 | 0.011 | 0.01 | 0.007 |
| 24 | 0.012 | 0.014 | 0.015 |

Table 3.1: Summary of polynomial coefficients of a variable `hour`.

*For the linear independent vectors $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_n$, the algorithm calculates an orthogonal system of n pairwise orthogonal vectors. It generates the same subspace.*

*Due to the rounding errors, the process is numerically unstable. To stabilize the Gram-Schmidt process as defined in the algorithm above (3.6), we do a small modification. If the process is implemented for $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_{k-1}$, the vector $\boldsymbol{v}_k$ for $1 \leq k \leq n$ is then computed by*

$$\boldsymbol{v}_k = \boldsymbol{w}_k - \sum_{i=1}^{k-1} \frac{\langle \boldsymbol{v}_i, \boldsymbol{w}_k \rangle}{\langle \boldsymbol{v}_i, \boldsymbol{v}_i \rangle} \boldsymbol{v}_i,$$

*so that $\boldsymbol{v}_k$ is orthogonal to all the vectors $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_{k-1}$, i.e. $\langle \boldsymbol{v}_j, \boldsymbol{v}_k \rangle = 0 \ \forall j = 1, \ldots, k-1$.*

**Example 13** *For random variables $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_d$, we yield with the Gram-Schmidt Orthogonalization and orthogonal system of $\boldsymbol{X}_1^*, \ldots, \boldsymbol{X}_d^*$ with $\boldsymbol{X}_k^{*T} \boldsymbol{X}_j^* = 0 \forall k \neq j$.*

*To give an idea of an output of the process, see the following example of an order three polynomial of a covariate `hour`, i.e. poly(`hour`, 3), calculated by R (c.f. Table 3.1).*

### 3.3.2 Interactions between covariates

If there exist a coupling effect between two or more covariates, that is a covariate effect which depends on the value of at least one other independent variable, it is called an interaction between covariates.

**Example 14** *To give a foundation of interactions, consider the simple regression model with response variable* $\mathbf{Y}$ *and two predictors* $\boldsymbol{x}_1 = (x_{i1})_{i=1,\dots,n}$ *and* $\boldsymbol{x}_2 = (x_{i2})_{i=1,\dots,n}$ *and an interaction between these two predictors*

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1} x_{i2} + \varepsilon_i, \quad i = 1, \dots, n. \tag{3.7}$$

*The terms* $\beta_1 x_{i1}$ *and* $\beta_2 x_{i2}$ *depending only on one variable are called the main effects, whereas the term* $\beta_3 x_{i1} x_{i2}$ *is called the interaction between the two covariates* $\boldsymbol{x}_1$ *and* $\boldsymbol{x}_2$*.*

*To interpret the interaction term, examine the change of* $E[\mathbf{Y}]$ *when one variable change by* $u$ *units, e.g. adding* $u$ *to the first covariate* $\boldsymbol{x}_1$ *and we have*

$$\begin{aligned} E[Y_i|x_{i1}+u, x_{i2}] - E[Y_i|x_{i1}, x_{i2}] &= \beta_0 + \beta_1(x_{i1}+u) + \beta_2 x_{i2} + \beta_3(x_{i1}+u)x_{i2} \\ &\quad - \beta_0 - \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1} x_{i2} \\ &= \beta_1 u + \beta_3 u x_{i2} \end{aligned}$$

*Now we have the distinction of the two cases* $\beta_3 = 0$ *and* $\beta_3 \neq 0$*.*

- $\underline{\beta_3 = 0}$*: The interaction is canceled from the model, just main effects are included. The expected change* $\beta_1 u$ *is independent from the value of the second predictor* $\boldsymbol{x}_2$*.*
- $\underline{\beta_3 \neq 0}$*: The expected change* $\beta_1 u + \beta_3 u x_{i2}$ *depends on the added amount* $u$ *and also on the value of the second covariate* $\boldsymbol{x}_2$*.*

Therefore adding an interaction term is required when the effect of changing a covariate depends on the value of another covariate.

An important aspect of the interaction terms is that we should always check the interaction term first, like we did in the case-by-case analysis in our example. After the interaction term is included, we can proceed with significance testing of the main effects. A removed main variable should not be included in any interaction term involving this main effect.

**Example 15** *Returning to our modified multiple linear regression equation (3.5), we can add an interaction term between the covariates* `T1.kitchen` *and* `T2.living`*. Doing so, we can detect possible interaction between these two predictors.*

$$\begin{aligned} \log(\texttt{app}_i) = \; & \beta_0 + \beta_1 \texttt{T1.kitchen}_i + \beta_2 \texttt{T2.living}_i \\ & + \beta_3(\texttt{T1.kitchen}_i \times \texttt{T2.living}_i) + \varepsilon_i, \quad i = 1, \dots, 19735. \end{aligned} \tag{3.8}$$

*After we modeled the regression (3.8), we have to estimate the regression coefficient to see whether our interaction coefficient* $\beta_3$ *differs from zero, i.e. there exists an interaction, or equals zero, i.e. there is no interaction.*

In the next section we take a look on the estimation procedure of the model parameter.

## 3.4 Model parameters, estimation, and residuals

From our model assumptions we get the following formula where we use the estimator $\hat{\boldsymbol{Y}}$, that is the estimated linear function in (3.1), to predict $\mathbf{Y}$:

$$\hat{\mathbf{Y}} = \hat{f}(x_1, \ldots, x_p) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_p x_p \tag{3.9}$$

But how we compute our estimator?

Before we answer that question, we give some preparations and important settings. To simplify the representation, let's use the matrix notation. So recap, the estimator of the mean $E[Y_i]$ of $Y_i$ is given by

$$\begin{aligned} \hat{Y}_i &= \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_p x_{ip} \\ &= \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}. \end{aligned} \tag{3.10}$$

Furthermore, with the residual, which is the deviation between the true value $y_i$ and estimated value $\hat{y}_i$ denoted by $\hat{\boldsymbol{\varepsilon}} = (\hat{\varepsilon}_1, \ldots, \hat{\varepsilon}_n)'$, we have

$$\hat{\boldsymbol{\varepsilon}} = \mathbf{y} - \hat{\mathbf{y}} = \mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}. \tag{3.11}$$

Note that $\hat{\beta}_i \neq \beta_i$ for all $i = 1, \ldots, n$, as a consequence of the unknown true parameter $\beta$ without the error, c.f. the multiple linear model (3.2).

As for the regression parameter vector $\boldsymbol{\beta}$, the residuals $\hat{\boldsymbol{\varepsilon}}$ is not fully identical to our error $\boldsymbol{\varepsilon}$ we introduced in the last section. The residuals $\hat{\varepsilon}_i$ can be identified as predictions of $\varepsilon_i$.

### 3.4.1 Method of least squares - estimation of the regression coefficient

Now that we presented the important assumption we answer the question that arose in the previous section on how we develop estimators for unknown parameters $\boldsymbol{\beta}$ and $\sigma^2$ of the linear model and their statistical properties. The most common method for estimating regression parameters $\boldsymbol{\beta}$ is the method of least squares.

In this section we give special statistical properties of the typical method of least squares. Although the method has much advantages, like simple mathematical formula which can be differentiate, the estimators are highly sensitive to outliers.

Minimizing the sum of the squared deviations from residual equation (3.11), we get the estimated values for the regression coefficients $\boldsymbol{\beta}$:

**Definition 3.7** *(Sum of squared deviations). The sum of the squared deviations to obtain the estimated values for the regression coefficients $\boldsymbol{\beta}$ is defined as*

$$LS(\boldsymbol{\beta}) = \sum_{i=1}^{n} (y_i - \boldsymbol{x}_i^T \boldsymbol{\beta})^2 = \sum_{i=1}^{n} \hat{\varepsilon}_i^2 = \hat{\boldsymbol{\varepsilon}}^T \hat{\boldsymbol{\varepsilon}} \tag{3.12}$$

*for given data $(y_i, x_i)$, $i = 1, 2, \ldots, n$ and with respect to $\boldsymbol{\beta} \in \mathbb{R}^{p+1}$.*

To minimize $LS(\boldsymbol{\beta})$ (3.12), we take the derivative and getting the result easily that is

$$\frac{\partial LS\,(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = -2\mathbf{X}^T\mathbf{y} + 2\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} \tag{3.13}$$

From taking the second derivative with outcome $\frac{\partial^2 LS(\boldsymbol{\beta})}{\partial^2 \boldsymbol{\beta}} = 2\mathbf{X}^T\mathbf{X}$ and the fact that matrix $\mathbf{X}^T\mathbf{X}$ is positive definite due to $\text{rank}(\mathbf{X}) = p + 1$, we know that there exists a solution.

Finally by solving the *normal equations*

$$\mathbf{X}^T\mathbf{y} = \mathbf{X}'\mathbf{X}\boldsymbol{\beta}, \tag{3.14}$$

which have a unique solution, we yield our least squares estimator.

**Lemma 3.8** *(Least squares estimator). The resulting least squares estimate from the normal equations is given by*

$$\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}(\mathbf{y}) = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \tag{3.15}$$

*The associated estimator $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}(\mathbf{Y})$ is given by*

$$\hat{\boldsymbol{\beta}}(\mathbf{Y}) = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} \tag{3.16}$$

**Example 16** *Reviewing our Example 15, we already fitted our regression model*
$\widehat{log(\texttt{app})} = \hat{\beta}_0 + \hat{\beta}_1 \texttt{T1.kitchen} + \hat{\beta}_2 \texttt{T2.living} + \hat{\beta}_3 \texttt{T1.kitchen} \times \texttt{T2.living}$.

*Looking at the summary of this interaction model we derived in R, we get the estimated model parameter*

$$\begin{aligned} \hat{\beta}_0 &= -1.310893, \quad \hat{\beta}_1 = 0.171617, \\ \hat{\beta}_2 &= 0.315548, \quad \hat{\beta}_3 = -0.010191. \end{aligned}$$

*These estimated regression coefficients are calculated by the method of the least squares, i.e. the least squares estimate (3.15), where*
$\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3)^T \in \mathbb{R}^4$,
$\mathbf{y} = ((log(\texttt{app}_i))_{i=1,\ldots,19735}) \in \mathbb{R}^{19735}$ *and*
$\boldsymbol{X} = (\mathbf{1}, \texttt{T1.kitchen}, \texttt{T2.living}, \texttt{T1.kitchen} \times \texttt{T2.living}) \in \mathbb{R}^{\mathbf{19735\times 4}}$.

*Continuing our example examine the interaction, we first consider the interaction term $\hat{\beta}_3$, which is very small. But observing all the small estimated model parameter, we can say that an interaction term could be needed, since $\hat{\beta}_3 \neq 0$. So the effect of changing* `T1.kitchen` *depends on the value of* `T2.living`*, i.e. adding $u := 1$ unit to the covariate* `T1.kitchen`*, the expected change on the response variable is $\hat{\beta}_1 u + \hat{\beta}_3 u \texttt{T2.living}_i = 0.171617 - 0.010191 \cdot \texttt{T2.living}_i$.*

*Whether or not an inclusion of an interaction effect is really necessary can be statically tested using the hypothesis*

$$H_0 : \hat{\beta}_3 = 0 \quad versus \quad H_1 : \hat{\beta}_3 \neq 0,$$

*see later in Section Statistical inference and F-test.*

*Note that if the interaction terms is needed thus we can not remove the main effect regardless of the $\hat{\beta}_1$ and $\hat{\beta}_2$.*

Taking a look at the next method to estimate another term, our goal will be to show the method of least squares estimator with Gaussian error coincide with maximum likelihood estimator of the regression coefficients.

### 3.4.2 Maximum likelihood estimation

Now that we have the least squares estimator, we have to specify the distributional assumptions with regard to the error term $\varepsilon$. As we introduced in the model equation settings we assume normally distributed errors $\boldsymbol{\varepsilon} \sim N_n(0, \sigma^2 \mathbf{I}_n)$ and so we have $\mathbf{Y} \sim N_n(\boldsymbol{X\beta}, \sigma^2 \mathbf{I}_n)$.

It follows the likelihood equation.

**Definition 3.9** *The likelihood of* $(\boldsymbol{\beta}, \boldsymbol{\sigma})$ *given the data values* $\mathbf{y}$ *is*

$$L(\boldsymbol{\beta}, \boldsymbol{\sigma}|\mathbf{y}) = \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} \exp\left(-\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right) \tag{3.17}$$

The corresponding log-likelihood is thus given by

$$l(\boldsymbol{\beta}, \boldsymbol{\sigma}|\mathbf{y}) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}). \tag{3.18}$$

Maximizing the log-likelihood with respect to $\boldsymbol{\beta}$, we only have to consider

$$M = -\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}),$$

since the first two terms of log-likelihood (3.18) are independent of $\boldsymbol{\beta}$.

Note that the maximum likelihood estimates of the regression parameter $\boldsymbol{\beta}$ under normality assumption is equivalent to minimize the least squares criterion (3.12) $LS(\boldsymbol{\beta}) = \hat{\boldsymbol{\varepsilon}}^T\hat{\boldsymbol{\varepsilon}} = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$.

Therefore the maximum likelihood estimate $\hat{\boldsymbol{\beta}}$ is also the least squares estimate.

Not only the $\boldsymbol{\beta}$ can be estimated using maximum likelihood, also the variance $\sigma^2$ is estimated by differentiation of the log-likelihood with respect to $\sigma^2$ and setting to zero.

$$\frac{\partial l(\boldsymbol{\beta}, \sigma^2)}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \stackrel{!}{=} 0$$

The estimate for the variance $\sigma^2$ is obtained by substituting the estimate $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$ into the differential of the log-likelihood

$$-\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})^T(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}(\mathbf{y} - \hat{\mathbf{y}})'(\mathbf{y} - \hat{\mathbf{y}}) \stackrel{!}{=} 0$$

which yields the estimate

$$\hat{\sigma}^2_{ML} = \frac{1}{n}(\mathbf{y} - \hat{\mathbf{y}})^T(\mathbf{y} - \hat{\mathbf{y}}) = \frac{1}{n}\hat{\boldsymbol{\varepsilon}}^T\hat{\boldsymbol{\varepsilon}}. \tag{3.19}$$

The corresponding estimator is

$$\hat{\sigma}^2_{ML}(\mathbf{Y}) = \frac{1}{n}(\mathbf{Y} - \hat{\mathbf{Y}})^T(\mathbf{Y} - \hat{\mathbf{Y}}) = \frac{1}{n}\hat{\boldsymbol{\varepsilon}}^T\hat{\boldsymbol{\varepsilon}}. \tag{3.20}$$

To create the unbiased estimator $\hat{\boldsymbol{\sigma}}^2_{REML}$, the expectation of the sum of squared residual $E[\hat{\boldsymbol{\varepsilon}}^T\hat{\boldsymbol{\varepsilon}}] = (n-p)\sigma^2$ is used. Thus we get

$$\hat{\boldsymbol{\sigma}}^2_{REML} = \frac{1}{n-p}\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}. \tag{3.21}$$

**Example 17** *Calculate the estimate of $\sigma^2$ using the corresponding formula of the unbiased estimator $\hat{\boldsymbol{\sigma}}^2_{REML}$ in Equation (3.21).*

*Again utilize Example 15, we can determine the unbiased estimate of $\sigma^2$ by substituting the response, covariates, the number of data points and predictors into the corresponding unbiased estimate $\hat{\boldsymbol{\sigma}}^2_{REML}$, analogously we did in the last Example 16 . Doing this in the program R, we yield the result*

$$\hat{\boldsymbol{\sigma}}^2_{REML} = \frac{1}{n-p}(\mathbf{y}-\hat{\mathbf{y}})^T(\mathbf{y}-\hat{\mathbf{y}}) = 0.40855.$$

### 3.4.3 Distribution of the estimators

Define the vector of the fitted random values $\hat{Y}_i$ with the help of the calculated parameter estimators as

$$\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} \tag{3.22}$$

Further, define the *hat matrix* which presents the projection of $\mathbf{Y}$ onto the space that spanned by the columns of $\mathbf{X}$ by

$$\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T \tag{3.23}$$

The projection matrix $\mathbf{H} \in \mathbb{R}^{n\times n}$ is symmetric, i.e. $\mathbf{H}^T = \mathbf{H}$, and idempotent, i.e. $\mathbf{H}^2 = \mathbf{H}$.

Since the estimator of regression coefficients, fitted values and raw residuals are linear functions, the transformation rule for expectation and variance-covariance matrix can be applied so that

$$\begin{array}{ll} E[\hat{\boldsymbol{\beta}}] = \boldsymbol{\beta}, & Var[\hat{\boldsymbol{\beta}}] = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}, \\ E[\hat{\mathbf{Y}}] = \mathbf{X}\boldsymbol{\beta}, & Var[\hat{\mathbf{Y}}] = \sigma^2(\mathbf{H}), \\ E[\hat{\boldsymbol{\varepsilon}}] = \mathbf{0}, & Var[\hat{\boldsymbol{\varepsilon}}] = \sigma^2(\mathbf{I}_n - \mathbf{H}) \end{array}$$

Thus under normality assumption we have

$$\begin{array}{l} \hat{\boldsymbol{\beta}} \sim N_{p+1}(\boldsymbol{\beta}, \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}), \\ \hat{\mathbf{Y}} \sim N_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2(\mathbf{H})), \\ \hat{\boldsymbol{\varepsilon}} \sim N_n(0, \sigma^2(\mathbf{I}_n - \mathbf{H})). \end{array} \tag{3.24}$$

## 3.5 Performance of regression models

### 3.5.1 Analysis of Variance

With the help of the given properties of the least squares estimator, we introduce a fundamental analysis of variance formula for the empirical variance of the observed response

$y_i$. This formula can be used as goodness-of-fit measure. In the analysis of variance, we have many ways to test the explanatory power by the regression model. For our purpose we study the coefficient of determination. Note that all the assumption of the linear regression model have to be fulfilled.

Consider the following additive decomposition formula:

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}\hat{\varepsilon}_i^2, \tag{3.25}$$

where $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$.

The employed sum of squares are introduced to quantify the amount of variation explained by the regression, which are defined as

(i) $\text{SST} := \sum_{i=1}^{n}(y_i - \bar{y})^2$ as total sum of squares

(ii) $\text{SSR} := \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$ as regression sum of squares

(iii) $\text{SSE} := \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}\hat{\varepsilon}_i^2$ as error sum of squares.

The decomposition formula follows from the fact that $\sum_{i=1}^{n}(\hat{y}_i - \bar{y})(y_i - \hat{y}_i) = 0$. So from this, it follows that

$$\text{SST} = \text{SSR} + \text{SSE} \tag{3.26}$$

### 3.5.2 Coefficient of Determination - $R^2$ statistic

The quantity $R^2$ is a measure of how well our model predicts $\mathbf{Y}$.

Using decomposition Formula (3.25) or (3.26), we obtain the coefficient of determination $R^2$ as a goodness-of-fit measure.

**Definition 3.10** *(Multiple coefficient of determination).*
*Define the coefficient of determination as*

$$R^2 := \frac{\text{SSR}}{\text{SST}} = 1 - \frac{\text{SSE}}{\text{SST}}. \tag{3.27}$$

*Additionally we define the adjusted multiple coefficient of determination as*

$$R^2_{adj} := 1 - \frac{\text{SSE}/(n-p)}{\text{SST}/(n-1)}. \tag{3.28}$$

From the decomposition formula, the coefficient of determination is in the range of $0 \leq R^2 \leq 1$.

**Example 18** *The limit case $R^2 = 1$ implies $\text{SSE} = 0$, i.e. all residuals are zero and fits our data perfectly. Whereas the limit $R^2 = 0$ implies $\text{SSR} = 0 \Rightarrow \hat{y}_i = \bar{y}_i$ for all $i$. Thus the prediction of $Y_i$ is independent of the explanatory variables which means that the covariates do not have a explanatory power for the mean of $\mathbf{Y}$. In this case we also have to consider the uncovered explanatory power that is a non-linear relationship.*

Consequently, the closer $R^2$ is to 1, the better our variability is explained by our model.

Finally, we come to the weaknesses of $R^2$. The $R^2$ typically increase when more predictor variables are added to a model, which makes the $R^2$ not really appropriate for comparing two or more models. In fact, $R^2$ is the fundamental model comparison statistic, but only for models with the same number of predictors. So it is essential developing a criterion which allows to compare models with different numbers of predictors and do not penalize adding more regression parameter or comparing reduced models.

Including a correction term for the number of regression parameter, we get the adjusted $R^2$.

$$R^2_{adj} := 1 - \frac{n-1}{n-p}(1-R^2) = 1 - \frac{\text{SSE}/(n-p)}{\text{SST}/(n-1)}$$

**Example 19** *We compare the modified linear regressions (3.5) and (3.8) we set. Table 3.2 contains the estimated models and their corresponding adjusted coefficients of determination. We observe that every $R^2_{adj}$ is very small. One reason is the strong variability in the data. (see scatter-plot in Figure 3.1). Further, there are many important covariates of the original energy use data set missing in the fitted models we introduced in the examples. The models in the application parts later are containing at least ten of our explanatory variables, which leads to an higher adjusted coefficient of determination.*

| Model | Equation | | $R^2_{adj}$ |
|---|---|---|---|
| M1: | $\widehat{\log(\texttt{app})} =$ | $3.235 - 0.026 \cdot \texttt{T1.kitchen} + 0.080 \cdot \texttt{T2.living}$ | 0.04722 |
| M2: | $\widehat{\log(\texttt{app})} =$ | $-1.311 + 0.172 \cdot \texttt{T1.kitchen} + 0.316 \cdot \texttt{T2.living}$ $-0.01 \cdot (\texttt{T1.kitchen} \times \texttt{T2.living})$ | 0.05152 |

Table 3.2: Comparing the adjusted coefficient of determination $R^2_{adj}$ of different models for the relationship between appliances energy use and the temperatures.

*If we compare these two models, we find a higher $R^2_{adj}$ for the interaction model M2. This means that the interaction model should be preferred. (Compare later with residual analysis Figure 3.2).*

*Further, the two models are nested, since M1 is included in M2. While the normal coefficient of determination increases typically with increasing number of parameters for nested models, the adjusted coefficient of determination punishes the growing parameters. For a comparison see the original coefficient of determination of these two models $R^2(M1) = 0.04732$ and $R^2(M2) = 0.05166$. Even though the values are very small and hard to compare, the difference between the original coefficient of determination is greater than the difference of the adjusted coefficient of determination of M1 and M2.*

*Note that we only compared the model with the modified response variable, since it is not possible to compare models with different response variables, i.e. the original response $y$ and the transformed response $\log(y)$.*

### 3.5.3 Model selection - AIC and BIC

We have already seen a way to choose a better fitted model based on the adjusted coefficient of determination. Regarding to prediction quality in linear models, now we learn a another way to compare different (parametric) regression model with model choice criteria. Review some assumptions, that is

- independent observations $y_i,\ \ i = 1, \ldots, n$ with

  expectation $E[y_i] = \mu_i$ and

  variance $Var[y_i] = \sigma^2$, and

- given potential regressors $(1, \mathbf{x}_1, \ldots, \mathbf{x}_p)$.

For estimations, a subset of included covariates $M \subset \{0, 1, \ldots, p\}$ are used. We obtain the least square estimator

$$\hat{\boldsymbol{\beta}}_M = (\mathbf{X}_M^T \mathbf{X}_M)^{-1} \mathbf{X}_M^T \mathbf{Y},$$

with the corresponding design matrix $\mathbf{X}_M$. The estimator for the vector $\boldsymbol{\mu}$ of means $\mu_i = E[Y_i]$ is now given as

$$\hat{\mathbf{Y}}_M = \mathbf{X}_M \hat{\boldsymbol{\beta}}_M.$$

The Akaike information criterion (AIC) is one of the most commonly used criteria for choosing a model. For the AIC formula within the scope of likelihood-based inference, see (Fahrmeir et al., 2013, Appendix B, p. 664).

**Definition 3.11** *(Akaike information criterion, AIC).*
*Let the notion be given as above. In general, AIC is defined by*

$$AIC = -2 \cdot l(\hat{\boldsymbol{\beta}}_M, \hat{\sigma}^2) + 2(|M| + 1), \tag{3.29}$$

*where $l(\hat{\boldsymbol{\beta}}_M, \hat{\sigma}^2)$ is the maximum value of the log-likelihood, i.e. the ML estimators $\hat{\beta}_M$ and $\hat{\sigma}^2$ are inserted into the log-likelihood.*

*Note that $(|M| + 1)$ is the total number of parameter and the error variance $\sigma^2$ which is counted as a parameter.*

In case of Gaussian errors in a linear model, we have

$$\begin{aligned} -2l(\hat{\boldsymbol{\beta}}_M, \hat{\sigma}^2) &= n \log(\hat{\sigma^2}) + \tfrac{1}{\hat{\sigma}^2} (\mathbf{y} - \mathbf{X}_M \hat{\boldsymbol{\beta}}_M)^T (\mathbf{y} - \mathbf{X}_M \hat{\boldsymbol{\beta}}_M) \\ &= n \log(\hat{\sigma}^2) + \tfrac{n\hat{\sigma}^2}{\hat{\sigma}^2} \\ &= n \log(\hat{\sigma}^2) + n. \end{aligned}$$

Ignoring the constant $n$, we yield for the AIC formula

$$AIC = n \log(\hat{\sigma}^2) + 2(|M| + 1)$$

AIC is considering the ML estimator $\hat{\sigma}^2 = \hat{\boldsymbol{\varepsilon}}^T \hat{\boldsymbol{\varepsilon}}/n$ for the computation, not the usual unbiased variances estimator $\hat{\sigma}^2_{REML}$.

**Definition 3.12** *(Bayesian Information Criterion, BIC). The BIC is generally defined by*

$$BIC = -2 \cdot l(\hat{\boldsymbol{\beta}}_M, \hat{\sigma}^2) + \log(n)(|M| + 1), \tag{3.30}$$

*where $l(\hat{\boldsymbol{\beta}}_M, \hat{\sigma}^2)$ again is the maximum value of the log-likelihood as given for the AIC.*

*Multiplied by 1/2, it is known as Schwartz criterion:*

$$BIC_S = \frac{1}{2} BIC \tag{3.31}$$

Assuming an Gaussian error, we obtain analogously for BIC

$$BIC = n \cdot \log(\hat{\sigma}^2) + \log(n)(|M| + 1).$$

We recognize a similar form of AIC and BIC, that is the criteria select models which gives the highest likelihood. But the main difference is that the BIC is penalizing a complex model much more than the AIC (for models with more than eight observations). The penalization in AIC and BIC is set in order to avoid over-fitting. In both cases, smaller values indicate a better model fit, i.e. the model which minimizes the information criterion.

For a deeper understanding, we refer to (Akaike, 1974 and Schwarz, 1978).

**Example 20** *(Model choice with AIC). We illustrate the approaches for model choice using the model (3.5), where our goal is to model the relationship between the transformed appliances energy use, with the variable* $\log(app)$*, and the two explanatory variables, i.e. temperature in the kitchen* `T1.kitchen` *and temperature in the living room* `T2.living`*. We already produced scatter-plots, and box-plots to see the relationship between the response variable and covariates. From these plots we assume an approximately linear increasing, but linear weak effect, for both covariates. With rising temperature in living room and kitchen we have a slightly rising appliances energy use, so there seems to be a relationship which makes it arguable. Based on these variables, we want to test possible model combinations. We examine the simple linear regression, our multiple linear model (3.5), and the model with added interaction term (3.8).*

*Using AIC criterion, we obtain the resulting AIC values for the considered models that are summarized in Table 3.3. The last model, Model D with the added interaction term, is the preliminary best model.*

| Model | Equation | | DF | AIC |
|---|---|---|---|---|
| A: | $\widehat{\log(\texttt{app}_i)} = \hat{\beta}_0$ | $+\hat{\beta}_1 \texttt{T1.kitchen}_i$ | 3 | 38871.98 |
| B: | $\widehat{\log(\texttt{app}_i)} = \hat{\beta}_0$ | $+\hat{\beta}_1 \texttt{T2.kitchen}_i$ | 3 | 38456.79 |
| C: | $\widehat{\log(\texttt{app}_i)} = \hat{\beta}_0$ | $+\hat{\beta}_1 \texttt{T1.kitchen}_i + \hat{\beta}_2 \texttt{T2.living}_i$ | 4 | 38433.94 |
| D: | $\widehat{\log(\texttt{app}_i)} = \hat{\beta}_0$ | $+\hat{\beta}_1 \texttt{T1.kitchen}_i + \hat{\beta}_2 \texttt{T2.living}_i$ $+\hat{\beta}_3 \texttt{T1.kitchen}_i \times \texttt{T2.living}_i$ | 5 | 38345.81 |

Table 3.3: Model selection with criterion AIC.

Note that we are not using BIC. As mentioned above, the only difference between AIC and BIC is the way they penalize, i.e. $2(|M|+1)$ in AIC and $\log(n)(|M|+1)$ in BIC. Thus the BIC produces higher penalty if the sample size $n$ is higher which might cause an under-fitting, whereas AIC might over-fit. There are many researches on the choice between AIC and BIC. The preference depends on the researcher and the goal. While AIC is best suited for prediction as it aims to find the best approximating model, the method BIC is best for explanation as it allows consistent estimation of the underlying data generating process. For more on the differences, see also the paper of Shmueli et al. (2010). In overview, we can say that in many papers, such as those to which we refer for our purposes in this thesis, or programs, the method AIC is used. In addition, our second statistical model chooses based on AIC model selection and therefore we are also using AIC to ensure consistency.

Either way, BIC leads to the same ranks and choices, but with higher values caused by the penalty term depending on the sample size.

## 3.6 Residual analysis

Finally we examine the statistical properties of the random residuals $\hat{\varepsilon}_i = Y_i - \mathbf{X}_i{}'\hat{\boldsymbol{\beta}}$. Substituting the regression parameter estimator, the expression is as follows:

$$\hat{\boldsymbol{\varepsilon}} = \mathbf{Y} - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} \tag{3.32}$$

Thus we get the following statistical properties of the random residuals:

$$\begin{aligned} &E[\hat{\boldsymbol{\varepsilon}}] = E[\mathbf{Y}] - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T E[\mathbf{Y}] = \mathbf{X}\boldsymbol{\beta} - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} = 0, \\ &\mathrm{Cov}[\hat{\boldsymbol{\varepsilon}}] = \sigma^2(\mathbf{I}_n - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T) = \sigma^2(\mathbf{I}_n - \mathbf{H}). \end{aligned}$$

Note that in comparison to the error term, the residuals also have the mean value zero, but the residuals are not uncorrelated and have heteroscedastic variances, which is obvious from taking a look at the $i$-th diagonal element of the covariance matrix.

Further from the assumption of a normally distributed error, we then have distribution of residuals

$$\hat{\boldsymbol{\varepsilon}} \sim N_n(0, \sigma^2(\mathbf{I}_n - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T)) = N_n(0, \sigma^2(\mathbf{I}_n - \mathbf{H})). \tag{3.33}$$

So consequently the distribution of residual sum of squares is

$$\frac{\hat{\boldsymbol{\varepsilon}}^T\hat{\boldsymbol{\varepsilon}}}{\sigma^2} = (n-p)\frac{\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-(p+1)}. \tag{3.34}$$

Moreover the residual sum of squares and the least squares estimator are independent. For the proof, we refer to Fahrmeir et al. (2013, Chapter 3, page 171)

The last statements are important for the hypothesis tests regarding to regression coefficients.

### 3.6.1 Standardized and studentized residuals

As we have seen, the results of the statistical properties of residuals saying that residuals are neither homoscedastic nor uncorrelated. But when we analyzing data, the residuals are used to validate model assumptions in a linear model.

Since the correlation are neglectable, we are dealing with the heteroscedasticity problem in the next steps. An obvious solution is standardization by dividing through the estimated standard deviation of the residuals.

Then resulting standardized residuals is given as

$$r_i = \frac{\hat{\varepsilon}_i}{\hat{\sigma}\sqrt{1-h_{ii}}}, \tag{3.35}$$

where $\hat{\sigma}^2$ is the estimate of $\sigma^2$ as defined in (3.21) and $h_{ii}$ is the $i$-th diagonal element of the hat matrix.

So when we suppose now that our model assumptions are true, the standardized residuals are homoscedastic. With the help of standardized residuals we can analyze the variances and conclude whether the assumption of homoscedasticity is violated or not. In practice we plot the standardized residuals versus the predicted values $\hat{y}_i$.

**Example 21** *The residual plots in Figure 3.2 show dependencies. So we have to transform* $\mathbf{Y}$*, so that the variance homogeneity is fulfilled. From the distribution, the majority of the standardized residuals, 95 % should lie in the interval [-2,2]. More on distribution explained in the next steps.*

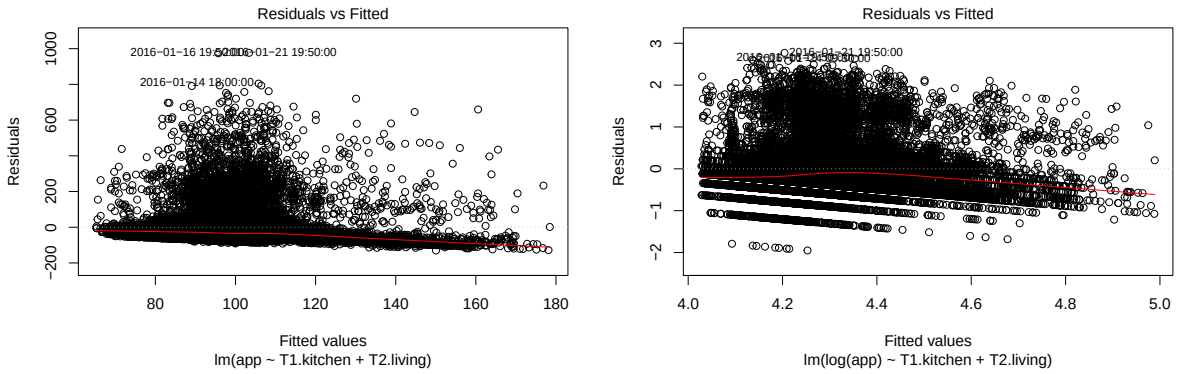



Figure 3.2: Residual plot. The left figure shows the regression (3.3) and the right figure shows the regression (3.5). The comparison reveals an improvement through the logarithmic transformation to $\log(\texttt{app}_i)$, $i = 1, \ldots, 19735$.

From the facts that the residuals are normally distributed and $(n-p)\frac{\hat{\sigma}^2}{\sigma^2}$ is $\chi^2_{n-p}$-distributed, the definition of the t-distribution (c.f. Definition 2.1) leads to the interest to assume a $t$-distributed standardized residuals.

The numerator and denominator in equation (3.35) are not stochastically independent as $\hat{\varepsilon}_i$ is a part of the expression of $\hat{\sigma}$. This means we can not assume a $t$-distributed standardized residual. To solve the problem of dependence, define "leave-one-out" estimators

which contains all observations except the $i$-th observation. Hence define the residuals based on these "leave-one-out" estimators to use the definition of the $t$-distribution and obtain the valid $t$-distributed studentized residuals $r_i^*$.

Define $\mathbf{X}_{(-i)}$ as the design matrix and $\mathbf{Y}_{(-i)}$ as the response vector with removed $i$-th row. The corresponding least squares estimator based on all observations except the $i$-th one equals

$$\hat{\boldsymbol{\beta}}_{(-i)} = (\mathbf{X}_{(-i)}^T \mathbf{X}_{(-i)})^{-1} \mathbf{X}_{(-i)}^T \mathbf{Y}_{(-i)}$$

Therefore we yield the predicted values

$$\hat{y}_{(-i)} = \mathbf{x}_i^T \hat{\boldsymbol{\beta}}_{(-i)}$$

and thus our residuals for the $i$-th observation

$$\hat{\varepsilon}_{(-i)} = y_{(i)} - \hat{y}_{(-i)} = y_{(-i)} - \mathbf{x}_i^T (\mathbf{X}_{(-i)}^T \mathbf{X}_{(-i)})^{-1} \mathbf{X}_{(-i)}^T \mathbf{y}_{(-i)}$$

is then distributed as follows

$$\hat{\epsilon}_{(-i)} \sim N(0, \sigma^2 (1 + \mathbf{x}_i^T (\mathbf{X}_{(-i)}^T \mathbf{X}_{(-i)})^{-1} \mathbf{x}_i))$$

or transformed to

$$\frac{\hat{\epsilon}_{(-i)}}{\sigma (1 + \mathbf{x}_i' (\mathbf{X}_{(-i)}' \mathbf{X}_{(-i)})^{-1} \mathbf{x}_i)^{1/2}} \sim N(0, 1).$$

From the distribution of the residual sum of squares (3.34) we yield

$$(n - p - 1) \frac{\hat{\sigma}_{(-i)}^2}{\sigma^2} \sim \chi_{n-(p+1)-1}^2,$$

where

$$\hat{\sigma}_{(-i)}^2 = \frac{1}{n-p-1} [(Y_1 - \mathbf{x}_1^T \hat{\boldsymbol{\beta}}_{(-i)})^2 + \cdots + (Y_{i-1} - \mathbf{x}_{i-1}^T \hat{\boldsymbol{\beta}}_{(-i)})^2 + (Y_{i+1} - \mathbf{x}_{i+1}^T \hat{\boldsymbol{\beta}}_{(-i)})^2 + \cdots + (Y_n - \mathbf{x}_n^T \hat{\boldsymbol{\beta}}_{(-i)})^2]$$

an estimator for $\sigma^2$ which is based on all observation except the $i$-th one.

Finally we have the studentized residuals

$$r_i^* = \frac{\hat{\epsilon}_{(-i)}}{\hat{\sigma}_{(-i)} (1 + \mathbf{x}_i^T (\mathbf{X}_{(-i)}^T \mathbf{X}_{(-i)})^{-1} \mathbf{x}_i)^{1/2}} \sim t_{n-(p+1)-1} \tag{3.36}$$

as now $\hat{\epsilon}_{(-i)}$ and $\hat{\sigma}_{(-i)}$ are independent which holds due to the above formulation which does not consider the $i$-th observation of $Y_i$ in the calculations.

The studentized residuals helps to confirm the model assumption and to detect outliers.

**Example 22** *The first plots in Figure 3.2 are the classical residual plots to see whether the residual points are randomly spread around the zero line, which means they do not have a certain pattern, that is our goal.*

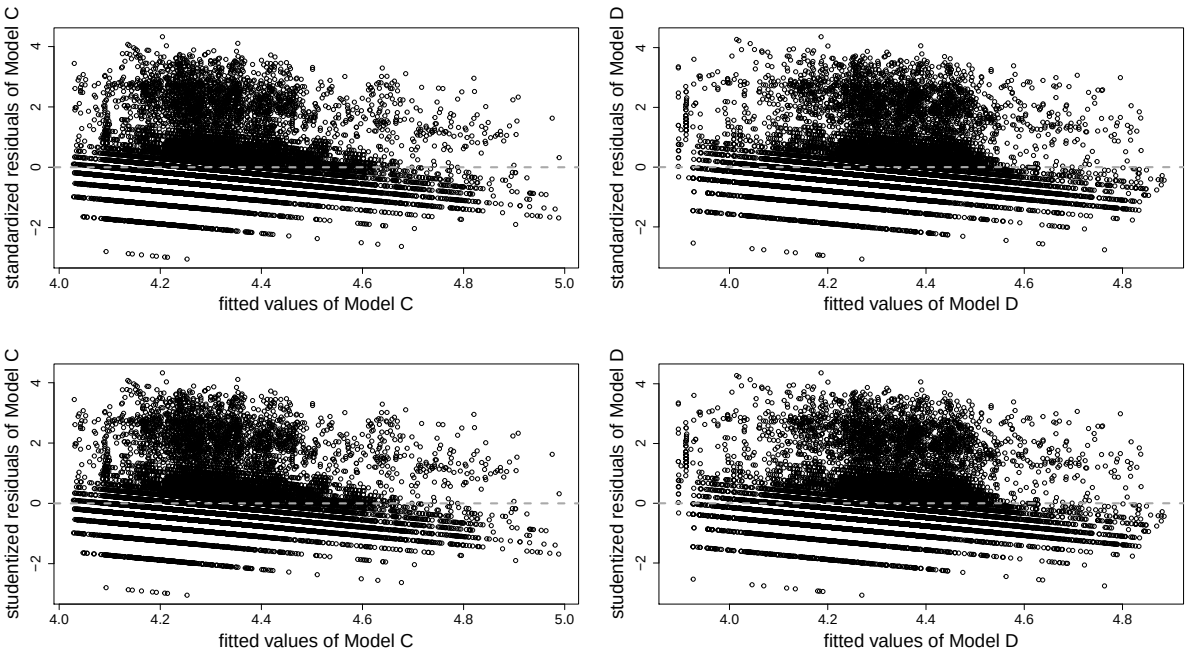



Figure 3.3: Standardized and studentized residuals versus fitted values of Model C on the left side and Model D on the right side. The models are as we defined in Table 3.3 of Section 3.5.3.

*Next, we can calculate the standardized, studentized residuals for each data point, as we defined in (3.35) and (3.36) respectively. To compare the results, we display them in Figure 3.3.*

*As we can see, we can not identify large differences, but we have a lighter amount of data for the studentized residual plot.*

### 3.6.2 Stationary models and autocorrelation function

Let $(X_i)_{i\in N}$ be the multivariate time series process with $X_i = ((X_{i1}, \dots X_{ip})^T)_{i\in N}$, and the corresponding observations $x_{i_1 j}, \dots, x_{i_n j}$ for $j = 1, \dots, p$. N denotes the index set, for example here $N = \mathbb{N}$, or $N = \mathbb{R}, [0, 1]$.

**Definition 3.13** *(Stationarity). A stochastic process $(X_i)_{i\in N}$ is called strictly (or strongly) stationary, if*

$$(X_{i_1}, \dots, X_{i_n}) \overset{d}{=} (X_{i_1+h}, \dots, X_{i_n+h}) \quad \text{for all } i_1, \dots, i_n \in N, n \in \mathbb{N},$$

*and $h$ such that $i_1 + h, \dots, i_n + h \in N$.*

*In particular, $X_i$ is identically distributed for all $i \in N$.*

**Definition 3.14** *(Autocovariance function). Let $(X_i)_{i\in N}$ be a time series with $Var(X_i) < \infty$ for all $i \in N$. Then*

$$\gamma_X(r, s) = Cov(X_r, X_s) = E\left[(X_r - E[X_r])(X_s - E[X_s])\right], \quad r, s \in N$$

*is called autocovariance function of* $(X_i)_{i\in N}$.

**Definition 3.15** $(X_i)_{i\in N}$ *is (weakly) stationary, if*

*(i)* $E[X_i]$ *is independent of* $i$,

*(ii)* $E[|X_i|^2] < \infty \ \forall i \in N$ *and*

*(iii)* $\gamma_X(i+h,i)$ *is independent of* $i$ *for each* $h$.

This means that for a weakly stationary process $(X_i)_{i\in N}$, and set $r = i+h$ and $s = i$, we have

$$\begin{array}{ll} & \gamma_X(i+h,i) = \gamma_X(i+h-i,0) \qquad \forall i,h : i, i+h \in N \\ \Leftrightarrow & \gamma_X(h) := \gamma_X(h,0) = Cov(X_{i+h}, X_i) \quad \forall i,h \in N. \end{array} \tag{3.37}$$

So the autocovariance function (acvf) $\gamma_X(h)$ is the covariance between stochastic process $(X_i)_{i\in N}$ ( later with observations) at a distance $h$, or precisely with lag $h$.

**Definition 3.16** $(X_i)_{i\in N}$ *is called a Gauss process, if all finite-dimensional distributions of* $(X_i)_{i\in N}$ *are multivariate normal. Further, if* $(X_i)_{i\in N}$ *is weakly stationary, with* $E[|X_i <$ $|^2] < \infty$, *then* $(X_i)_{i\in N}$ *is also strictly stationary, with* $Var(X_i) < \infty$.

**Definition 3.17** *(Autocorrelation function). Again, let* $(X_i)_{i\in N}$ *be a stationary time series. Using the autocovariance function* $\gamma_X(h)$ *of* $(X_i)_{i\in N}$ *at lag* $h$ *defined in Equation (3.37), we obtain the autocorrelation function (acf) of* $(X_i)_{i\in N}$ *at lag* $h$ *as*

$$\rho_X^{(N)} = \frac{\gamma_X(h)}{\gamma_X(0)} = Corr(X_{i+h}, X_i) \tag{3.38}$$

The corresponding estimate of the autocorrelation function (3.38) is defined as follows. Let $x_{1j}, \ldots, x_{nj}$ be observations of a time series for $j = 1, \ldots, p$. Recall the sample mean of our observations is given as $\bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}$ for $j = 1, \ldots, p$.

The sample autocovariance function is

$$\hat{\gamma}_j(h) = \frac{1}{n}\sum_{i=1}^{n-|h|}(x_{i+|h|,j} - \bar{x}_j)(x_{ij} - \bar{x}_j), \quad -n < h < n.$$

And thus the sample autocorrelation function is

$$\hat{\rho}_j(h) = \frac{\hat{\gamma}_j(h)}{\hat{\gamma}_j(0)}, \quad -n < h < n.$$

**Example 23** *For simplicity, let the stochastic process be one-dimensional. Let's take our response variable* **app** *as an example. We have the sequence* $((\mathbf{app}_i)_{i=1,\ldots,19735})$, *that is the observed appliances energy consumption over a time period. The Figure 3.4 shows the corresponding sample autocorrelation function at lag 1, . . . , 25000. We have the highest spike at lag 0, i.e.* $\hat{\rho}(0) = 1$, *since it is the correlation with the same time point here. The horizontal dashed lines on the graph are the bounds* $\pm 1.96/\sqrt{n}$ *with* $n = 19735$. *These dashed lines represents the significance level. In our Figure 3.4 all the spikes rises above the level which means that the time points are correlated with each other dependent of the lag (x-axes). After lag 25000 the spikes fall below the significance level and therefore the current appliances use has no effect on the appliances use at lag 25000.*

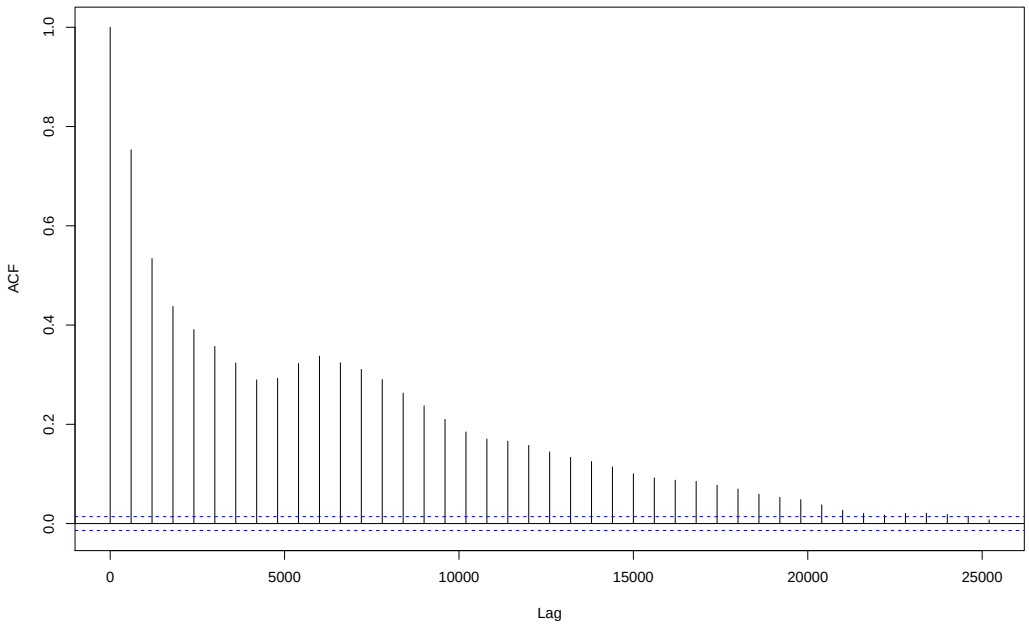



Figure 3.4: The sample autocorrelation function $\hat{\rho}(h)$ for the energy consumption.

For an extension and more theory on time series, we refer to Brockwell et al. (1991) and for more examples see Brockwell and Davis (2016).

## 3.7 Statistical inference

### 3.7.1 F-test

In this paragraph we describe statistical tests for hypothesis regarding to the unknown regression parameter $\boldsymbol{\beta}$. For these parameters we are able to construct confidence intervals. Besides of the duality between the statistical tests and confidence intervals which holds, the assumption of an independent and identically normally distributed error, i.e. $\varepsilon_i \sim N(0, \sigma^2)$, is required. Even for large sample size with non-normal errors, the tests and confidence intervals stay valid what we will see in the following.

There are several statistical hypothesis of interest for linear models.

**Example 24** *(i) The tests of significance, i.e. whether a variable should be included or not*

$$H_0 : \beta_j = 0 \text{ versus } H_1 : \beta_j \neq 0, \text{ with } j = 1, \ldots, p$$

*(ii) The composite test of a subvector, i.e. test to sparse and avoid over-parameterization in the model*

$$H_0 : \boldsymbol{\beta_l} = \mathbf{0} \text{ versus } H_1 : \boldsymbol{\beta_l} \neq \mathbf{0}, \text{ with subvector } \boldsymbol{\beta_l} = (\beta_1, \ldots, \beta_k) \text{ and } l = 1, \ldots, p$$

*(iii) The test of equality, i.e. test the necessary of differentiate and specify or subsetting some variables*

$$H_0 : \beta_j - \beta_k = 0 \text{ versus } H_1 : \beta_j - \beta_k \neq 0, \text{ with } j, k = 1, \ldots, p \text{ and } j \neq k$$

The generalized testing problems is called general linear hypothesis

$$H_0 : \mathbf{C}\boldsymbol{\beta} = \mathbf{d} \quad \text{versus} \quad H_1 : \mathbf{C}\boldsymbol{\beta} \neq \mathbf{d}, \tag{3.39}$$

where $\mathbf{C} \in \mathbb{R}^{k\times(p+1)}$ with $rank(\mathbf{C}) = k \leq p+1$, which means $k$ represents the number of linear independent restriction, and $\boldsymbol{d} \in \mathbb{R}^k$.

Assuming Gaussian error to generate a test statistic for the general test problem (3.39). Then calculate

(i) residual sum of squares $\text{SSE} = \hat{\boldsymbol{\varepsilon}}^T\hat{\boldsymbol{\varepsilon}}$ for the full model

(ii) residual sum of squares $\text{SSE}_{H_0} = \hat{\boldsymbol{\varepsilon}}_{H_0}^T\hat{\boldsymbol{\varepsilon}}_{H_0}$ for the model under the null hypothesis $\mathbf{C}\beta = \mathbf{d}$

(iii) statistic $\frac{\Delta\text{SSE}}{\text{SSE}} = \frac{\text{SSE}_{H_0}-\text{SSE}}{\text{SSE}}$, which is the relative distance of the residual sum of squares between the restricted and full model.

The idea of proving $\text{SSE}_{H_0} - \text{SSE} \geq 0$ and $\Delta\text{SSE} \geq 0$, is that the smaller the distance the more likely is a that we will not reject the null hypothesis. For more explanation, see Fahrmeir et al. (2013), Section 3.3.1, p.129.

Now to derive the test statistic and the distribution of it under $H_0$ of (3.39), we use our results from the distribution of the estimator parameter $\hat{\boldsymbol{\beta}}$ in (3.24).

Assume that $H_0$ is true, then we have

$$\mathbf{C}\boldsymbol{\beta} - \mathbf{d} \overset{H_0}{\sim} N(0, \sigma^2\mathbf{C}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{C}^T).$$

Since the regression coefficient $\hat{\boldsymbol{\beta}}$ and residuals are independent, we are applying the definition of $\chi^2$-distribution to obtain

$$\frac{(\mathbf{C}\boldsymbol{\beta} - \mathbf{d})^T(\mathbf{C}\boldsymbol{\beta} - \mathbf{d})}{\sigma^2(\mathbf{C}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}')} \overset{H_0}{\sim} \chi^2_k. \tag{3.40}$$

We recognize the sum of squares error under the restriction of $\mathbf{C}\boldsymbol{\beta} - \mathbf{d}$ in the latest expression. So setting the SSE with the least squares estimator $\hat{\boldsymbol{\beta}}_{H_0}$ under $H_0$ in matrix-vector notation

$$\text{SSE}_{H_0} := (\mathbf{C}\boldsymbol{\beta} - \mathbf{d})^T(\mathbf{C}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{C}^T)^{-1}(\mathbf{C}\boldsymbol{\beta} - \mathbf{d}).$$

Finally with yield the desired $F$-distribution (c.f. Definition 2.2) with the following Theorem.

**Theorem 3.18** *Assuming the normality assumption. From the stochastic properties and distribution of residual sum of squares (3.34) we derive and obtain*

$$\text{SSE}/\sigma^2 \sim \chi^2_{n-(p+1)}$$

*and with equation (3.40)*

$$\text{SSE}_{H_0}/\sigma^2 \sim \chi^2_k$$

A proof of the theorem can be found in Fahrmeir et al. (2013, Section 3.5.2).

To transform the statistic to a test statistic under $H_0$ with the distribution $F$, we add the constant factors $\frac{1}{k}$ as the numerator and $\frac{1}{n-p}$ as the denominator degree of freedom (df).

The test statistic under the null hypothesis is called the *F-test* and is defined as

$$F = \frac{\text{SSE}_{H_0}/k}{\text{SSE}/(n-(p+1))} \overset{H_0}{\sim} F_{k,n-(p+1)} \tag{3.41}$$

From this we obtain the following test. Let $\alpha$ be the significance level. Then the null hypothesis can be rejected if the test statistic is larger than the $(1-\alpha)$-quantile of the corresponding F-distribution

$$F > F_{k,n-(p+1)}(1-\alpha)$$

**Example 25** *(F-Tests for specific test problems).*

*(i)* *Test of significance (t-test):*

$H_0 : \beta_j = 0$ *versus* $H_1 : \beta_j \neq 0$, *with* $j = 0, \ldots, p$. *In this case we obtain the F-test from (3.41) :*

$$F_j = \frac{(\hat{\beta}_j^2 - 0)/((\mathbf{X}^T\mathbf{X})^{-1})_{jj}}{\text{SSE}/(n-(p+1))} \overset{H_0}{\sim} F_{1,n-(p+1)}$$

*with the rejection rule for our test statistic* $F_j > F_{1,n-(p+1)}(1-\alpha)$ *with the* $(1-\alpha)$*-quantile of the corresponding F-distribution.*

*Equivalently, with* $F_j = t_j^2$, *the F-test with just one regression coefficient can be based on the t-test, that is compare to studentized residual,*

$$t_j = \frac{\hat{\beta}_j}{se(\hat{\beta}_j)} \sim t_{n-(p+1)},$$

*where* $se(\hat{\beta}_j) = \sqrt{\hat{Var}(\hat{\beta}_j)}$ *is the estimated standard deviation or standard error of* $\hat{\beta}_j$, *compare to the unbiased estimator* $\hat{\sigma}^2_{REML}$ *in (3.24). Note that* $t_j$ *is t-distributed with* $(n-(p+1))$ *degree of freedom.*

*Thus the critical value for the rejection region of* $H_0$ *is deduced from the* $(1-\alpha/2)$*-quantile of the t-distribution with (n-(p+1)) degree of freedom. Which leads us to the decision rule that*

*we reject the null hypothesis, if* $|t_j| > t_{1-\alpha/2}(n-(p+1))$

*This test can be performed to select the covariates stepwise. Either the forward selection by adding covariates to the model based on F-test, or backward section by removing covariates till we have significant covariates.*

*(ii)* *Test for significance of regression: Test if there is a linear relationship between the response and any of the covariates. The null hypothesis is then*

$H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$. *Under the* $H_0$ *the least squares estimator can be given by* $\hat{\beta}_0 = \bar{Y}$, *so we obtain the F-test statistic*

$$\begin{aligned} F &= \frac{n-(p+1)}{p} \frac{\sum(\hat{y}_i - \bar{y}_i)^2}{\sum \hat{\varepsilon}_i^2} \\ &\overset{*}{=} \frac{n-(p+1)}{p} \frac{\sum(\hat{y}_i - \bar{y})^2 / \sum(y_i - \bar{y})^2}{1 - \sum(\hat{y}_i - \bar{y})^2 / \sum(y_i - \bar{y})^2} \quad \sim F_{p,n-(p+1)} \\ &\overset{*}{=} \frac{n-(p+1)}{p} \frac{R^2}{1-R^2} \end{aligned}$$

*with the decision rule:*

> *Reject* $H_0$, *if we obtain for our test statistic* $F$ *that* $F > F_{p,n-(p+1)}(1-\alpha)$
> *with the* $(1-\alpha)$*-quantile of the corresponding* $F$*-distribution.*

*In* $*$ *we used the composition formula (3.25) and the Definition 3.10 defining* $R^2$. *Use for interpretation that* $R^2_{adj} = 1 - (1-R^2)\frac{n-1}{n-(p+1)}$ *or equivalently* $R^2_{adj} = R^2 - (1-R^2)\frac{p}{n-(p+1)}$.
*So there is an connection between the F-test and the coefficient of determination. If* $R^2$ *is small enough, we expect the null hypothesis, i.e. no linear relationship, will not be rejected as F is also small. Otherwise F becomes comparably large when* $R^2$ *is close to its limit one.*

## 3.7.2 Confidence regions and prediction intervals

### Confidence intervals for regression coefficients

In order to construct a confidence interval for a $\beta_j$ under normality, we utilize the $t$-statistic $t_j = \frac{\hat{\beta}_j - d_j}{se_j}$ corresponding to the test $H_0 : \beta_j = d_j$. Remember that we reject the null hypothesis when $|t_j| > t_{n-(p+1)}(1-\alpha/2)$. So the probability of rejecting $H_0$ when $H_0$ is true equals $\alpha$. Hence, under $H_0$ we have

$$P(|t_j| > t_{n-(p+1)}(1-\alpha/2)) = P(|\frac{\hat{\beta}_j - \beta_j}{se_j}| > t_{n-(p+1)}(1-\alpha/2)) = \alpha.$$

On the other hand, the test is constructed such that the probability of not rejecting $H_0$, given $H_0$ is true, is provided by

$$\begin{aligned} & P(|t_j| < t_{n-(p+1)}(1-\alpha/2)) = 1-\alpha \\ \Leftrightarrow \quad & P(|\tfrac{\hat{\beta}_j - \beta_j}{se_j}| < t_{n-(p+1)}(1-\alpha/2)) = 1-\alpha \\ \Leftrightarrow \quad & P(\hat{\beta}_j - t_{n-(p+1)}(1-\alpha/2)se_j < \beta_j < \hat{\beta}_j + t_{n-(p+1)}(1-\alpha/2)se_j) = 1-\alpha \end{aligned}$$

and we yield the $(1-\alpha)$-confidence interval for $\beta_j$

$$\left[\hat{\beta}_j - t_{n-(p+1)}(1-\alpha/2)se_j, \hat{\beta}_j + t_{n-(p+1)}(1-\alpha/2)se_j\right]$$

Analogously we can create the $(1-\alpha)$-confidence region for a $k$-dimensional subvector $\boldsymbol{\beta}_l$ of $\boldsymbol{\beta}$.

**Prediction intervals**

Similarly we can construct the prediction intervals for a future observation. We already derived the prediction for a new observation, say $y_0$ with a given covariate vector $\mathbf{x}_0$, so the corresponding estimator of the expectation $E[Y_0] = \mu_0$ (application of Gauß-Markov-Theorem, see Fahrmeir et al. (2013) on p. 119 and its proof on p. 170) is given by

$$\hat{Y}_0 = \mathbf{x}_0^T \hat{\boldsymbol{\beta}}$$

Not only the point estimate, but also the interval estimation for $\mu_0$ is of interest. For the construction we use that $\hat{\boldsymbol{\beta}} \sim N_n(\boldsymbol{\beta}, \sigma^2(\mathbf{X}^T\mathbf{X})^{-1})$, and by the property of linear combinations and standardization it follows

$$\begin{array}{c} \mathbf{x}_0^T\hat{\boldsymbol{\beta}} \sim N(\mathbf{x}_0^T\boldsymbol{\beta}, \sigma^2\mathbf{x}_0^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_0) \\ \overset{standardizing}{\Leftrightarrow} \quad \frac{\mathbf{x}_0^T\hat{\boldsymbol{\beta}}-\mu_0}{\sigma(\mathbf{x}_0^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_0)^{1/2}} \sim N(0,1) \end{array}$$

Finally substituting $\sigma^2$ with the estimator $\hat{\sigma}^2$ (3.21) we obtain

$$P\left(-t_{n-(p+1)}(1-\alpha/2) \leq \frac{\mathbf{x}_0^T\hat{\boldsymbol{\beta}} - \mu_0}{\hat{\sigma}(\mathbf{x}_0^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_0)^{1/2}} \leq t_{n-(p+1)}(1-\alpha/2)\right) = 1-\alpha$$

which follows a $t$-distribution with $n-(p+1)$ degree of freedom.

This results in a $(1-\alpha)$-confidence interval for $\mu_0$.

For the purpose of getting the prediction interval which contains the random future observation $Y_0$ with a high probability, we look at the prediction error estimator $\hat{\varepsilon}_0 = Y_0 - \mathbf{x}_0^T\hat{\boldsymbol{\beta}}$ with

$$\hat{\varepsilon}_0 \sim N(0, \sigma^2 + \sigma^2\mathbf{x}_0^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_0).$$

Again, by replacing $\hat{\sigma}^2$ for $\sigma^2$ and standardizing we obtain

$$\frac{Y_0 - \mathbf{x}_0^T\hat{\boldsymbol{\beta}}}{\hat{\sigma}(1+\mathbf{x}_0^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_0)^{1/2}} \sim t_{n-(p+1)}$$

and thus

$$P\left(-t_{n-(p+1)}(1-\alpha/2) \leq \frac{Y_0 - \mathbf{x}_0^T\hat{\boldsymbol{\beta}}}{\hat{\sigma}(1+\mathbf{x}_0^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_0)^{1/2}} \leq t_{n-(p+1)}(1-\alpha/2)\right) = 1-\alpha$$

Eventually, with $(1-\alpha)$-confidence, we find the future observation at $\mathbf{x}_0$ within the prediction interval

$$\mathbf{x}_0^T\hat{\boldsymbol{\beta}} \pm t_{n-(p+1)}(1-\alpha/2)\hat{\sigma}(1+\mathbf{x}_0^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_0)^{1/2}.$$

By the construction of the prediction interval, it is always wider than the corresponding confidence interval for $\mu_0$. This is due to the addition of a possibly high error variance. The confidence interval is constructed for $E[Y_0] = \mu_0$, so that the random interval overlaps the fixed and constant mean $\mu_0$ with probability $1-\alpha$. On the other hand, the prediction interval with probability $1-\alpha$ contains the random future observation $Y_0$.

# Chapter 4

# Generalized additive model (GAM)

So far we introduced and described statistical models with a univariate response modeled as the sum of linear or transformed predictors depending linearly on the estimated parameters and a zero mean random error term. Furthermore, statistical inference is usually based on assumption, that is the response variable is normally distributed. This approach of multiple linear regression is the most widely used method in the analysis of designed experiments and for other modeling task such as polynomial regression.

At this point we finally come to a more flexible and automated method to fit a model. The generalized additive model (GAM) relaxes the strict linear assumptions. It allows not only the expected value of the response variable to depend linearly on smooth functions of predictor variables, but also allows any distribution from the exponential family. The exact parametric form of these functions are unknown as the degree of smoothness are appropriate for each of them.

In practice the usage of GAM requires some extension to the multiple linear model methods.

(i) smooth functions must be represented in a certain way

(ii) degree of smoothness of the function must be made controllable

$\Rightarrow$ models with varying degree of smoothness can be explored

(iii) means from data are required for selecting the most appropriate degree of smoothness

$\Rightarrow$ models are then useful for more than purely exploratory work

In general the model is structured like this

$$g(\mu_i) = \mathbf{A}_i\boldsymbol{\beta} + f_1(x_{i1}) + f_2(x_{i2}) + f_3(x_{i3}, x_{i4}) + \ldots, \qquad i = 1, \ldots, n, \tag{4.1}$$

with $\mu_i \equiv E[Y_i]$ and $Y_i \sim EF(\mu_i, \phi)$ denotes again the response variable with $EF(\mu_i, \phi)$ as an exponential family distribution with mean $\mu_i$ and scale parameter $\phi$. Moreover $\mathbf{A}_i$ is a row of the model matrix for any strictly parametric model components and $\boldsymbol{\beta}$ the

corresponding parameter vector, and finally $f_j$ are the smooth functions of the covariate $x_{ij}$ which are estimated non-parametrically.

In abstract, a generalized additive model allows to fit a non-linear function for each predictor rather than parametric relationships and therefore allows more flexible specification of the dependence of the response on the covariates. But this flexibility and convenience leads to two new theoretical issues. How to represent the smooth functions and how to choose the smoothness?

In the following we will discuss the foundations of GAM and the estimation of the model parameters. For a detailed insight to GAM, we refer to the book of Wood (2017).

## 4.1 Additive models - An introductory example

Suppose we have two independent variables, `T1.kitchen` and `T2.living` with $n = 19735$ data points, which are observed for the response variable `app`. For the response variable, we again use the logarithmic transformation $y_i := \log(\texttt{app}_i)$, for $i = 1, \ldots, 19735$, as we did for the multiple linear regression chapter after Example 10. We introduce our new model with a simple additive model structure,

$$Y_i = \alpha + f_1(\texttt{T1.kitchen}_i) + f_2(\texttt{T2.living}_i) + \varepsilon_i, \quad i = 1, \ldots, 19735, \tag{4.2}$$

with the intercept $\alpha$, the smoothing functions $f_j$ for $j = 1, 2$, and the error terms $\varepsilon_i$ as independent $N(0, \sigma^2)$ random variables.

If we the model contains only one function, we could just using methods covered in multiple linear models, that is representing $f$ such that it becomes a linear model and then just focusing on the single smoothing function. But the fact that the model contains more than one leads to an identifiability problem. $f_1$ and $f_2$ are only estimable inside of an additive constant. This means, any constant added to $f_1$ have to be simultaneously subtracted from $f_2$, so that the prediction of the model does not change. This implies that the identifiability constraints have to be established on the model before fitting the model.

With this identifiability constraints we can now proceed exactly as intended for the univariate model with just one smoothing function $f$, i.e. representing the model with the help of penalized regression splines, which estimated by penalized least squares, and with selected degree of smoothing by cross validation or restricted maximum likelihood (REML) approach. We will now explore these points step by step.

### 4.1.1 Penalized regression representation

The smooth functions in (4.2) can be described by using penalized piecewise linear basis

$$\begin{aligned} f_1(\texttt{T1.kitchen}) &= \textstyle\sum_{j=1}^{k_1} b_j(\texttt{T1.kitchen})\delta_j, \\ f_2(\texttt{T2.living}) &= \textstyle\sum_{j=1}^{k_2} c_j(\texttt{T2.living})\gamma_j, \end{aligned}$$

where $\delta_j$ and $\gamma_j$ are the corresponding unknown coefficients, $b_j(\texttt{T1.kitchen})$ and $c_j(\texttt{T2.living})$ are basis functions of the form

$$b_j(x) = \begin{cases} (x - x^\star_{j-1})/(x^\star_j - x^\star_{j-1}) & x^\star_{j-1} < x \le x^\star_j \\ (x^\star_{j+1} - x)/(x^\star_{j+1} - x^\star_j) & x^\star_j < x \le x^\star_{j+1} \\ 0 & \text{otherwise} \end{cases} \tag{4.3}$$

for $j = 2, \ldots, k-1$ and knots defined by $\{x^\star_j : j = 1, \ldots, k\}$ with $x^\star_j > x^\star_{j-1}$. This is defined for our two covariates using sequences $k_1$ and $k_2$ knots, $\texttt{T1.kitchen}^\star_j$ and $\texttt{T2.living}^\star_j$ evenly spaced over the whole range of `T1.kitchen` and `T2.living`, respectively.

Furthermore, define the $n$-dimensional vectors, in our example we have $n = 19735$, with

$$\boldsymbol{f}_1 = [f_1(\texttt{T1.kitchen}_1), \ldots, f_1(\texttt{T1.kitchen}_{19735})]^T,$$
$$\boldsymbol{f}_2 = [f_2(\texttt{T2.living}_1), \ldots, f_2(\texttt{T2.living}_{19735})]^T,$$

so that we have the matrix notation $\boldsymbol{f}_1 = \boldsymbol{X}_1\boldsymbol{\delta}$ and $\boldsymbol{f}_2 = \boldsymbol{X}_2\boldsymbol{\gamma}$. Note that $b_j(\texttt{T1.kitchen}_i)$ is the $i, j$-th element of $\boldsymbol{X}_1$, similarly for $\boldsymbol{X}_2$.

For each function there is a penalty. Let $x$ be a covariate which can be set to $x =$ `T1.kitchen` or $x =$ `T2.living`. To set this penalty we have to establish a term that measures the wiggliness, i.e. $\sum_{j=2}^{k-1}\{f(x^\star_{j-1}) - 2f(x^\star_j) + f(x^\star_{j+1})\}^2$, as this sum squared measures differences of the function at the knots. So if $f$ is wiggly, the term will take high values, and if $f$ is smooth, the term will be low.

Note that for the basis of tent function as in (4.3), the coefficients of $f$ are simply the function value at the knots, i.e. $\delta_j = f_1(\texttt{T1.kitchen}^\star_j)$ and $\gamma_j = f_2(\texttt{T2.living}^\star_j)$. Now we can straightforward formulate a penalty as a quadratic form using the term that measures the wiggliness, that is for $\delta_j = f_1(\texttt{T1.kitchen}^\star_j)$:

$$\sum_{j=2}^{k-1}(\delta_{j-1} - 2\delta_j + \delta_{j+1})^2 = \boldsymbol{\delta}^T\boldsymbol{D}^T\boldsymbol{D}\boldsymbol{\delta} = \boldsymbol{\delta}^T\bar{\boldsymbol{S}}\boldsymbol{\delta}$$

Here we define $\bar{\boldsymbol{S}} := \boldsymbol{D}^T\boldsymbol{D}$ with the $(k-2)\times k$ matrix $\boldsymbol{D} := \begin{pmatrix} 1 & -2 & 1 & 0 & \cdot & \cdot & \cdot \\ 0 & 1 & -2 & 1 & 0 & \cdot & \cdot \\ 0 & 0 & 1 & -2 & 1 & 0 & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \end{pmatrix}$.

Doing this analogously for $f_2(\texttt{T2.living}^\star_j)$, we obtain the following penalties of the form

$$\boldsymbol{\delta}^T\boldsymbol{D}_1^T\boldsymbol{D}_1\boldsymbol{\delta} = \boldsymbol{\delta}^T\bar{\boldsymbol{S}}_1\boldsymbol{\delta} \quad \text{for } f_1,$$
$$\boldsymbol{\gamma}^T\boldsymbol{D}_2^T\boldsymbol{D}_2\boldsymbol{\gamma} = \boldsymbol{\gamma}^T\bar{\boldsymbol{S}}_2\boldsymbol{\gamma} \quad \text{for } f_2.$$

At this point we will address the aforementioned identifiability problem. Purposes for estimation are some linear constraints, but to avoid wide confidence intervals, the best option is a sum-to-zero constraint

$$\sum_{i=1}^{19735} f_1(\texttt{T1.kitchen}_i) = 0, \quad \Leftrightarrow \quad \mathbf{1}^T\boldsymbol{f}_1 = 0,$$

with $\mathbf{1}$ is an $n = 19735$ vector of 1's. Note that this constraints allows $f_1$ to have the same shape with the same penalty value, since it shifts only vertically so that its mean value is zero.

For the application of the identifiability constraints, the following has to hold

$$\mathbf{1}^T \boldsymbol{X}_1 \boldsymbol{\delta} = 0 \ \forall \boldsymbol{\delta} \quad \Rightarrow \quad \mathbf{1}^T \boldsymbol{X}_1 = \mathbf{0}.$$

To yield the latter condition, we have to define a column centered matrix by subtracting the column mean from each column of $\boldsymbol{X}_1$, i.e.

$$\tilde{\boldsymbol{X}}_1 = \boldsymbol{X}_1 - \mathbf{1}\mathbf{1}^T \boldsymbol{X}_1 / n,$$

and set $\tilde{\boldsymbol{f}}_1 = \tilde{\boldsymbol{X}}_1 \boldsymbol{\delta}$.
Checking this constraints

$$\tilde{\boldsymbol{f}}_1 = \tilde{\boldsymbol{X}}_1 \boldsymbol{\delta} = \boldsymbol{X}_1 \boldsymbol{\delta} - \mathbf{1}\mathbf{1}^T \boldsymbol{X}_1 \boldsymbol{\delta} / n = \boldsymbol{X}_1 \boldsymbol{\delta} - \mathbf{1} c = \boldsymbol{f}_1 - c,$$

defining the scalar $c = \mathbf{1}^T \boldsymbol{X}_1 \boldsymbol{\delta} / n$. This shows that the constraint only shift in the level of $\boldsymbol{f}_1$. Additionally note that the column centered matrix $\tilde{\boldsymbol{X}}_1$ is reduced by one rank, i.e. only $k_1 - 1$ elements of the $k_1$ vector $\boldsymbol{\delta}$ can be uniquely estimated. A simple identifiability constraint deals with the problem that a single element of $\boldsymbol{\delta}$ is set zero and leads to removing the corresponding column of $\tilde{\boldsymbol{X}}_1$ and $\boldsymbol{D}$.

The column centered matrix in this rank reduced basis form automatically fulfills the identifiability constraint. Henceforward it is supposed that the matrices $\boldsymbol{X}_j$, $\boldsymbol{D}_j$, etc. are the constrained versions, so the tildes are dropped in the following.

Now that we have deployed the constrained bases for the $f_j$, we can represent (4.2) as a linear version

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where $\boldsymbol{X} = (\mathbf{1}, \boldsymbol{X}_1, \boldsymbol{X}_2)$ and $\boldsymbol{\beta} = (\alpha, \boldsymbol{\delta}^T, \boldsymbol{\gamma}^T)$.

In this notation we can easily express the penalties as quadratic forms

$$\boldsymbol{\beta}^T \boldsymbol{S}_1 \boldsymbol{\beta} = (\alpha, \boldsymbol{\delta}^T, \boldsymbol{\gamma}^T) \begin{pmatrix} 0 & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \bar{\boldsymbol{S}}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \alpha \\ \boldsymbol{\delta} \\ \boldsymbol{\gamma} \end{pmatrix} = \boldsymbol{\delta}^T \bar{\boldsymbol{S}}_1 \boldsymbol{\delta}.$$

### 4.1.2 Fitting model by penalized least squares

To determine the coefficient estimates $\hat{\boldsymbol{\beta}}$ of the model (4.2) is the next goal.

Analogously to the multiple linear case, we have to minimize the penalized least squares objective function

$$\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 + \lambda_1 \boldsymbol{\beta}^T \boldsymbol{S}_1 \boldsymbol{\beta} + \lambda_2 \boldsymbol{\beta}^T \boldsymbol{S}_2 \boldsymbol{\beta}, \tag{4.4}$$

where $\lambda_1$ and $\lambda_2$ are the smoothing parameter that control the weight to smooth $f_1$ and $f_2$, relative to the objective of tightly fitting the response data $\boldsymbol{y}$. Note that for this calculation, the weight has to be given to the objective function. Thus we assume at this point that these smoothing parameters are given.

Thus our coefficient estimates $\hat{\boldsymbol{\beta}}$ are given by

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X} + \lambda_1\boldsymbol{S}_1 + \lambda_2\boldsymbol{S}_2)^{-1}\boldsymbol{X}^T\boldsymbol{y}. \tag{4.5}$$

Compare the estimate with the linear model case (3.15) and note the penalization terms for the additive models.

Set the influence matrix as $\boldsymbol{A} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X} + \lambda_1\boldsymbol{S}_1 + \lambda_2\boldsymbol{S}_2)^{-1}\boldsymbol{X}^T$, which is comparable to the hat matrix (3.23) in the linear regression case.

To give more computational stability, re-write the objective function (4.4) as

$$\left\| \begin{pmatrix} \boldsymbol{y} \\ \boldsymbol{0} \end{pmatrix} - \begin{pmatrix} \boldsymbol{X} \\ \boldsymbol{B} \end{pmatrix} \boldsymbol{\beta} \right\|^2,$$

where we have $\boldsymbol{B} = \begin{pmatrix} \boldsymbol{0} & \sqrt{\lambda_1}\boldsymbol{D}_1 & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \sqrt{\lambda_2}\boldsymbol{D}_2 \end{pmatrix}$ such that the matrix satisfies $\boldsymbol{B}^T\boldsymbol{B} = \lambda_1\boldsymbol{S}_1 + \lambda_2\boldsymbol{S}_2$.

Note that in a single smooth case, we have simply the un-penalized least squares objective function for an augmented model version and corresponding response data and thus the model can be fitted by linear regression with the help of stable orthogonal matrix based method.

### 4.1.3 Choosing smoothing parameter and setting distributions

At this point we give hints on the choice of $\lambda$. Examine the penalized least squares objective function (4.4), it is obvious that the smoothing parameter $\lambda$ controls the trade-off between smoothness of the estimated $f$ and precision to the data.

- $\lambda \to \infty$: This choice will lead to a straight line estimate for $f$.
- $\lambda \to 0$: This choice will result in an un-penalized piecewise linear regression estimate.

For choosing the smoothing parameter $\lambda$, there are some methods, the cross validation and the REML, the Bayesian model approach. The cross validation method will be described in the next GAM theory section. As for Bayesian model approach, we operate on beliefs for rather smooth than wiggly true models and formulate prior beliefs, for example that a simple choice of a prior distribution on function wiggliness is an exponential prior $\propto exp(-\boldsymbol{\lambda}\boldsymbol{\beta}^T\boldsymbol{S}\boldsymbol{\beta}/\sigma^2)$, which is equivalent to an improper multivariate normal prior $\boldsymbol{\beta} \sim N(\boldsymbol{0}, \sigma^2\boldsymbol{S}^-/\boldsymbol{\lambda})$ with pseudo-inverse $\boldsymbol{S}^-$, i.e. for an eigen-decomposition $\boldsymbol{S} = \boldsymbol{U}\boldsymbol{E}\boldsymbol{U}^T$ with $\boldsymbol{E}^-$ denoting a diagonal matrix of the inverse of non-zero eigenvalues and with zeros in place for zero eigenvalues of corresponding matrix $\boldsymbol{E}$, then the pseudo-inverse is $\boldsymbol{S}^- = \boldsymbol{U}\boldsymbol{E}^-\boldsymbol{U}^T$.

Now, the quantity we introduce is an estimate of the Bayesian covariance matrix for the model coefficients

$$\hat{\boldsymbol{V}}_\beta = (\boldsymbol{X}^T\boldsymbol{X} + \lambda_1\boldsymbol{S}_1 + \lambda_2\boldsymbol{S}_2)^{-1}\hat{\sigma}^2,$$

with $\hat{\sigma}^2$ is the residual sum of squares for the fitted model divided by the effective residual degree of freedom, i.e. $\hat{\sigma}^2 = \frac{\|\boldsymbol{y}-\boldsymbol{A}\boldsymbol{y}\|^2}{tr((\boldsymbol{I}-\boldsymbol{A})^T(\boldsymbol{I}-\boldsymbol{A}))}$. In Section 4.2.2 we will give more definitions and explanations.

By assuming the above prior, the penalized least squares objective function (4.4) and the coefficient estimate (4.5), the posterior distribution for $\boldsymbol{\beta}$ is

$$\boldsymbol{\beta}|\boldsymbol{Y} \sim N(\hat{\boldsymbol{\beta}}, \boldsymbol{V}_\beta),$$

this result is used for later inferences about $\boldsymbol{\beta}$ in the following GAM theory part about posterior distribution (Section 4.4).

We also have the possibility of estimating $\sigma^2$ and $\lambda$ using marginal likelihood maximization or REML. These points will be discussed next in the GAM theory section.

## 4.2 Theory of generalized additive models

### 4.2.1 Model setting

A standard form of a *generalized additive model* is

$$g(\mu_i) = \mathbf{A}_i\boldsymbol{\gamma} + \sum_j f_j(x_{ij}), \quad Y_i \sim EF(\mu_i, \phi), \quad i = 1, \ldots, n, \tag{4.6}$$

where $\mathbf{A}_i$ is the $i$-th row of the parametric model matrix with corresponding parameters $\boldsymbol{\gamma}$, $f_j$ is a smooth function of covariate $\mathbf{x}_j$, and $EF(\mu_i, \phi)$ denotes an exponential family distribution with mean $\mu_i$ and scale parameter $\phi$. For given $\mu_i$ the $Y_i$ are modeled independently.

For every form of the GAM model, we need to choose smoothing bases and penalties for each $f_j$, implying model matrices $\mathbf{X}^{[j]}$ and penalties $\mathbf{S}^{[j]}$. Consider our basic model, if $b_{kj}$ is the $k$-th basis function for $f_j$, then $X^{[j]}_{ik} = b_{kj}(x_{ij})$. Comparable to hat matrix of the multiple linear model (3.23) and to avoid the smooth terms confound with the intercept, included in $\mathbf{A}$, any smooth $\mathbf{X}^{[j]}$ contains an $\mathbf{1}$ in the span. Therefore we need identifiability constraints.

The identifiability constraints of the form

$$\sum_i f_j(x_{ij}) = 0$$

are re-parameterized and absorbed into the basis in a suitable way.

(An example of the only exception is the Gaussian random effects, where we have a null space so that $f_j \to 0$. For more, see Wood (2017), Section 6.5.)

But in practice, the application of the constraints to all the smooths in the basic forms of the model is the usual procedure.

For a detailed and summarized explanation, please see Wood (2017), Section 5.4.1.

After re-parameterization, we denote $\boldsymbol{\mathcal{X}}^{[j]}$ and $\boldsymbol{\mathcal{S}}^{[j]}$ as the model matrix and penalty matrix for $f_j$, respectively. Then combine column-wise $\mathbf{A}$ and $\boldsymbol{\mathcal{X}}^{[j]}$ to yield the whole

model matrix

$$\boldsymbol{X} = (\mathbf{A}, \boldsymbol{\mathcal{X}}^{[1]}, \boldsymbol{\mathcal{X}}^{[2]}, \dots).$$

With the corresponding model coefficient vector $\boldsymbol{\beta}$, which contains $\boldsymbol{\gamma}$ and the individual smooth term coefficient vectors, we yield our *over-parameterized (generalized) linear model*

$$g(\mu_i) = \mathbf{X}_i\boldsymbol{\beta}, \qquad Y_i \sim EF(\mu_i, \phi). \tag{4.7}$$

Now we can write a total smoothing penalty for the model

$$\sum_j \lambda_j \boldsymbol{\beta}^T \mathbf{S}^{[j]} \boldsymbol{\beta}, \tag{4.8}$$

where $\lambda_j$ is a smoothing parameter and $\mathbf{S}^{[j]}$ is $\boldsymbol{\mathcal{S}}^{[j]}$ embedded as a diagonal block in a matrix with zero entries otherwise. So the penalty for the smooth $f_j$ has the form $\lambda_j \boldsymbol{\beta}^T \mathbf{S}^{[j]} \boldsymbol{\beta}$.

In case that we have $f_j$ to be a tensor product or an adaptive smoother, there are more than one $\mathbf{S}^{[j]}$ for each $f_j$. We will deal with this case later.

To control the model smoothness of the over-parameterized (generalized) linear model (4.7), we have to estimate the fit by maximization of the *penalized log likelihood*

$$l_p(\boldsymbol{\beta}) = l(\boldsymbol{\beta}) - \frac{1}{2\phi} \sum_j \lambda_j \boldsymbol{\beta}^T \mathbf{S}^{[j]} \boldsymbol{\beta}, \tag{4.9}$$

where $l(\boldsymbol{\beta})$ is the log likelihood of the (generalized) linear model. We can see that $\lambda_j$ controls the balance between the goodness of fit of the model and the model smoothness.

**Prior distribution** Therefore during the fitting, the smoothing process employs the smoothing penalty $\boldsymbol{\beta}^T \boldsymbol{S} \boldsymbol{\beta}$. This is due to the belief that the true function is more likely to be smooth than wiggly. Defining a prior distribution on function wiggliness $f(\boldsymbol{\beta}) \propto \exp(-\lambda \boldsymbol{\beta}^T \boldsymbol{S} \boldsymbol{\beta}/2)$, based on the Bayesian manner, we can conclude that the improper Gaussian prior is distributed as follows

$$\boldsymbol{\beta} \sim N(\mathbf{0}, \boldsymbol{S}_\lambda^- \phi), \tag{4.10}$$

where $\boldsymbol{S}^-$ is a pseudoinverse of $\boldsymbol{S}$.

So given the smoothing basis, the smooth model (4.6) and its transformation to (4.7), with $\varepsilon_i \sim N(0, \sigma^2)$, the distribution can be expressed as $\boldsymbol{Y} \sim N_n(\boldsymbol{X}\boldsymbol{\beta}, \boldsymbol{I}\sigma^2)$.

To get a more comprehensive explanation, we refer to Wood (2017), Section 5.8.

### 4.2.2 Smoother and parameter estimations

To formulate the model more precisely, we now look at the formally simplified GAM to concentrate at the smoothers, with $i = 1, \dots, n$, $g$ is a link function, which can be identical, logarithmic or inverse, $\boldsymbol{Y}$ is the response variable, $\boldsymbol{x}_1, \dots, \boldsymbol{x}_p$ are the $p$ independent variables, $\mathbf{A} = 1$ so $\beta_0$ is an intercept, $f_1, \dots, f_p$ are unknown smooth functions and $\boldsymbol{\varepsilon} \in \mathbb{R}^n$ is an i.i.d. random error.

$$\begin{aligned} g(E[Y_i]) &= \beta_0 + f_1(x_{i1}) + \cdots + f_p(x_{ip}) + \varepsilon_i, \\ Y_i &\sim \text{some exponential family distribution.} \end{aligned} \tag{4.11}$$

The smooth function $f$ is constructed by the sum of basis functions $b$ and their corresponding regression coefficients $\beta$. It can be formally written as

$$f(x) = \sum_{i=1}^{q} b_i(x)\beta_i,$$

where $q$ is basis dimension.

These smooth functions are also known as splines. Smoothing splines are real functions which are piecewise-defined by their basis functions, i.e. polynomial functions, and their connection is called knots.

**Example 26** *There are several kind of smoothing splines that exit. Some suitable smoothing bases $b$ are*

- *thin plate regression splines*
- *cubic regression spline*
- *cyclic cubic regression spline*
- *Penalized splines (P-splines)*

*The most common used spline is the cubic basis function. For example, we fix the dimension $q = 3$, i.e. the number of knots are 3. Thus we set the basis function as follows*

$$b_{cubic}(x) = \begin{cases} \frac{1}{4}(x+2)^3 & if\ -2 \le x \le -1, \\ \frac{1}{4}(3|x|^3 - 6x^2 + 4) & if\ -1 \le x \le +1, \\ \frac{1}{4}(2-x)^3 & if\ +1 \le x \le +2. \end{cases}$$

*To give further example, in practice, for daily seasonality cubic regression spline and for a lower frequency such as weekly seasonality P-splines are used. Both are knot-based.*

For the regularization of the spline smoothness, the penalized regression splines are deployed.

Thus, we have the model in our linear way as written in (4.7). Accordingly, the objective function with respect to the smooth function $f$ is

$$\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 + \lambda \int_0^1 [f''(x)]^2 dx,$$

with $\lambda$ be a smoothing parameter. To complete the circle, the integral of squares of second derivatives can be written as the introduced smoothing penalty in (4.8) for the model:

$$\int_0^1 [f''(x)]^2 dx = \boldsymbol{\beta}^T \boldsymbol{S} \boldsymbol{\beta},$$

where $\boldsymbol{S}$ is the matrix of known coefficients.

The penalized least squares objective function is therefore defined as

$$\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 + \sum_j \lambda_j \boldsymbol{\beta}^T \boldsymbol{S}^{[j]} \boldsymbol{\beta}. \tag{4.12}$$

Finally, the equations, mathematical calculations and conclusions imply that the regression coefficients can be obtained by the estimator

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X} + \boldsymbol{\lambda}\boldsymbol{S})^{-1}\boldsymbol{X}^T\boldsymbol{Y}. \tag{4.13}$$

Contrary to the estimator $\hat{\boldsymbol{\beta}}$ of the linear model in (3.16), there are added a penalization in the GAM approach. Therefore $\hat{\boldsymbol{\beta}}$ is called penalized least squares estimator.

Now some important theoretical questions are to be answered. In the next step, we introduce the method on how to obtain the estimate of $\boldsymbol{\beta}$. This method is called PIRLS, i.e. penalized iteratively re-weighted least squares. For this method we return to the general formally matrix representation (4.7) as formulated in the model setting.

### Estimation of $\boldsymbol{\beta}$ given $\lambda$ with method PIRLS

Consider objective (4.9) as an optimization problem like we had in the linear model maximum likelihood.

Maximize objective (4.9) through the following penalized iteratively re-weighted least squares iteration:

(i) Initialize $\hat{\mu}_i = y_i + \delta_i$ and $\hat{\eta}_i = g(\hat{\mu}_i)$.

$\delta_i$ is usually zero, but may also be a small constant to ensure a finite $\hat{\eta}_i$.

(ii) Compute pseudo data $z_i = g'(\hat{\mu}_i)(y_i - \hat{\mu}_i)/\alpha(\hat{\mu}_i) + \hat{\eta}_i$
and iterative weights $w_i = \alpha(\hat{\mu}_i)/g'(\hat{\mu}_i)^2 V(\hat{\mu}_i)$.

(iii) Find $\hat{\boldsymbol{\beta}}$, so that the weighted least squares objective function

$$\|\boldsymbol{z} - \boldsymbol{X}\boldsymbol{\beta}\|_W^2 + \sum_j \lambda_j \boldsymbol{\beta}^T \boldsymbol{S}^{[j]} \boldsymbol{\beta}$$

is minimized.

(iv) Update $\hat{\boldsymbol{\eta}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$ and $\hat{\mu}_i = g^{-1}(\hat{\eta}_i)$.

In the iteration steps we used that $\|a\|_W^2 = \boldsymbol{a}^T\boldsymbol{W}\boldsymbol{a}$ with $\boldsymbol{W} = diag(w_i)$ and $V(\mu)$ is the variance function which is calculated by the exponential family distribution or defining the quasi-likelihood. Finally we employed $\alpha(\mu_i) = [1+(Y_i-\mu_i)\{V'(\mu_i)/V(\mu_i)+g''(\mu_i)/g'(\mu_i)\}]$.

**Degree of freedom and scale parameter estimation**

An appropriate REML estimator of the scale parameter $\phi$, occurring in model setting (4.6) and penalized log likelihood (4.9), is (c.f. Wood (2017), Section 3.4.2)

$$\hat{\phi} = \frac{\|\boldsymbol{Z} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\|_W^2}{n - \tau}, \tag{4.14}$$

where

$$\tau = tr\{(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X} + \boldsymbol{S}_\lambda)^{-1}\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X}\}, \tag{4.15}$$

and $\boldsymbol{S}_\lambda = \sum_j \lambda_j \boldsymbol{S}^{[j]}$.

Thus interpret $\tau$ as the effective degree of freedom of the model which (roughly) coincide with the degree of freedom, see Wood (2017), Section 5.4.2. Let $\boldsymbol{F} = (\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X} + \boldsymbol{S}_\lambda)^{-1}\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X}$ which takes the weights into account. We can interpret the matrix as a mapping from the un-penalized coefficient estimator to the penalized coefficient estimator. The trace is the average shrinkage realized by the coefficients and multiplied by the number of the coefficients. So the effective degree of freedom is now obtained by summing the $F_{ii}$ values.

Note that the REML estimator $\hat{\phi}$ is the Pearson estimator of the scale parameter, since $\|\boldsymbol{Z} - \boldsymbol{X}\boldsymbol{\beta}\|_W^2$ corresponds to the Pearson statistic. (see Wood (2017), Section 3.1.5, but since of the susceptible to problems, discussed on p.110, it is safer to use correct estimator (3.11) presented in this book)

**Example 27** *(Alternative definition of effective degree of freedom).*
*Consider the simple Gaussian additive model case*

$$Y_i = \beta_0 + \sum_j f_j(x_{ij}) + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2). \tag{4.16}$$

*This is structured like the Gaussian linear model (see Chapter 3) but with the presence of smoothing terms. Proceed with the Gaussian additive model as in the model setting of Section 4.2.1 to yield the model matrix $\boldsymbol{X}$. Here we have the influence matrix $\boldsymbol{A} := \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X} + \boldsymbol{S}_\lambda)^{-1}\boldsymbol{X}^T$ and $\boldsymbol{F} = (\boldsymbol{X}^T\boldsymbol{X} + \boldsymbol{S}_\lambda)^{-1}\boldsymbol{X}^T\boldsymbol{X}$.*

*The expected residual sum of squares for this model is then*

$$E(\|\boldsymbol{Y} - \boldsymbol{A}\boldsymbol{Y}\|^2) = \sigma^2\{n - 2tr(\boldsymbol{A}) + tr(\boldsymbol{A}\boldsymbol{A})\} + \boldsymbol{b}^T\boldsymbol{b},$$

*where $\boldsymbol{b} = \boldsymbol{\mu} - \boldsymbol{A}\boldsymbol{\mu}$ is the smoothing bias. The smoothing bias is estimated as $\hat{\boldsymbol{b}} = \hat{\boldsymbol{\mu}} - \boldsymbol{A}\hat{\boldsymbol{\mu}}$. These results lead us to the variance or scale estimator*

$$\hat{\sigma}^2 = \frac{\|\boldsymbol{Y} - \boldsymbol{A}\boldsymbol{Y}\|^2 - \hat{\boldsymbol{b}}^T\hat{\boldsymbol{b}}}{n - 2tr(\boldsymbol{A}) + tr(\boldsymbol{A}\boldsymbol{A})},$$

*and thus set the effective degree of freedom of the model as $\tau_1 = 2tr(\boldsymbol{A}) - tr(\boldsymbol{A}\boldsymbol{A}) = 2tr(\boldsymbol{F}) - tr(\boldsymbol{F}\boldsymbol{F})$. The term-specific effective degree of freedom can be calculated by taking the corresponding elements of $diag(2\boldsymbol{F} - \boldsymbol{F}\boldsymbol{F})$ and summing these up.*

*An alternative way to get $\tau_1$ is to consider the bias corrected fitted values. For this way, see Wood (2017), Section 6.1.2. This alternative is useful for computing our p-values of the smooths.*

For stable least squares in case of negative weights, see Wood (2017), Section 6.1.3.

Henceforth, we will use the case of Gaussian additive models as the basis for further GAM calculations.

### 4.2.3 Tensor product smooth interactions

Before we become acquainted with choosing the optimal smoothing parameter $\lambda$ and setting the basis dimension, there is an important part of the regression model to show. That is the interaction between covariates.

In contrary to the linear regression, GAM gives four possible ways to include the interaction term to the model.

For simplicity, consider the model (4.11) with just two covariates $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$.

(i) $\underline{\boldsymbol{x}_1 \times \boldsymbol{x}_2}$: Multiplication of two covariates, like we already seen in Section 3.3.2 with the multiple linear regression.

(ii) $\underline{f_1(\boldsymbol{x}_1) \times \boldsymbol{x}_2}$: Interaction between a smoothed function to one covariate.

(iii) $\underline{f_1(\boldsymbol{x}_1, \boldsymbol{x}_2) := f_1(\boldsymbol{x}_1) \times f_2(\boldsymbol{x}_2)}$: Interaction between smoothers, that is use the same smoothed function for both covariates.

(iv) $\underline{f_1(\boldsymbol{x}_1) \otimes f_2(\boldsymbol{x}_2)}$: The most complex interaction term in GAM is to set the tensor product interactions. In contrast to the third possibility, this option uses different smoothing basis for two covariates and penalize it in two different ways.

The most interesting interaction is the tensor product case, which we use later when we include interactions in our application part. The tensor product interactions has the following representation:

$$f_{12}(\boldsymbol{x}_1, \boldsymbol{x}_2) = \sum_{i=1}^{I} \sum_{j=1}^{J} \delta_{ij} b_{i1}(\boldsymbol{x}_1) b_{i2}(\boldsymbol{x}_2), \tag{4.17}$$

where $b_1$ and $b_2$ are the two basis functions, $I$ and $J$ are two corresponding basis dimensions and $\delta$ is a vector of unknown coefficients.

**Example 28** *Considering our simple additive model structure (4.2). As for the linear models, we can include some interaction terms in the regression equation (4.2):*

$$Y_i = \alpha + f_1(\texttt{T1.kitchen}_i) + f_2(\texttt{T2.living}_i) + f_{12}(\texttt{T1.kitchen}_i, \texttt{T2.living}_i) + \varepsilon_i, \tag{4.18}$$

*for* $i = 1, \ldots, 19735$.

*Due to the structure of* $f_{12}$*, formally in (4.17), there are more than one penalization* $\boldsymbol{S}^{[12]}$*. In all other aspects, the interaction term is treated exactly as shown above for the smooth functions* $f_j$*.*

## 4.3 Selection criterion of the smoothness

In this section we will provide some smoothness selection criteria.

We already set up the model and considered the estimation for $\boldsymbol{\beta}$. But to estimate $\boldsymbol{\beta}$ we assumed the given smoothing parameter $\boldsymbol{\lambda}$. In the following we estimate the smoothing parameter $\boldsymbol{\lambda}$. We will see that this part of model estimation is challenging.

In general use, we have two classes of method, the prediction error method, based on GCV and AIC, or marginal likelihood method, based on the Bayesian/mixed model.

### 4.3.1 Un-biased risk estimator (UBRE) for known scale parameter

Assume we want to estimate the smoothing parameter in a simple case of an additive model, which is for data with constant known variance.

Our goal here is to provide a $\hat{\boldsymbol{\mu}}$ that is as close as possible to the true $\boldsymbol{\mu} \equiv E[\boldsymbol{Y}]$.

This can be achieved by taking the expected mean square error of the model, i.e. $MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)$ or the estimator in matrix formulation $MSE = \frac{1}{n}\|\boldsymbol{Y} - \hat{\boldsymbol{Y}}\|^2$, and the fact that a scalar is its own trace (c.f. Wood (2017), Section 1.8.6, which is leading to (1.13) from this book) which implies that

$$M = E\left(\|\boldsymbol{\mu} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\|^2/n\right) = E\left(\|\boldsymbol{Y} - \boldsymbol{A}\boldsymbol{Y}\|^2\right)/n - \sigma^2 + 2tr(\boldsymbol{A})\sigma^2/n. \tag{4.19}$$

Remember, $\boldsymbol{A}$ is the influence matrix or equivalently the hat matrix (3.23).

The estimate $M$ has to be minimized. An appropriate way is to choose the smoothing parameter so that the un-biased risk estimator (UBRE) is as follows

$$\mathcal{V}_u(\boldsymbol{\lambda}) = \|\boldsymbol{Y} - \boldsymbol{A}\boldsymbol{Y}\|^2/n - \sigma^2 + 2tr(\boldsymbol{A})\sigma^2/n. \tag{4.20}$$

Note that the risk estimator depends on the smoothing parameter which is included in $\boldsymbol{A}$. Furthermore we want to indicate that the risk estimator is equivalent with the Mallows's $C_p$, but for more details we refer to Christensen, 2018 or (Mallows, 1973).

Estimating $\boldsymbol{\lambda}$ by minimizing $\mathcal{V}_u(\boldsymbol{\lambda})$ over $\boldsymbol{\lambda}$ works good for a known $\sigma^2$, but for an unknown $\sigma^2$ there arises problems using the MSE estimator $M$.

**Example 29** *(Using MSE estimator M for unknown $\sigma^2$) Using the scale parameter estimator (4.14) and substitute the resulting approximation*

$$E\left(\|\boldsymbol{Y} - \boldsymbol{A}\boldsymbol{Y}\|^2\right) = \sigma^2\{n - tr(\boldsymbol{A})\}$$

*into the MSE estimator (4.19) which yields*

$$M = E\left(\|\boldsymbol{\mu} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\|^2/n\right) = \frac{tr(\boldsymbol{A})}{n}\sigma^2$$

*and thus we have our MSE estimator for unknown $\sigma^2$ with $\tilde{M} = \frac{tr(\boldsymbol{A})}{n}\hat{\sigma}^2$.*

*If we now consider a un-penalized models with one and two parameters. Before the selection would judge an improvement, the two parameter models has to reduce $\hat{\sigma}^2$ to less than half the one parameter $\sigma^2$ estimate. This is not a suitable basis model selection.*

### 4.3.2 Cross validation for unknown scale parameter

Since minimizing the average square error in model predictions of $E[\boldsymbol{Y}]$ does not work well for unknown $\sigma^2$, we have to look for an alternative. An appropriate way is to base smoothing parameter estimation on mean square prediction error. So adding a new observation $y$ using the fitted model, the expected mean square prediction error can be represented as

$$P = \sigma^2 + M.$$

Since the direct dependence on $\sigma^2$ tends to the mean, the criteria based on $P$ are more resistant to over-smoothing.

The estimation of $P$ is now provided by the cross validation (e.g. Stone (1974)). To use the method we omit a single datum $y_i$ from the process to fit the model. Doing this, the one response variable $Y_i$ becomes independent of the model fitted with the remaining data points. By omitting all data in turn, we yield the ordinary cross validation (OCV) estimator of $P$,

$$\mathcal{V}_0 = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{\mu}_i^{[-i]}\right)^2,$$

where $\mu_i^{[-i]}$ denotes the prediction of $E[Y_i]$ which is the result of fitting the model without $y_i$.

Note that to obtain the $n$ terms $\hat{\mu}_i^{[-i]}$ for calculating $\mathcal{V}_0$, there is no need to perform $n$ model fits. Consider the penalized least squares objective function (4.12) and minimizing it to get the $i$-th term in the OCV score

$$\sum_{j=1,j\neq i}^{n}\left(y_j - \hat{\mu}_j^{[-i]}\right)^2 + penalties.$$

(The added penalties in the sum of squares term does not depend on the included observations.)

Now including the zero term $(\hat{\mu}_i^{[-i]} - \hat{\mu}_i^{[-i]})^2$ to obtain the unchanged objective function

$$\sum_{j=1}^{n}\left(y_j^\star - \hat{\mu}_j^{[-i]}\right)^2 + penalties, \tag{4.21}$$

with $\boldsymbol{y}^\star = \boldsymbol{y} - \bar{\boldsymbol{y}}^{[i]} + \bar{\boldsymbol{\mu}}^{[i]}$. $\bar{\boldsymbol{y}}^{[i]}$ and $\bar{\boldsymbol{\mu}}^{[i]}$ are vectors with the $i$-th elements are $y_i$ and $\hat{\mu}_i^{[-i]}$, respectively, and the remaining entries are zeros.

Again, minimizing objective function (4.21) results in the $i$-th prediction $\hat{\mu}_i^{[-i]}$.

Observe that the difference now is that the fitting objective function (4.21) has the structure for the model with the whole data. Hence the fitting by minimizing (4.21) also yield the influence matrix $\boldsymbol{A}$ for the model fitted to all the data. To check this statement, consider the $i$-th prediction

$$\begin{aligned}\hat{\mu}_i^{[-i]} &= \boldsymbol{A}_i\boldsymbol{y}^\star \\ &= \boldsymbol{A}_i\boldsymbol{y} - A_{ii}y_i + A_{ii}\hat{\mu}_i^{[-i]} \\ &= \hat{\mu}_i - A_{ii}y_i + A_{ii}\hat{\mu}_i^{[-i]}\end{aligned} \tag{4.22}$$

As a reminder, $\hat{\mu}_i$ is obtained from the fit of the whole vector $\boldsymbol{y}$. Subtracting $y_i$ on both sides of the equation (4.22) and doing a transformation yields

$$y_i - \hat{\mu}_i^{[-i]} = \frac{(y_i - \hat{\mu}_i)}{(1 - A_{ii})},$$

which leads to the OCV score with just one single fit

$$\mathcal{V}_0 = \frac{1}{n} \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{(1 - A_{ii})^2}. \tag{4.23}$$

With this improved OCV score (4.23) there is clearly no need to calculate $n$ fits. (Stone (1977) demonstrates the asymptotic equivalence of cross validation and AIC which supports the results.)

There is also the leave-several-out cross validation that works analogously, but leaving out subsets of the data and also in this case only a single model fit is needed for the computations.

OCV is suitable method for estimating the smoothing parameter. But there are problems with ordinary cross validation, that is the expensive computation of minimizing in the case of the additive model with its several smoothing parameters and otherwise the lack of invariance. See Wood (2017), Section 6.2.(2) for more explanation.

### 4.3.3 Generalized cross validation

A solution is the generalized cross validation which does not suffer from this problem of the lack of invariance. The parameter estimation, effective degree of freedom (EDF) and expected prediction error are invariant to a rotation of $\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}$ by any orthogonal matrix $\boldsymbol{Q}$. The problem is the leading diagonal of the influence matrix, i.e. elements $A_{ii}$, are not invariant and neither are the individual terms in the sum of (4.23).

First, focusing on the rotations of $\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}$ and performing cross validation on rotated problems. As we want to base the cross validation on data, the one problem is the sensitivity to some data points which are very highly leverage relative to the others.

**Example 30** *Consider a highly uneven $A_{ii}$ values. They tend to cause the cross validation score (4.23) not to be based on the whole data. That is, due to the outliers, the cross validation score is dominated by a small proportion of the data.*

The idea is to choose the rotation $\boldsymbol{Q}$ so that the elements $A_{ii}$ are as even as possible. So taking the influence matrix $\boldsymbol{A}$ for the original problem, then the influence matrix for the rotated problem is

$$\boldsymbol{A}_Q = \boldsymbol{Q}\boldsymbol{A}\boldsymbol{Q}^T.$$

Further, if $\boldsymbol{B}$ is any matrix such that $\boldsymbol{B}\boldsymbol{B}^T = \boldsymbol{A}$, then the influence matrix now can be written as

$$\boldsymbol{A}_Q = \boldsymbol{Q}\boldsymbol{B}\boldsymbol{B}^T\boldsymbol{Q}^T.$$

All the elements on the leading diagonal of the influence matrix $\boldsymbol{A}_Q$ have the same value, if the chosen orthogonal matrix $\boldsymbol{Q}$ is such that each row of $\boldsymbol{QB}$ has the same Euclidean length. Since

$$tr(\boldsymbol{A}_Q) = tr(\boldsymbol{QAQ}^T) = tr(\boldsymbol{AQ}^T\boldsymbol{Q}) = tr(\boldsymbol{A}),$$

the values must be $tr(\boldsymbol{A})/n$.

For a detailed explanation that this neat row-length-equalizing property actually exist, see Wood (2017), Section 6.2.3.

Now, with the best rotation of the fitting problem, the adjusted ordinary cross validation score (4.23) can be generalized to the *generalized cross validation score (GCV)*

$$\mathcal{V}_g = \frac{n\|\boldsymbol{y} - \hat{\boldsymbol{\mu}}\|^2}{[n - tr(\boldsymbol{A})]^2}. \tag{4.24}$$

**Example 31** *Again consider the representation with the smoothing functions, i.e. the model (4.11). For the important procedure in choosing or estimating the optimal smoothing parameter $\lambda$ and the number of basis dimensions, i.e. the degree of freedom, we have to minimize the generalized cross validation score (GCV)*

$$\mathcal{V}_g = \frac{n\sum_{i=1}^{n}(y_i - \hat{f}_i)^2}{[tr(\boldsymbol{I} - \boldsymbol{A})]^2},$$

*where $\boldsymbol{A}$ is the influence matrix.*

*Let's look at two cases:*

*(i)* *$\lambda \to 1$: The spline is over-smoothed and therefore the model is highly smoothed*

*(ii)* *$\lambda \to 0$: The spline is not penalized and therefore the model fits most wiggles*

*In this case, we have a classical ordinary least squares regression behavior where the sum of squared residuals are minimal.*

*As for the number of basis dimensions, i.e. estimated degrees of freedom, we have the following behavior of fitted values:*

*(i)* *Higher basis dimension: the fit is less smoothed.*

*(ii)* *Lower basis dimension: the fit is more smoothed.*

*In summary, the number of basis functions and the smoothing parameters interact to control the wiggliness of a smooth function. To visualize our finding, we will take the additive model (4.2) into account. To simplify the visualization, we only plot* $\widehat{\log(\texttt{app}_i)} = \hat{\alpha} + \hat{f}_1(\texttt{T1.kitchen}_i)$ *with* $i = 1, \ldots, 19735$, *c.f. Figure 4.1. It becomes clear how changing both together affects model behavior. When the number of basis functions is high and the smoothing parameter is too low to smooth the model, it ends up over-fitting the data.*

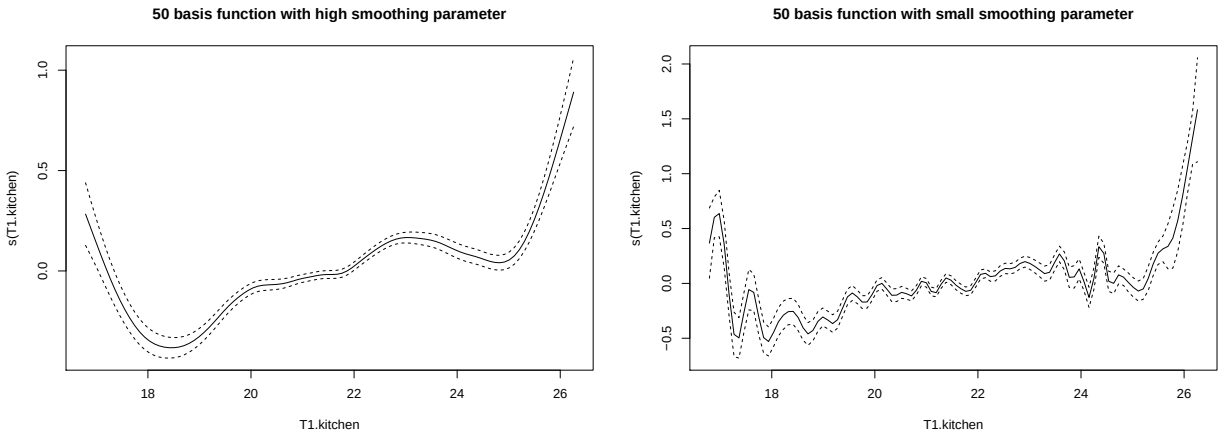



Figure 4.1: Fit of GAM model $\widehat{\log(\texttt{app}_i)} = \hat{\alpha} + \hat{f}_1(\texttt{T1.kitchen}_i)$, i.e. a model with logarithmic transformed appliances as a smooth function of the kitchen temperature with 50 basis functions and a smoothing parameter of $\lambda = 0.9999$ on the left panel and $\lambda = 0.0001$ on the right panel. Note that this representation, with fixed basis function and smoothing parameter, can be created across all methods, e.g. GCV and REML. All these methods will show similar results.

Review the tensor product interaction term in (4.17). Unlike the usage of one smoothing function, the advantage in using the tensor product is that the form is invariant to a re-scaling of its covariates, since it allows for an overall anisotropic, i.e. different in each direction, penalty. This is also due to the benefit of allowing different metrics of variables in the interaction term.

Note that the expected prediction error is not affected by the rotation which means the GCV is the OCV on the rotation problem. Not only that the GCV is also valid for estimate of prediction error, but also it is invariant.

Another adjustment is the double cross validation which focus of the sensitivity to over-fitting, see Wood (2017), Section 6.2.4.

### 4.3.4 Prediction error criteria for the generalized case

For the generalized model case (4.6) there a many ways to yield the smoothing parameter selection criteria. This works by substituting the model deviance or the Pearson statistic for the residual sum of square in the UBRE score (4.20) or the GCV score (4.24). Since in practice, the Pearson statistic tend to under-smooth, the deviance based methods are preferred in general.

Setting the deviance into the scores, the UBRE score becomes

$$\mathcal{V}_u(\boldsymbol{\lambda}) = D(\hat{\boldsymbol{\beta}}) + 2\gamma\phi\tau,$$

where $\phi$ was known in this case, and the GCV score becomes

$$\mathcal{V}_g(\boldsymbol{\lambda}) = \frac{nD(\hat{\boldsymbol{\beta}})}{(n-\gamma\tau)^2}, \tag{4.25}$$

where $\tau$ is the model effective degree of freedom (4.15) and $\gamma$ is set to 1, usually, but can be increased to force smoother models.

These criteria are discussed in Hastie and Tibshirani (1990). Additional calculations and details can be found in Wood (2017), Section 6.2.5.

### 4.3.5 Marginal likelihood and REML

As we mentioned at the beginning of this section, the other class of method for smoothness selection criteria is the Bayesian view of smoothing.

The idea is taken from the traditional Bayesian inference to get posterior distributions of the model coefficient $\boldsymbol{\beta}$ conditioned on the observations $\boldsymbol{y}$ and, of course, the conditioning on the smoothing parameter $\boldsymbol{\lambda}$. This can be done by using the Bayes rule, i.e. $f(\boldsymbol{\beta}|\boldsymbol{y},\boldsymbol{\lambda}) = \frac{f(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{\lambda})\cdot f(\boldsymbol{\beta}|\boldsymbol{\lambda})}{f(\boldsymbol{y}|\boldsymbol{\lambda})}$, with $\boldsymbol{\beta}$ is the model coefficient vector, $f(\boldsymbol{\beta}|\boldsymbol{\lambda})$ is the prior density function of $\boldsymbol{\beta}$, $\boldsymbol{y}$ is the vector of observations and (the observation model) $f(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{\lambda})$ is the conditional data density given $\boldsymbol{\beta}$ and $\boldsymbol{\lambda}$. Next, we take the denominator as it is the density of the choices of the smoothing parameter. The calculation of the denominator is based on the Bayesian rule for inference, where the dependence of the model coefficient is marginalized out from the numerator, so that we obtain a non-dependency on $\boldsymbol{\beta}$. Thus it takes the form $f(\boldsymbol{y}|\boldsymbol{\lambda}) = \int f(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{\lambda})f(\boldsymbol{\beta}|\boldsymbol{\lambda})d\boldsymbol{\beta}$.

The goal is to compare the smoothing parameters $\boldsymbol{\lambda}$.

Here the smoothing penalties correspond to a Gaussian prior on the model coefficients. In this approach we choose the smoothing parameter that maximize the *Bayesian log marginal likelihood*

$$\mathcal{V}_r(\boldsymbol{\lambda}) = \log \int f(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{\lambda})f(\boldsymbol{\beta}|\boldsymbol{\lambda})d\boldsymbol{\beta}. \tag{4.26}$$

Now the score is the logarithm of the joint density of the data and coefficient $\boldsymbol{\beta}$, where the coefficients integrated out. The interpretation of this integral is that it is an average likelihood of random draws from the prior.

The approach is called *empirical Bayes* of estimating parameter by maximizing the marginal likelihood (4.26) which has the form of the REML criterion, i.e. choose $\lambda$ so that the prior variance is about right, compare Figure 4.2, and where the random coefficients have Gaussian distributions.

We want to find the maximum of the function to be integrated and for that apply a second order Taylor series approximation for the logarithm of that function. This is the basic idea of a Laplace approximation. (c.f. MacKay and Mac Kay (2003) or better Azevedo-Filho and Shachter (1994)). Further, computing an expectation of the posterior distribution, the maximum is the MAP (maximum penalized likelihood) solution and the second order Taylor series corresponds to a Gaussian distribution for which integrals can be determined analytically. So the Laplace approximation can be used to find the best solution in case multiple local maxima exist.

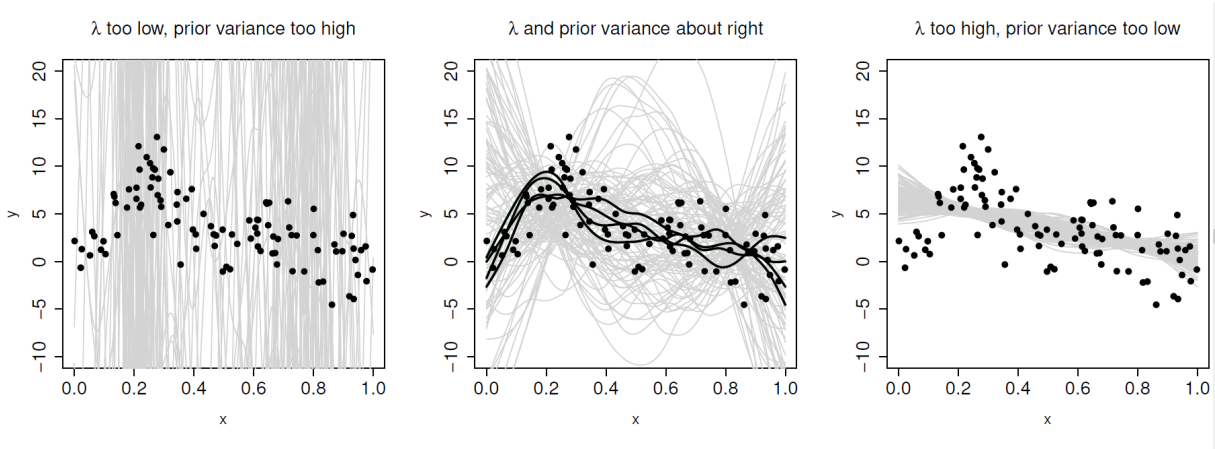



Figure 4.2: The figure is adopted from Wood (2017), Section 6.2, Figure 6.3. The simulation shows the workflow of method REML starting with its drawing from the prior. A low smoothing parameter leads to the drawing lines fail to closely pass the real data which is given on the left hand side. Contrary on the right side, a high smoothing parameter leads to a high smoothness and thus also fails to closely pass the real data. In the middle we have the right smoothing parameter and thus the smoothness is right so that the real data can be passed closely and a high likelihood is obtained. The black curves are the ones that have reached the highest likelihoods.

Now to evaluate the integral in (4.26), we use the Laplace approximation (c.f. Wood (2017), Section 3.4, with help of equation (3.17) presented in the book) which results for our Gaussian family case in

$$\mathcal{V}_r(\boldsymbol{\lambda}) = l(\hat{\boldsymbol{\beta}}) - \frac{\hat{\boldsymbol{\beta}}^T \boldsymbol{S}_\lambda \hat{\boldsymbol{\beta}}}{2\sigma^2} - \frac{\log |\boldsymbol{S}_\lambda/\sigma^2|_+}{2} - \frac{\log |\boldsymbol{X}^T\boldsymbol{X}/\sigma^2 + \boldsymbol{S}_\lambda/\sigma^2|}{2} + \frac{M}{2}\log(2\pi), \quad (4.27)$$

where $l$ is the log likelihood, $M$ is the dimension of the null space of $\boldsymbol{S}_\lambda$, and $|\boldsymbol{A}|_+$ denotes a product of the non-zero eigenvalues of $\boldsymbol{A}$.

Note that in the Gaussian additive model case the score (4.27) is exact. Due to the Laplace approximation we have a term in the Taylor expansion which do not have to be dropped in the Gaussian family case.

The use of the marginal likelihood/REML method for smoothing parameter selection is well-established, but there are also problems with $\log |\boldsymbol{S}_\lambda|_+$ when optimizing $\mathcal{V}_r$ which is discussed in Wood (2017), Section 6.2.7.

### 4.3.6 Prediction error criteria versus marginal likelihood

To give a short comparison of these methods, we look at two important conditions, the smoothness and performance. While REML asymptotically under-smooths relative to GCV which has a better asymptotic prediction error performance, the GCV seems to be much slower. But to summarize, REML is more resistant to occasional severe overfitting, when higher variability exists and a higher variable estimates of $\lambda$ as a consequence which fits most wiggles.

That is why in practice, the REML method is used more often. Furthermore Reiss and Todd Ogden (2009) focus on spline-based approaches to non-parametric and semi-parametric regression and examine the two preferred methods, GCV and REML. With the help of two data sets the ideas are illustrated which results in favoring the REML method. Also Wood, 2011 seize this suggestion and optimized the REML method with the mentioned Laplace approximation and finds out that the REML shows improvements in terms of mean square error performance and the numerical robustness relative to GCV. In addition the REML method achieve less severe under-smoothing failure.

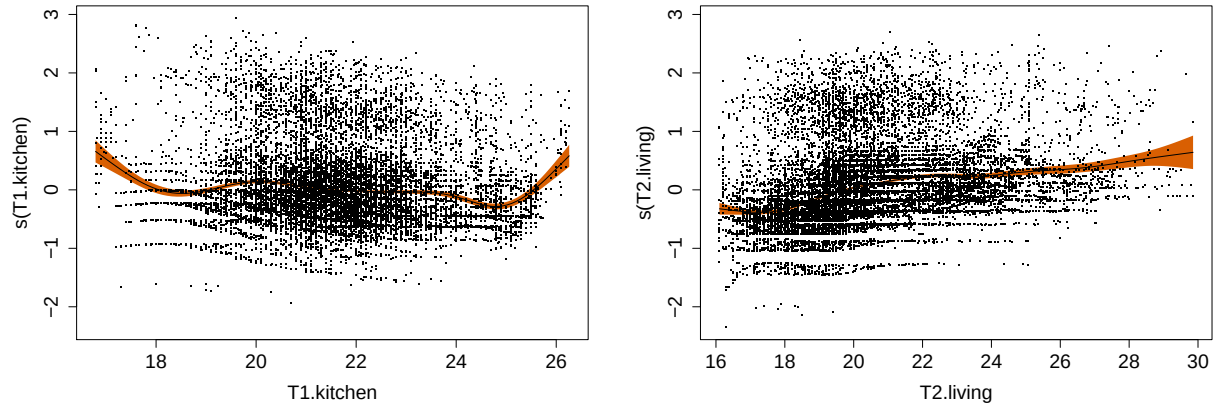
Another interesting article about the method differences is written by David Lawrence Miller and is published on github under the following website:

*https://github.com/DistanceDevelopment/dsm/wiki/Why-is-the-default-smoothing-method-%22REML%22-rather-than-%22GCV.Cp%22%3F* (seeMiller (4 04), [accessed: 2020-01-08]). He also argued with Simon Wood, the author our reference book Wood (2017), about the advantages of using the REML method when fitting models with finite sample size.

**Example 32** *Again using the GAM regression (4.2), we want to compare the two methods. In Figure 4.3 the GCV and REML objective function is shown. There is not an obvious difference, but looking closely at the tails of the right plots, we see that the the GCV method has a slightly wider confidence interval than the lower REML method which confirms the resistance of over-fitting.*

Thus we are going to use only the REML method in our application part later.

**Method: GCV**




**Method: REML**

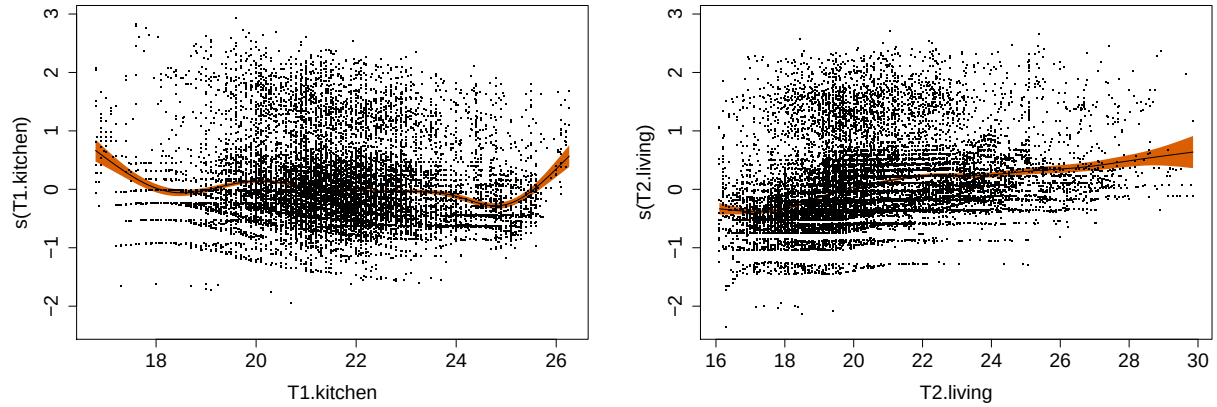



Figure 4.3: On the top panel we have fitted the GAM regression (4.2) with the method GCV and the lower panel the method REML is used.

## 4.4 Posterior distributions and confidence intervals

We already discussed the prior distributions. Also for a better understanding we refer to Wood (2017), Section 4.2.4., p. 172 and Section 5.8 p. 239.

Take the Bayesian view of the smoothing process and we get $\boldsymbol{\beta}$ with zero mean improper Gaussian prior distribution with precision matrix proportional to $\boldsymbol{S}_\lambda$ in (4.10). Then with help of our obtained distribution for linear models in (3.24), the posterior distribution of $\boldsymbol{\beta}$ is

$$\boldsymbol{\beta}|\boldsymbol{Y}, \boldsymbol{\lambda} \sim N(\hat{\boldsymbol{\beta}}, \boldsymbol{V}_\beta) \tag{4.28}$$

As we see, in the Gaussian identity link case we have $\boldsymbol{V}_\beta = (\boldsymbol{X}^T\boldsymbol{X} + \boldsymbol{S}_\lambda)^{-1}\sigma^2$. In general case estimated by PIRLS, like exponential family case, we have $\boldsymbol{V}_\beta = (\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X} + \boldsymbol{S}_\lambda)^{-1}\phi$, or more general regular likelihood at $\hat{\boldsymbol{\beta}}$ it becomes $\boldsymbol{V}_\beta = (\hat{\boldsymbol{\mathcal{I}}} - \boldsymbol{S}_\lambda)^{-1}$ with $\hat{\boldsymbol{\mathcal{I}}}$ as the Hessian of the negative log likelihood at $\hat{\boldsymbol{\beta}}$ or its expectation.

These results from the model are used for computing Bayesian credible intervals for any quantity $\alpha$ for prediction. From (4.28) we simulate replicate coefficient vectors from which we compute $\alpha$ element-wise and produce a sample from the posterior distribution of $\alpha|\boldsymbol{Y}$.

## 4.5 AIC and smoothing parameter uncertainty

By performing the estimation of smoothing parameters, the traditional way of model selection is considered. But it stops with removing terms from the model. Since we are also interested to compare models that are not necessarily nested, i.e. also non-nested, the issue accrue on how to select between all the models in a reasonable manner.

One very popular method for model selection with model regression is the Akaike information criterion, as we already introduced in the Chapter describing multiple linear models, Section 3.5.3. But we have to take care with the usage in connection with models containing random effects and smoother.

Two approaches are to be taken into account which differing in the way they deal with smooths:

- Marginal AIC

  This approach is based on the marginal likelihood of the model. The number of coefficients are the number of fixed effects and variance and smoothing parameters, which is used for the AIC penalty. So in practical problems the marginal likelihood underestimates variance components which means with respect to the smoother, that it can be made too smooth.

- Conditional AIC

  This traditional approach is based on the likelihood of all the coefficients, conditioned on their maximum penalized likelihood (MAP) estimates. The number of coefficients in this penalty are some estimate of the effective number of parameter. In the AIC penalty term the model degree of freedom $\tau$ from (4.15) is used. But it

is shown that the problem of neglecting the smoothing parameter of uncertainty in $\tau$, leads more likely to select a model which contains a random effect that is not in the true model.

Now, to correct the parameter $\tau$ for using it in the AIC penalty term there is an idea to solve the mentioned problem.

### 4.5.1 Uncertainty of smoothing parameter

Firstly, note that smoothing parameter are treated as fixed which ignores the uncertainty in the estimation. Wood (2017) summarized the proposed solution which was given from Kass and Steffey (1989). The idea is shown by computing a first order correction for this uncertainty in context of i.i.d. Gaussian random effects in a one way ANOVA type design. Their general approach can be extended.

Let $\rho_i = \log \lambda_i$ and $\boldsymbol{S}_\lambda = \sum_j \lambda_i \boldsymbol{S}_j$. Using the Bayesian large sample approximation as the Gaussian case

$$\boldsymbol{\beta}|\boldsymbol{Y}, \boldsymbol{\rho} \sim N(\hat{\boldsymbol{\beta}}_\rho, \boldsymbol{V}_\beta), \quad \text{where } \boldsymbol{V}_\beta = (\hat{\boldsymbol{\mathcal{I}}} + \boldsymbol{S}_\lambda)^{-1}, \tag{4.29}$$

where $\hat{\boldsymbol{\beta}}_\rho$ denotes the estimator of $\boldsymbol{\beta}$ obtained by using the smoothing parameters $\rho_i$. The large sample approximation is

$$\boldsymbol{\rho}|\boldsymbol{Y} \sim N(\hat{\boldsymbol{\rho}}, \boldsymbol{V}_\rho), \tag{4.30}$$

where $\boldsymbol{V}_\rho$ is the inverse of the Hessian of the negative log marginal likelihood with respect to $\boldsymbol{\rho}$.

For the improvement on using (4.29) with fixed $\boldsymbol{\rho}$ at its estimator, we assume correct (4.29) and (4.30), $\boldsymbol{Z} \sim N(0, \boldsymbol{I})$ and independently $\hat{\boldsymbol{\rho}}^\star \sim N(\hat{\boldsymbol{\rho}}, \boldsymbol{V}_\rho)$, then

$$\boldsymbol{\beta}|\boldsymbol{Y} \overset{d}{=} \hat{\boldsymbol{\beta}}_{\rho^\star} + \boldsymbol{R}_{\rho^\star}^T \boldsymbol{Z},$$

with $\boldsymbol{R}_{\rho^\star}^T \boldsymbol{R}_{\rho^\star} = \boldsymbol{V}_\beta$ and $\boldsymbol{V}_\beta$ depending on $\boldsymbol{\rho}^\star$. The computation when simulation from $\boldsymbol{\beta}|\boldsymbol{Y}$ is expensive, due to the re-computation of $\hat{\boldsymbol{\beta}}_{\rho^\star}$ and $\boldsymbol{R}_{\rho^\star}$ for each sample. So consider the first order Taylor expansion (c.f. Karpfinger et al. (2015), Section 10.4 Taylor) as an alternative way

$$\boldsymbol{\beta}|\boldsymbol{Y} \overset{d}{=} \hat{\boldsymbol{\beta}}_{\hat{\rho}} + \boldsymbol{J}(\boldsymbol{\rho} - \hat{\boldsymbol{\rho}}) + \boldsymbol{R}_{\hat{\rho}}^T \boldsymbol{Z} + \sum_k \left. \frac{\partial \boldsymbol{R}_\rho^T \boldsymbol{Z}}{\partial \rho_k} \right|_{\hat{\boldsymbol{\rho}}} (\rho_k - \hat{\rho}_k) + r,$$

where $r$ is a lower order remainder term and $\boldsymbol{J} = d\hat{\boldsymbol{\beta}}/d\boldsymbol{\rho}|_{\hat{\boldsymbol{\rho}}}$.

The expectation on the right hand side is $\hat{\boldsymbol{\beta}}_{\hat{\rho}}$ when leaving out $r$. Now taking care of the covariance matrix gives

$$\boldsymbol{V}_\beta^{'} = \boldsymbol{V}_\beta + \boldsymbol{V}^{'} + \boldsymbol{V}^{''}, \quad \boldsymbol{V}^{'} = \boldsymbol{J}\boldsymbol{V}_\rho\boldsymbol{J}^T \quad \text{and } V_{jm} = \sum_i^p \sum_l^M \sum_k^M \frac{\partial R_{ij}}{\partial \rho_k} V_{\rho,kl} \frac{\partial R_{im}}{\partial \rho_l}. \tag{4.31}$$

The computation of the derivative of the Cholesky factor is handled in Wood (2017), p. 422.

Finally dropping $\boldsymbol{V}^{''}$, we have the approximation from Kass and Steffey (1989):

$$\boldsymbol{\beta}|\boldsymbol{Y} \sim N(\hat{\boldsymbol{\beta}}_{\hat{\rho}}, \boldsymbol{V}^{\star}_{\beta}), \quad \text{where } \boldsymbol{V}^{\star}_{\beta} = \boldsymbol{V}_{\beta} + \boldsymbol{J}\boldsymbol{V}_{\rho}\boldsymbol{J}^{T}.$$

The first order Taylor expansion for $\hat{\boldsymbol{\beta}}$ about $\boldsymbol{\rho}$ yields similar formulation for the covariance matrix of $\hat{\boldsymbol{\beta}}$, i.e. $\boldsymbol{V}^{\star}_{\hat{\beta}} = (\hat{\boldsymbol{\mathcal{I}}} + \boldsymbol{S}_{\lambda})^{-1}\hat{\boldsymbol{\mathcal{I}}}(\hat{\boldsymbol{\mathcal{I}}} + \boldsymbol{S}_{\lambda})^{-1} + \boldsymbol{J}\boldsymbol{V}_{\rho}\boldsymbol{J}^{T}$ with $\hat{\boldsymbol{\mathcal{I}}}$ as the negative Hessian of the log likelihood.

### 4.5.2 Corrected AIC

Consider the AIC equation (3.29) and doing some substituting to improve the AIC. Replace the derivation with MLE by the penalized MLE, so that AIC can be represented as

$$\begin{aligned} AIC &= -2l(\hat{\boldsymbol{\beta}}) + 2E\left[(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_d)^T\hat{\boldsymbol{\mathcal{I}}}_d(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_d)\right] \\ &= -2l(\hat{\boldsymbol{\beta}}) + 2tr\left\{E\left[(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_d)(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_d)^T\right]\hat{\boldsymbol{\mathcal{I}}}_d\right\}, \end{aligned}$$

where $\boldsymbol{\beta}_d$ is the coefficient vector minimizing the $K - L$ divergence and $\hat{\boldsymbol{\mathcal{I}}}_d$ is the corresponding expected negative Hessian of the log likelihood.

Assuming an un-penalized situation, then $E[(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_d)(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_d)^T]$ is estimated as the observed inverse information matrix $\hat{\boldsymbol{\mathcal{I}}}^{-1}$ and $\tau^{'} = tr\left\{E[(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_d)(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_d)^T]\hat{\boldsymbol{\mathcal{I}}}_d\right\}$ is estimated as $tr(\hat{\boldsymbol{\mathcal{I}}}^{-1}\hat{\boldsymbol{\mathcal{I}}}) = k$. But since we have a penalized setting, $\hat{\boldsymbol{\mathcal{I}}}$ is not a good approximation of the expected inverse covariance matrix of $\hat{\boldsymbol{\beta}}$ any more. However we can approximate $\hat{\boldsymbol{\mathcal{I}}}_d$ by $\hat{\boldsymbol{\mathcal{I}}}$ and use the properties of the expectation which leads to $\tau^{'} = tr(\boldsymbol{V}_{\beta}\hat{\boldsymbol{\mathcal{I}}})$. For more we refer to Wood (2017), Section 6.10.1, equation (6.27) on p. 295. This expression has to be corrected for smoothing parameter uncertainty with the help of (4.31) to yield $\tau_2 = tr(\boldsymbol{V}^{'}_{\beta}\hat{\boldsymbol{\mathcal{I}}})$, and therefore the corrected AIC

$$AIC = -2l(\hat{\boldsymbol{\beta}}) + 2\tau_2. \tag{4.32}$$

For an illustration of the performance of (4.32) compared to the alternatives, we refer to Wood (2017), Figure 6.11.

## 4.6 Hypothesis testing and p-values

In this section we just give an idea on how the hypothesis testing works with generalized additive models.

An alternative to select models is the hypothesis testing for instance, in particular for choosing simpler models over complex ones. Meanwhile, the $p$-values for the parametric model effects are determined as for the un-penalized model. For this, review the Section 3.7 about hypothesis testing in the linear model case.

Again assume we want to test $H_0 : \boldsymbol{\beta}_l = 0$ with $\boldsymbol{\beta}_l$ is a subvector of $\boldsymbol{\beta}$ which contains the un-penalized (or fixed effects) coefficients. For this case, treating the smooths as random

effects, we can consult the frequentist or marginal covariance matrix for $\boldsymbol{\beta}_l$ are read from the Bayesian covariance matrix for $\boldsymbol{\beta}$, comparing to (4.13). Thus $\boldsymbol{V}_{\beta_l}$ denoting the block of $\boldsymbol{V}_\beta$ corresponding to $\boldsymbol{\beta}_l$.

So examine (3.41) for an involved scale parameter estimate, we have, approximately for the generalized case,

$$\hat{\boldsymbol{\beta}}_l^T \boldsymbol{V}_{\beta_l}^{-1} \hat{\boldsymbol{\beta}}_l / p_l \sim F_{p_l, n-(p+1)}$$

or no scale parameter estimate is involved, then

$$\hat{\boldsymbol{\beta}}_l^T \boldsymbol{V}_{\beta_l}^{-1} \hat{\boldsymbol{\beta}}_l \sim \chi^2_{p_l}.$$

We have $p$ and $p_l$ which are the dimension of $\boldsymbol{\beta}$ and $\boldsymbol{\beta}_l$, respectively.

For the generalized linear hypothesis testing, the null hypothesis is $H_0 : \boldsymbol{C}\boldsymbol{\beta}_l = \boldsymbol{d}$ and replace $\hat{\boldsymbol{\beta}}_l^T \boldsymbol{V}_{\beta_l}^{-1} \hat{\boldsymbol{\beta}}_l$ by $(\boldsymbol{C}\hat{\boldsymbol{\beta}}_l - \boldsymbol{d})^T (\boldsymbol{C}\boldsymbol{V}_{\beta_l}\boldsymbol{C}^T)^{-1} (\boldsymbol{C}\hat{\boldsymbol{\beta}}_l - \boldsymbol{d})$ in the above distributional outcome, as in (3.40) and (3.41).

For a single parameter test, we can equivalently using the reference distributions $t_{n-(p+1)}$ or $N(0, 1)$.

# Chapter 5

# Description of the energy consumption within a house data

The **appliance energy prediction** data set is made available by Luis Candanedo, *luismiguel.candanedoibarra@umons.ac.be*, University of Mons (UMONS) in the (UCI) machine learning repository. In particular the data can be obtained from

*https://archive.ics.uci.edu/ml/datasets/Appliances+energy+prediction*,

(see Luis Candanedo (2 15), [accessed: 2020-01-08]).

We are interested in modeling the use of energy for appliances and therefore lights over time accounting for external factors such as temperature and humidity.

## 5.1 House description

The data are recorded from a house which is located in Stambruges in Belgium. The low energy house was constructed under the condition of the passive house certification. Further, there are four occupants, two adults and two teenager, living in the house, with one adult regularly works in the home office.

Another interesting point that should be taken into account is the appliances energy metering in the different areas which is illustrated in the following Table 5.1.

As an additional insight and to underline our purposes, we take a look at the aggregated energy consumption per month with the percentage of each involved energy consumption of lights, appliances, electric heater, DHW and ventilation from January until May 2016. Lights range from one to four percent and appliances represents an energy use between 70 % and 79 % of monthly energy consumption. This fortifies our focus on appliances to analyze the energy consumption.

The temperature and humidity of the house were monitored with a wireless sensor network and is contained in our data set.

For a more detailed description of the house we refer to the above mentioned website and the paper of Candanedo et al. (2017).

| Room | Devices |
|---|---|
| Laundry | Fridge, freezer, wine cellar, washing machine, dryer, internet router, internet hub, network attached storage |
| Kitchen | Fridge, induction cook top, kitchen hood, microwave, oven, dishwasher, coffee machine |
| Living | TV, hard drive enclosure, DVD player, cable box, laptop, printer, electric blinds |
| Office | 2 desktop computers, 3 computer screens, router, laptop, printer, electric blind |
| ironing | Alarm clock, radio, iron, electric blind |
| parents | Alarm clock, radio, electric blind, 2 lamps |
| teenager | Computer (desktop and monitor or laptop), alarm clock, electric blind |
| bathroom | 2 electric toothbrushes, hair dryer |

Table 5.1: List of devices in all the relevant rooms or areas.

## 5.2 Description of recorded data

The data set contains 29 variables and 19735 observations over a period of 4.5 months (137 days). Most of the variables are numeric. The variables `app` and `lights` are a little bit different from the other numeric variables because it assumes very discrete, integer values. The `date` variable initially is of the structure factor, but is converted to a date-time class for our data analysis. If we now take a closer look at `date`, we see that the data were recorded at specific time points, that is every ten minutes from January 11th to May 27th, 2016. This ensures to capture even quick changes in energy consumption

Furthermore, we don't have to deal with any missing data, since all the time points are observed or, in the case of the variable `T.outstation`, interpolated. So we have a complete data set for our statistical regression models. Table 5.2 presents a list of all the variables, their structure and additional information as minimum, mean and maximum values.

In the application chapter of **Linear regression models**, in Section 7.1, we generate extra variables like `weekday` and `hours` so that we can include the time pattern in our regression model.

| variable | description | unit | class | min | mean | max |
|---|---|---|---|---|---|---|
| app | appliances energy use | Wh | integer | 10.00 | 97.69 | 1080 |
| lights | energy use of light fixtures in the house | Wh | integer | 0.00 | 3.80 | 70.00 |
| T1.kitchen | Temperature in kitchen area | °C | continuous | 16.79 | 21.69 | 26.26 |
| RH1.kitchen | Humidity in kitchen area | % | continuous | 27.02 | 40.26 | 63.36 |
| T2.living | Temperature in living room area | °C | continuous | 16.10 | 20.34 | 29.86 |
| RH2.living | Humidity in living room area | % | continuous | 20.46 | 40.42 | 56.03 |
| T3.laundry | Temperature in laundry room area | °C | continuous | 17.20 | 22.27 | 29.24 |
| RH3.laundry | Humidity in laundry room area | % | continuous | 28.77 | 39.24 | 50.16 |
| T4.office | Temperature in office room | °C | continuous | 15.10 | 20.86 | 26.20 |
| RH4.office | Humidity in office room | % | continuous | 27.66 | 39.03 | 51.09 |
| T5.bath | Temperature in bathroom | °C | continuous | 15.33 | 19.59 | 25.80 |
| RH5.bath | Humidity in bathroom | % | continuous | 29.82 | 50.95 | 96.32 |
| T6.outside | Temperature OUTSIDE the building (north side) | °C | continuous | -6.07 | 7.91 | 28.29 |
| RH6.outside | Humidity OUTSIDE the building (north side) | % | continuous | 1.00 | 54.61 | 99.90 |
| T7.ironing | Temperature in ironing room | °C | continuous | 15.39 | 20.27 | 26.00 |
| RH7.ironing | Humidity in ironing room | % | continuous | 23.20 | 35.39 | 51.40 |
| T8.teenager | Temperature in teenager room 2 | °C | continuous | 16.31 | 22.03 | 27.23 |
| RH8.teenager | Humidity in teenager room 2 | % | continuous | 29.60 | 42.94 | 58.78 |
| T9.parents | Temperature in parents room | °C | continuous | 14.89 | 19.49 | 24.50 |
| RH9.parents | Humidity in parents room | % | continuous | 29.17 | 41.55 | 53.33 |
| T.outstation | Temperature OUTSIDE (from Chievres weather station) | °C | continuous | -5.00 | 7.41 | 26.10 |
| RH.outstation | Humidity OUTSIDE (from Chievres weather station) | % | continuous | 24.00 | 79.75 | 100 |
| Pressure | Pressure (from Chievres weather station) | mm Hg | continuous | 729.30 | 755.50 | 772.30 |
| Windspeed | Wind speed (from Chievres weather station) | m/s | continuous | 0.00 | 4.04 | 14.00 |
| Visibility | Visibility (from Chievres weather station) | km | continuous | 1.00 | 38.33 | 66.00 |
| Tdewpoint | Dew Point | °C | continuous | -6.60 | 3.76 | 15.50 |
| date | time stamp | year-month-day hr:min:sec | index - factor | 2016-01-11 18:00:00 | | 2016-05-27 20:00:00 |

Table 5.2: Variable description of the appliance energy prediction data set.

# Chapter 6

# Exploration of the energy consumption within a house data

In the following we will focus on the variable app as the response variable and all the temperature and humidity variables as the covariates.

First we examine each variable individually and then pairwise. Finally, we analyze the pattern over time.

## 6.1 Marginal exploration

In this section we use the statistical method to describe each variable of the data set. We use the histogram to analyze the distribution and time series plots to detect pattern over time.

### 6.1.1 Response variable - Appliances

**Histogram of response variable app**

Figure 6.1 shows a histogram for the whole data sets which means $n = 19735$ observations. The appliance use of 50 Wh appears most often in our data set. We notice that the we have a right long tail in the count histogram (cf. top panel of Figure 6.1), which indicates a right-skewed data distribution. The average energy consumption is therefore on the left side of the middle of the data range limits, i.e.

$$\overline{\texttt{app}} > \texttt{app}_{med} > \texttt{app}_{mod}$$

with the mean of $\overline{\texttt{app}} = 97.695$Wh, the median of $\texttt{app}_{med} = 60$Wh and the mode of $\texttt{app}_{mod} = 50$Wh.

Thus the histogram visually gives the expected impression. In the higher energy use range, the Wh-values spread more than in the lower range. That is, the occupants are using the appliances more often in a lower watt-hour range, so they typically do not have an excessive energy usage.

Since many statistical techniques, such as linear regression, are based on the assumption that the variables have normal distribution, we transform app by the natural logarithm. Fortunately, we do not have any zeros in the measured set for app (see its minimum value in Table 5.2), we can simply use the natural logarithm without adding a small value. So taking the natural logarithm of this variable, we can transform our data to turn it maybe into a normal distribution or something more closely to a normal distribution, as we can see in the lower panel in Figure 6.1, as it is more centered now.

For this reason we will use the transformed variable lapp in the following sections, defined by

$$\texttt{lapp}_i = \ln(\texttt{app}_i), \quad \text{for } i = 1, \ldots, 19735.$$

The transformation results in the distribution points $\texttt{lapp}_{mod} = 3.9120$, $\texttt{lapp}_{med} = 4.0943$ and $\overline{\texttt{lapp}} = 4.3037$.

Again we have a slightly right-skewed distribution, but compared to the variable app we can make the rough conclusion that there is an approximately equivalent symmetry for the response variable lapp

$$\overline{\texttt{lapp}} \approx \texttt{lapp}_{med} \approx \texttt{lapp}_{mod}.$$

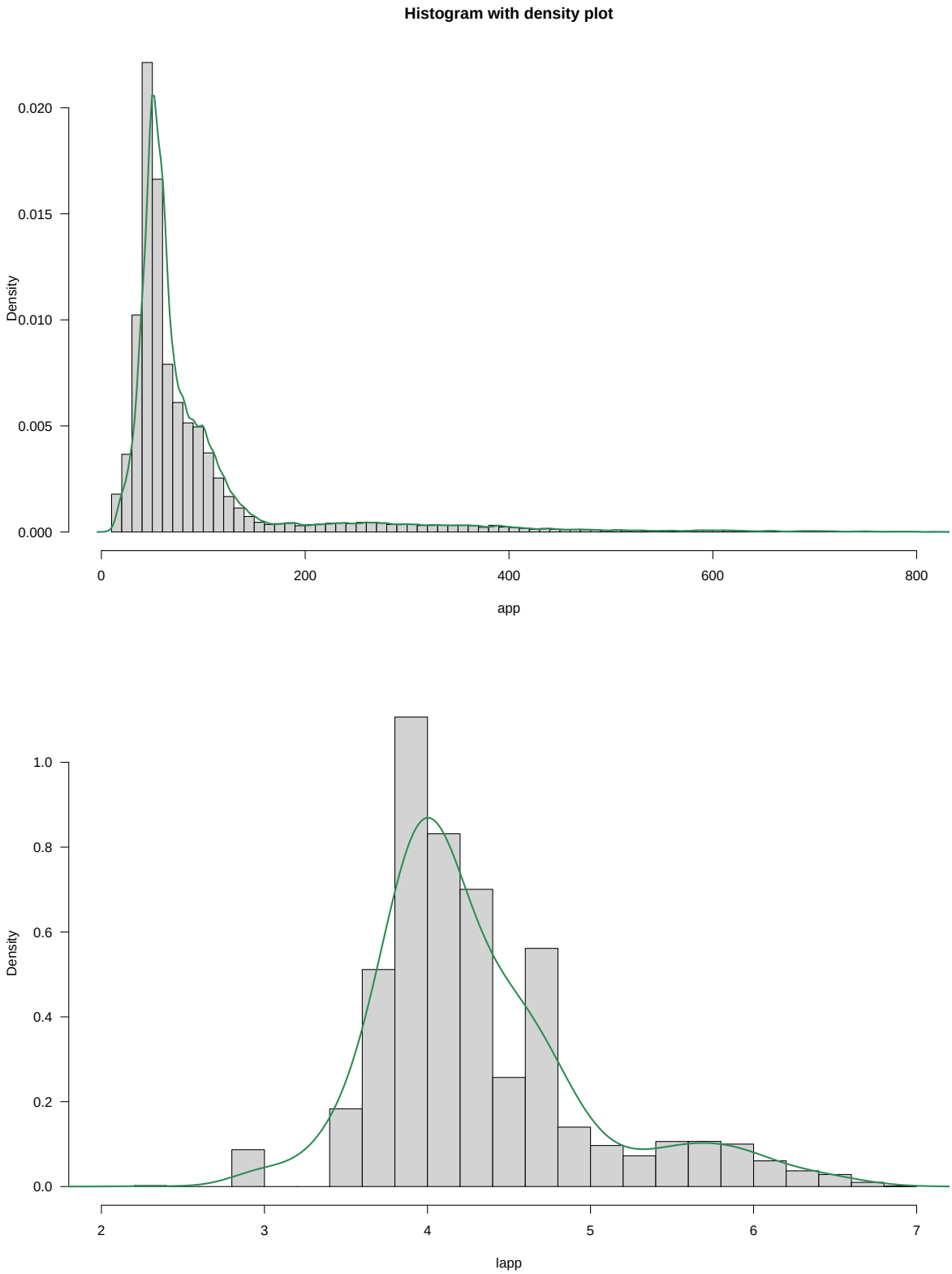



Figure 6.1: Appliances energy consumption distribution. Histograms with a density fit of `app` in the top panel and `lapp` in the lower panel. The histogram shows the frequency of energy use in intervals.

**Exploration of the appliances data over time**

First we plot time versus `app` and time versus `lapp` over the whole period to detect pattern of the energy consumption. To set more detail, we then select three monthly, weekly and finally four daily periods applying on our variable `lapp`.

Figure 6.4 and 6.5 show the energy consumption profile measured for the whole period. The time series of energy consumption has a high variability. A closer look at the figures is showing no clear pattern, except for the two time periods at the end of January and end of March, where we observe low energy consumption. But we can not tell, if the low values have a certain frequency or a random behavior of the occupants. Whereas in the monthly and weekly periods recognize some section wise pattern (see Figure 6.6 for monthly and Figure 6.9 for weekly periods). The figures indicate that energy consumption varies throughout the day and has its peaks during the days and its lows during the night. Additionally, it seems that the occupants been using appliances more often in winter months, since the fluctuations towards summer are shrinking. Compare the time series of January and May in Figure 6.6 and 6.9.

Now at last, as hinted in the weekly periods, we can see in the day periods that the appliances energy use has a certain pattern. The appliances tend to be in standby mode overnight and are in temporary use by the occupants from morning onward with its peaks. (cf. Figure 6.10). In March, the workdays Monday and Wednesday have a higher energy use over daytime, in contrast to Friday. A possible explanation could be that the occupants are at home or in the home office. Whereas the energy consumption is higher again on Sunday, perhaps due to domestic work on the weekend.

The expectation of the response variable `app` and `lapp` gives us an orientation for the following figures.

$$E[\texttt{app}] = 97.695, \qquad E[\texttt{lapp}] = 4.304.$$

Furthermore, for all the time series we give the corresponding autocorrelations in form of a correlogram. Remember that an autocorrelation plot is designed to show whether the data points of a time series are positively correlated, negatively correlated, or independent of each other. Thus the value of an autocorrelation function (acf) is in the range from $-1$ and 1 dependent of the lag between the data points of the time series.

The correlogram show the autocorrelation of the corresponding time series of the appliances energy consumption during the observation period. In summary, the spikes of the correlogram are statistically significant for the lags, since the spikes rises above the dashed significance level. This means that the energy consumption of the appliances are highly correlated with each other. In other words, when the energy use rises, it tends to continue rising. When the consumption falls, it tends to continue falling. The Figure 6.2, 6.3, 6.7 and 6.8 illustrate these autocorrelations and that there are no random time points, i.e. they are correlated in time.

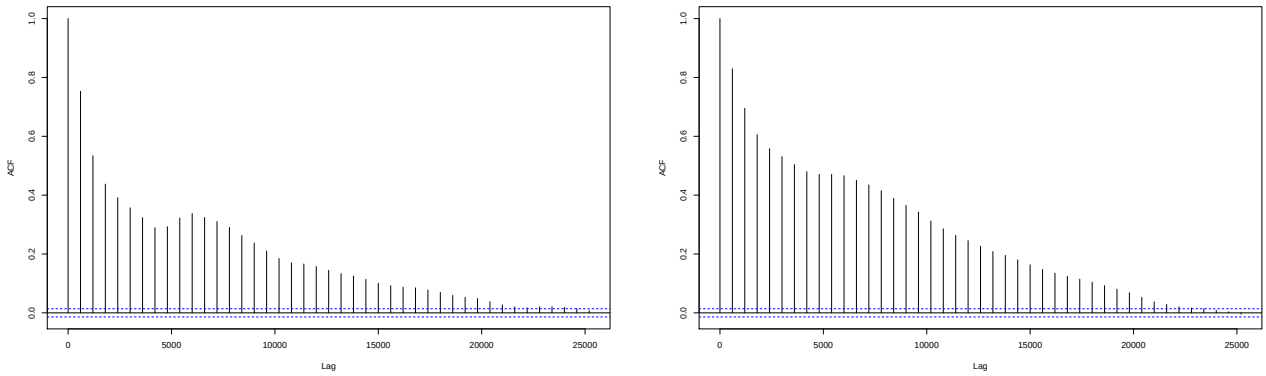



Figure 6.2: Correlogram. The autocorrelation plot of series `app` on the left side (corresponding to Figure 6.4), and autocorrelation plot of series `lapp` on the right side (corresponding to Figure 6.5) by lags. The dashed lines around zero showing the statistically significance level ($\alpha = 0.05$).

The autocorrelations of the series `lapp` by lag $h$ we see in the correlogram in Figure 6.2 are as follows

| $h$ | 0 | 600 | 1200 | 1800 | 2400 | 3000 | 3600 | 4200 | 4800 | 5400 | 6000 | 6600 | 7200 | 7800 | 8400 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $\hat{\gamma}_{\texttt{lapp}}(h)$ | 1.000 | 0.829 | 0.695 | 0.606 | 0.558 | 0.531 | 0.504 | 0.480 | 0.471 | 0.470 | 0.466 | 0.450 | 0.435 | 0.415 | 0.389 |

| $h$ | 9000 | 9600 | 10200 | 10800 | 11400 | 12000 | 12600 | 13200 | 13800 | 14400 | 15000 | 15600 | 16200 | 16800 | 17400 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $\hat{\gamma}_{\texttt{lapp}}(h)$ | 0.365 | 0.342 | 0.312 | 0.286 | 0.263 | 0.245 | 0.226 | 0.208 | 0.195 | 0.180 | 0.163 | 0.148 | 0.135 | 0.124 | 0.114 |

| $h$ | 18000 | 18600 | 19200 | 19800 | 20400 | 21000 | 21600 | 22200 | 22800 | 23400 | 24000 | 24600 | 25200 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $\hat{\gamma}_{\texttt{lapp}}(h)$ | 0.104 | 0.092 | 0.080 | 0.068 | 0.053 | 0.037 | 0.028 | 0.020 | 0.017 | 0.014 | 0.009 | 0.004 | $-0.005$ |

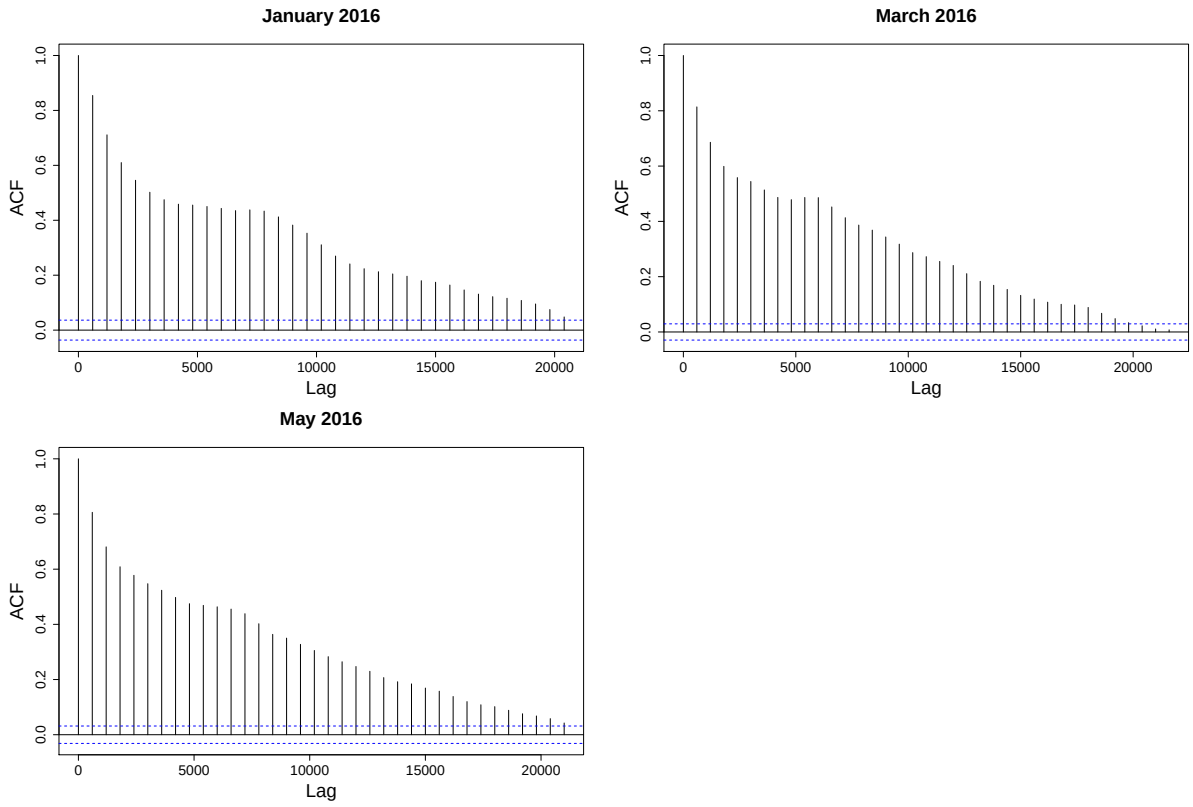



Figure 6.3: Correlogram. The monthly autocorrelation plots of series `lapp` by lags, corresponding to Figure 6.6. The dashed lines around zero showing the statistically significance level ($\alpha = 0.05$).

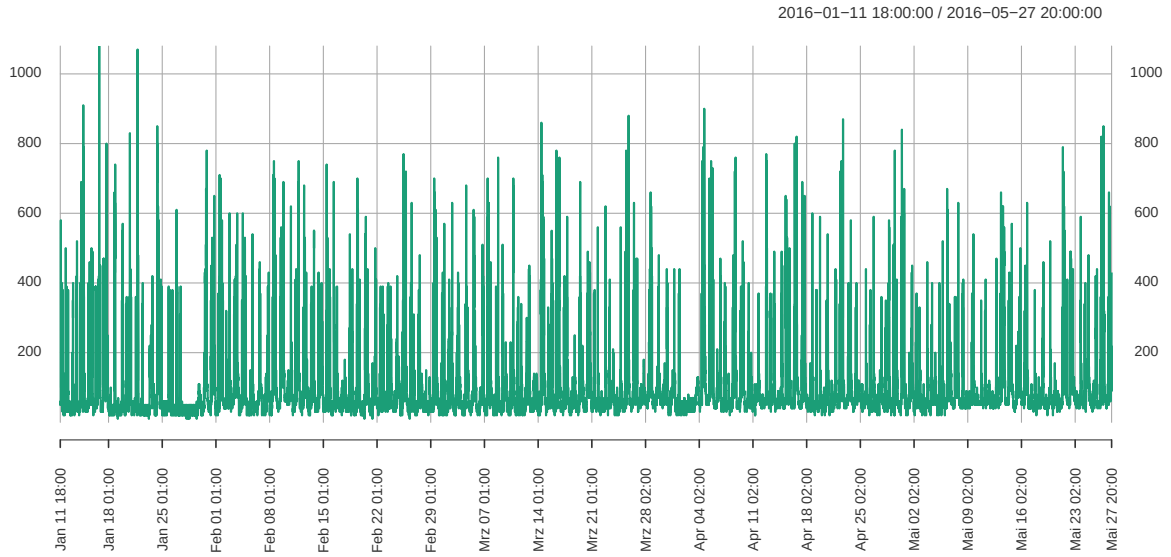



Figure 6.4: `app` time series for January to May 2016.

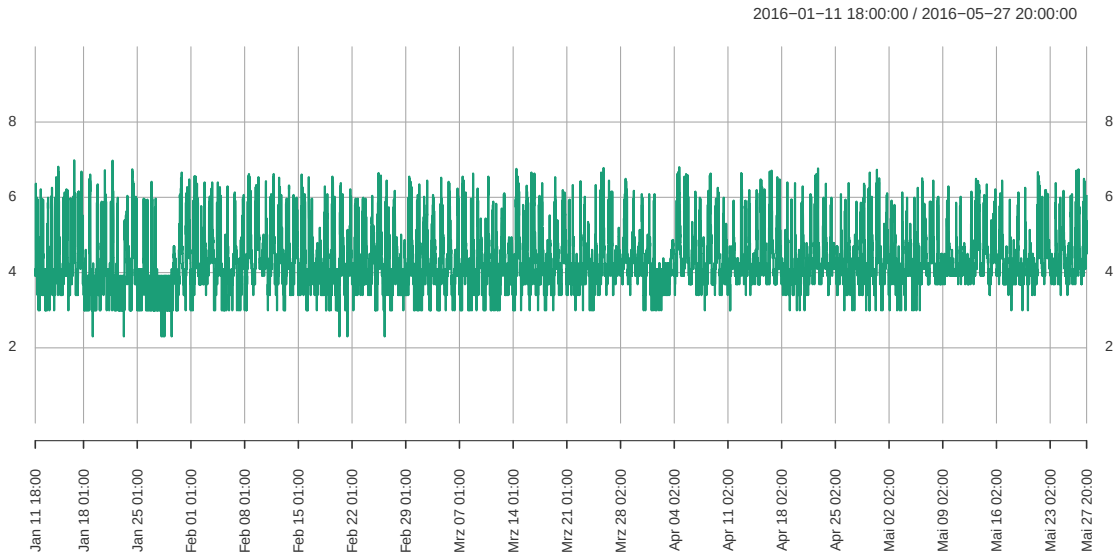



Figure 6.5: `lapp` time series for January to May 2016.

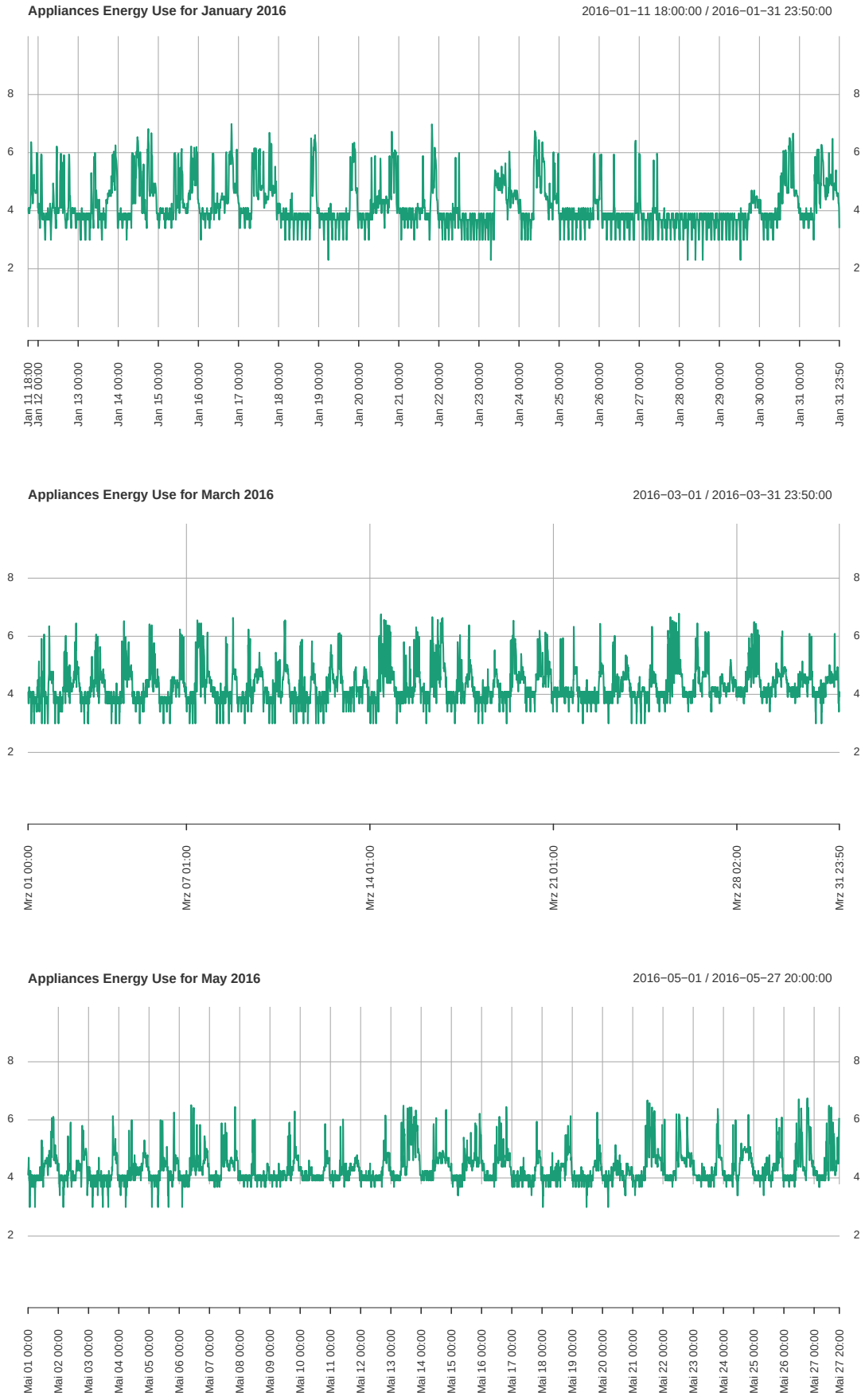



Figure 6.6: Time series of `lapp` for the months January, March and May 2016.

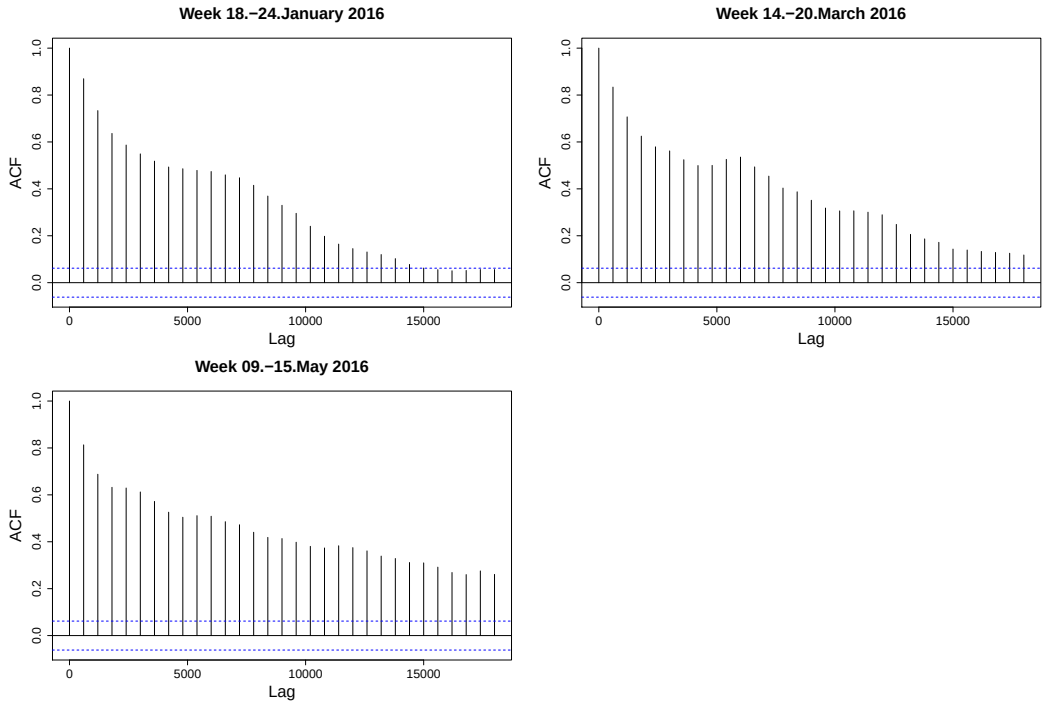



Figure 6.7: Correlogram. The weekly autocorrelation plots of series `lapp` by lags, corresponding to Figure 6.9. The dashed lines around zero showing the statistically significance level ($\alpha = 0.05$).

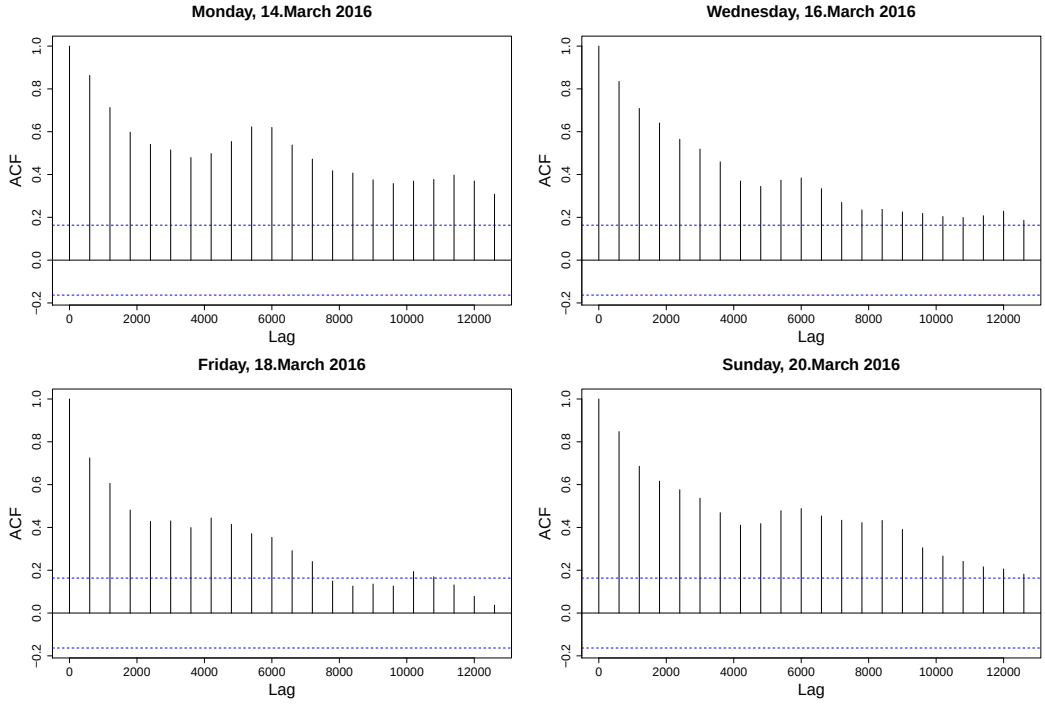



Figure 6.8: Correlogram. The daily autocorrelation plots of series `lapp` by lags, corresponding to Figure 6.10. The dashed lines around zero showing the statistically significance level ($\alpha = 0.05$).

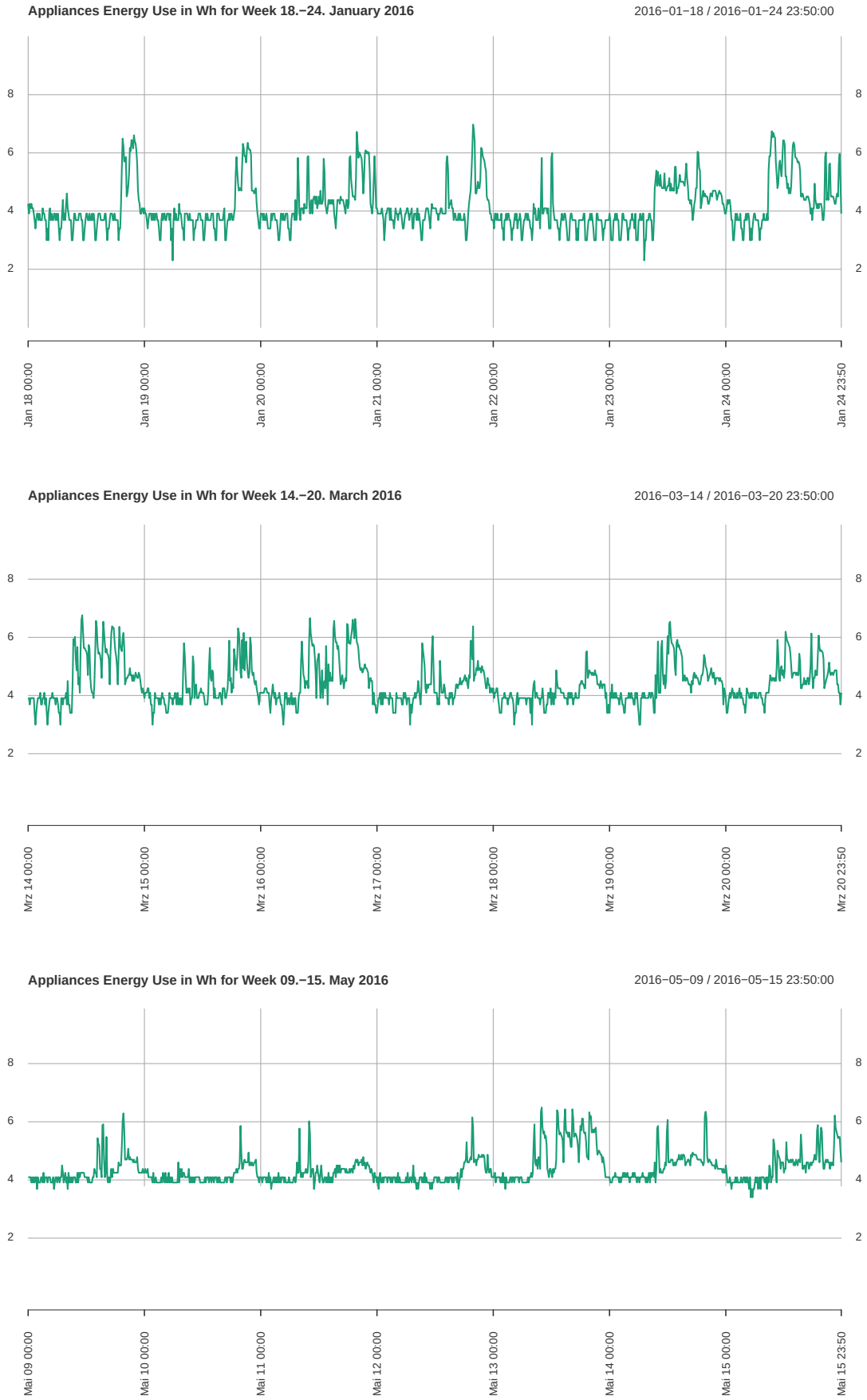



Figure 6.9: Time series of `lapp` for three equidistant weeks, where the week is starting on Monday and ending on Sunday.


Figure 6.10: Time series of variable `lapp` for four weekdays in March (Monday, Wednesday, Friday and Sunday).

### 6.1.2 Covariates

After the explorative analysis of the response variable, we are now studying our covariates, the temperatures and humidities. We analyze the temperature covariates first, then the humidity covariates.

#### Histogram of room temperatures

Starting with the histograms of the temperature covariates, Figure 6.11 shows that the different rooms have different properties in their distributions. On the one hand, the kitchen area and office room have a clear uni-modal distribution and symmetric character, whereas the living room temperatures are right skewed since the distribution is more centered on the lower °C-values and so the laundry and bathroom temperatures but a higher variability. On the other side the ironing and parents room temperatures also have multi-modal distribution.

#### Time series of temperature classified by area

Figure 6.13 shows a slightly rising temperature towards the summer months. In addition, it should be noted that the temperature in the house are relatively constant in their small variability, which supports the low-energy house model. Compare the minimum, maximum and mean of the temperature variables in Table 5.2 and the following expectations which supports the result.

$$
\begin{aligned}
&E[\texttt{T1.kitchen}] = 21.687, && E[\texttt{T2.living}] = 20.341,\\
&E[\texttt{T3.laundry}] = 22.268, && E[\texttt{T4.office}] = 20.855,\\
&E[\texttt{T5.bath}] = 19.592, && E[\texttt{T7.ironing}] = 20.267,\\
&E[\texttt{T8.teenager}] = 22.029, && E[\texttt{T9.parents}] = 19.486,\\
&E[\texttt{T6.outside}] = 7.911, && E[\texttt{T.outstation}] = 7.412.
\end{aligned}
$$

In a low-energy house like our model house, the temperature differences are kept to a minimum. This can be argued with the help of the house description in the paper of (Candanedo et al., 2017).

On the other hand, the fluctuations tell us that the occupants are in the area using the appliances.

The last two plots representing the outside temperatures show the much higher variability and a clear temperature rising towards June.

In summary it shows a small positive time effect. This conclusion is endorsed by the corresponding autocorrelation plots in Figure 6.12, where we have highly positive correlation between the time lags, i.e. with rising or falling temperature, the trend of rising or falling will continue, respectively.

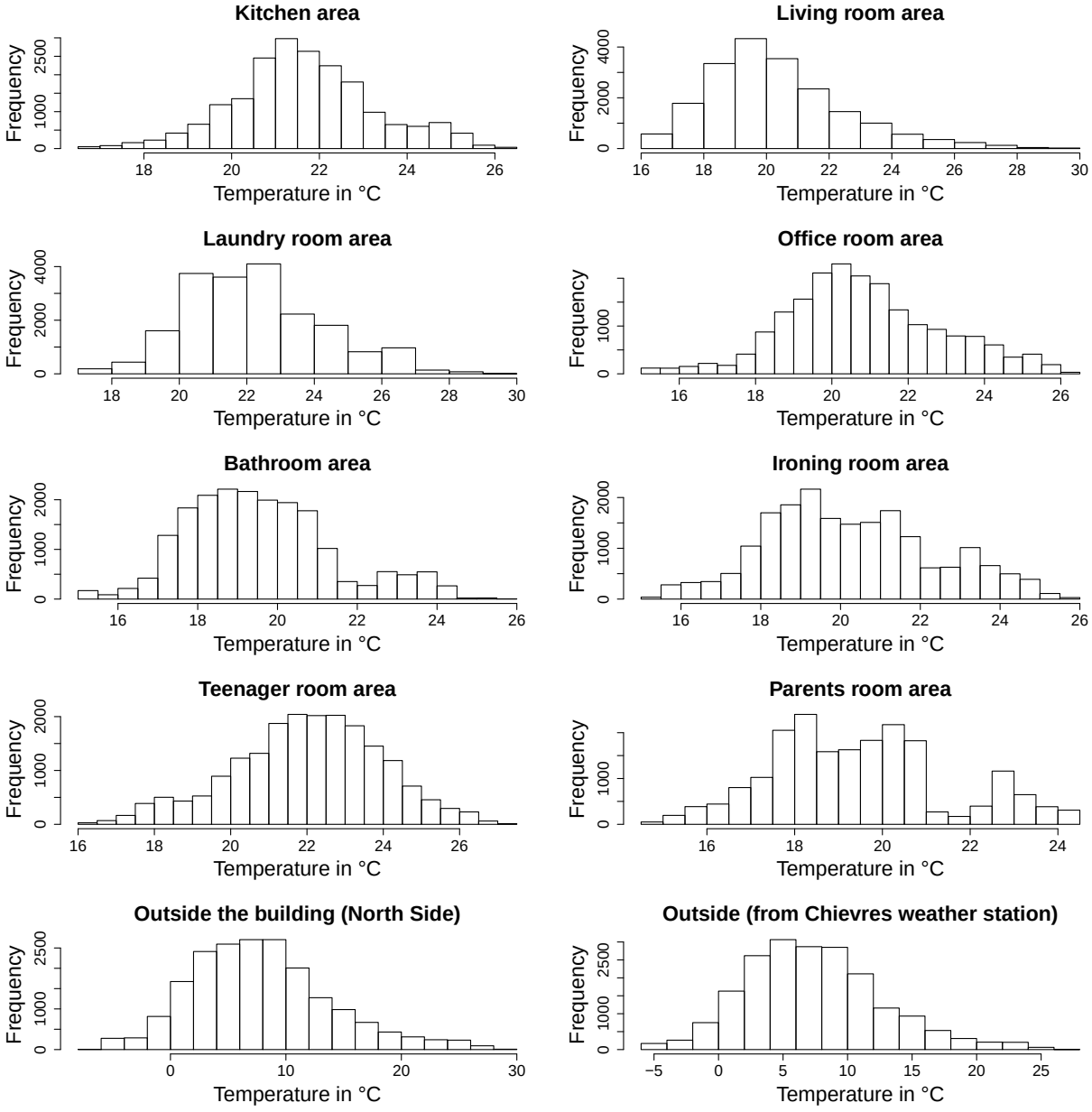



Figure 6.11: Temperature distribution. Histogram of all the temperature variables in °C.

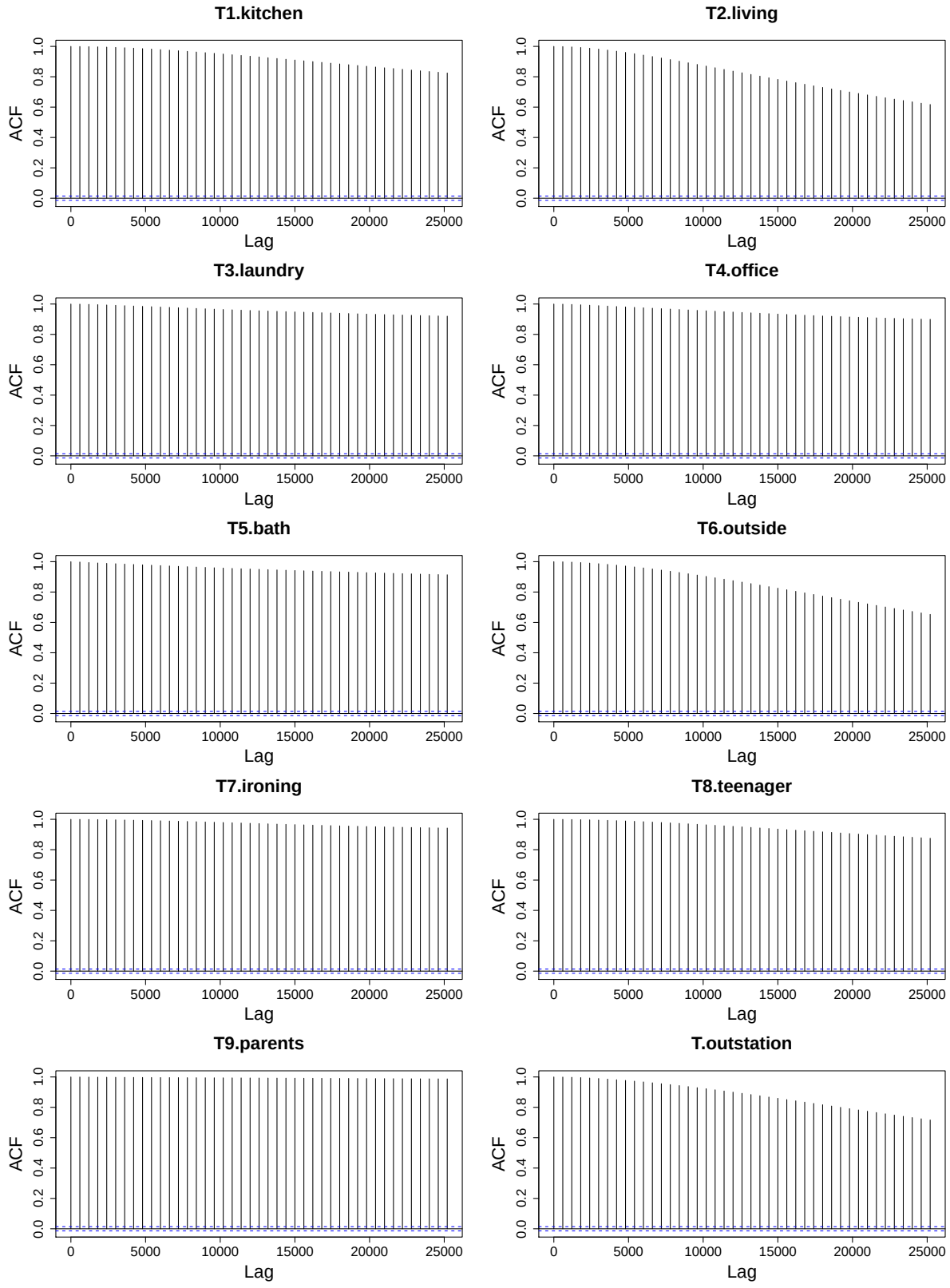



Figure 6.12: Correlogram. The autocorrelation plots of the given area temperature series by lags. The dashed lines around zero showing the statistically significance level ($\alpha = 0.05$).

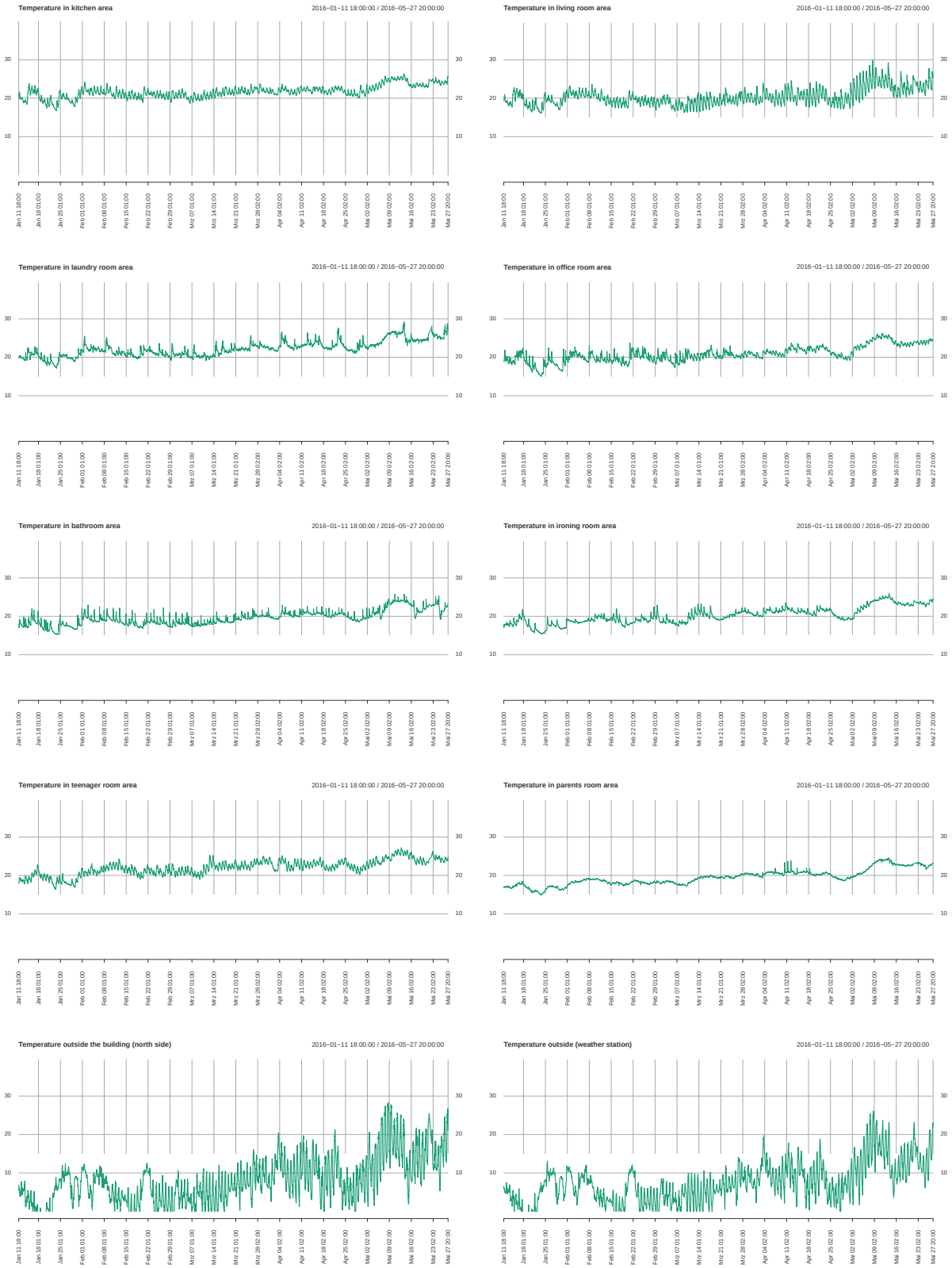



Figure 6.13: Time series of the temperature variable in °C over the entire period.

### Histogram of room humidities

Now we examine the distribution of the humidities. Also for the humidities we have different properties for the different rooms. All the inside humidities showing unimodal distributions. But for the laundry room, there is another very small peak which indicates the higher appliances use by the occupants in the room which rises the humidity. Overall we can say that the histograms in Figure 6.14 are slightly skewed.

The distribution which represents the outside the building humidity measurements is multimodal with peaks in the intervals $0-5\%$, $50-55\%$ and $95-100\%$. Whereas the outside humidity near the weather station takes all values of the humidity range, but only has one peak at $90-95\%$. So the distribution of outside the building could be influenced by the general outside humidity and also from the inside humidities and its isolation of the low energy house.

### Time series of humidity classified by area

Almost the same observation, as we had for the temperature time series, we can tell about the humidity time series, but with a downward trend. It can be explained by the well known psychrometric chart (see A.1) in the subject thermo dynamics. The higher the temperature, the more absolute humidity can be absorbed. So if the temperature goes up with a constant absolute humidity value, the relative humidity goes down. This statement can be deduced from the thermo dynamics lecture notes from Sattelmayer (2008, chapter 7).

Looking at Figure 6.16, the time series is relatively flat, except for the bathroom and outside humidity. In summary it shows a very small time effect. To support the statement, compare minimum, maximum and mean values of the Table 5.2 and following list of expectations.

$$
\begin{aligned}
&E[\texttt{RH1.kitchen}] = 40.260, && E[\texttt{RH2.living}] = 40.420,\\
&E[\texttt{RH3.laundry}] = 39.243, && E[\texttt{RH4.office}] = 39.027,\\
&E[\texttt{RH5.bath}] = 50.949, && E[\texttt{RH7.ironing}] = 35.388,\\
&E[\texttt{RH8.teenager}] = 42.936, && E[\texttt{RH9.parents}] = 41.552,\\
&E[\texttt{RH6.outside}] = 54.609, && E[\texttt{RH.outstation}] = 79.750.
\end{aligned}
$$

The correlogram in Figure 6.15 of the time series also supports our findings, i.e. positive autocorrelations and a more rapidly falling autocorrelations for the bathroom humidity `RH5.bath` and the outside humidity `RH.outstation`.

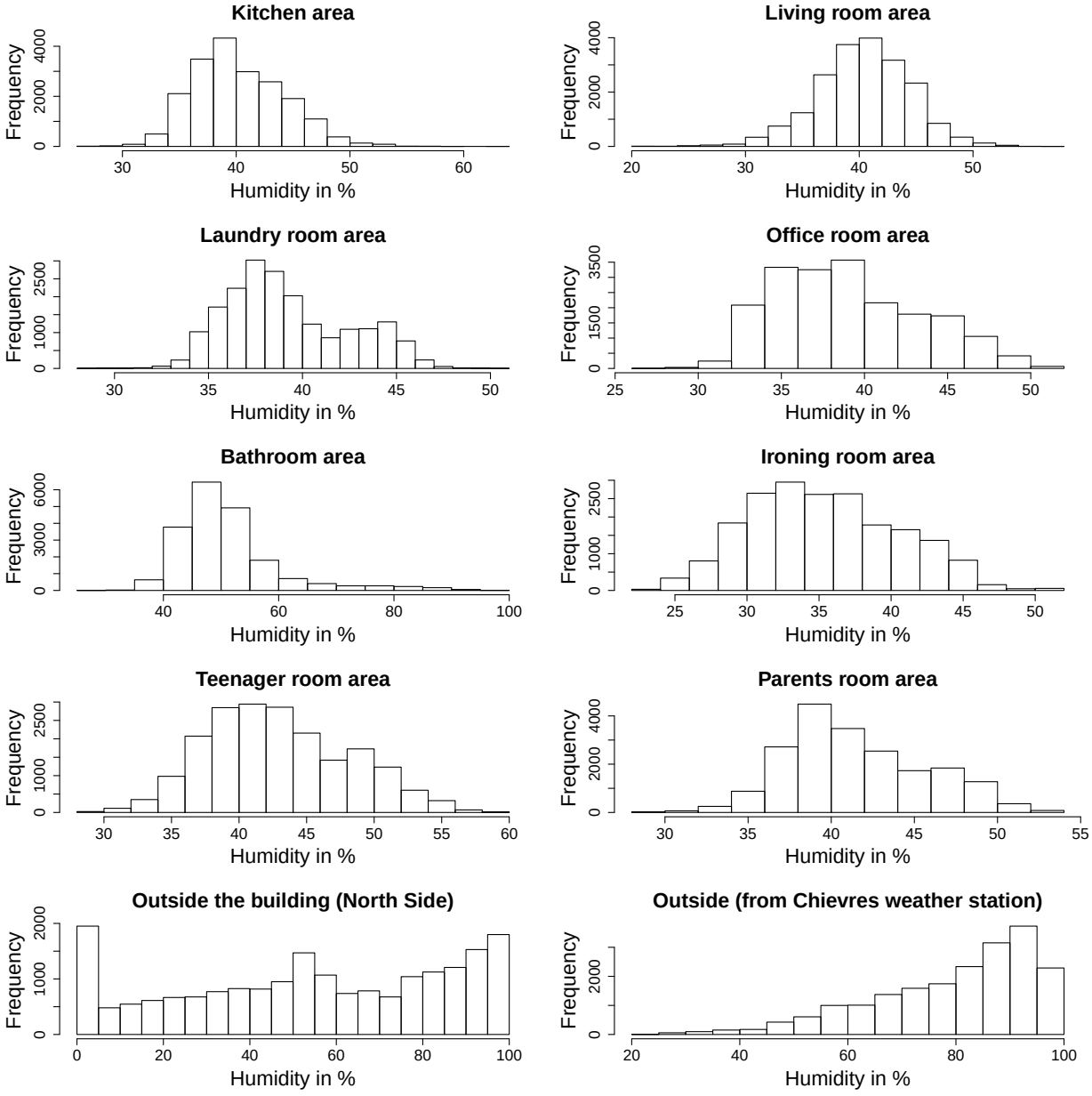



Figure 6.14: Humidity distribution. Histogram of all the humidity variables in %.

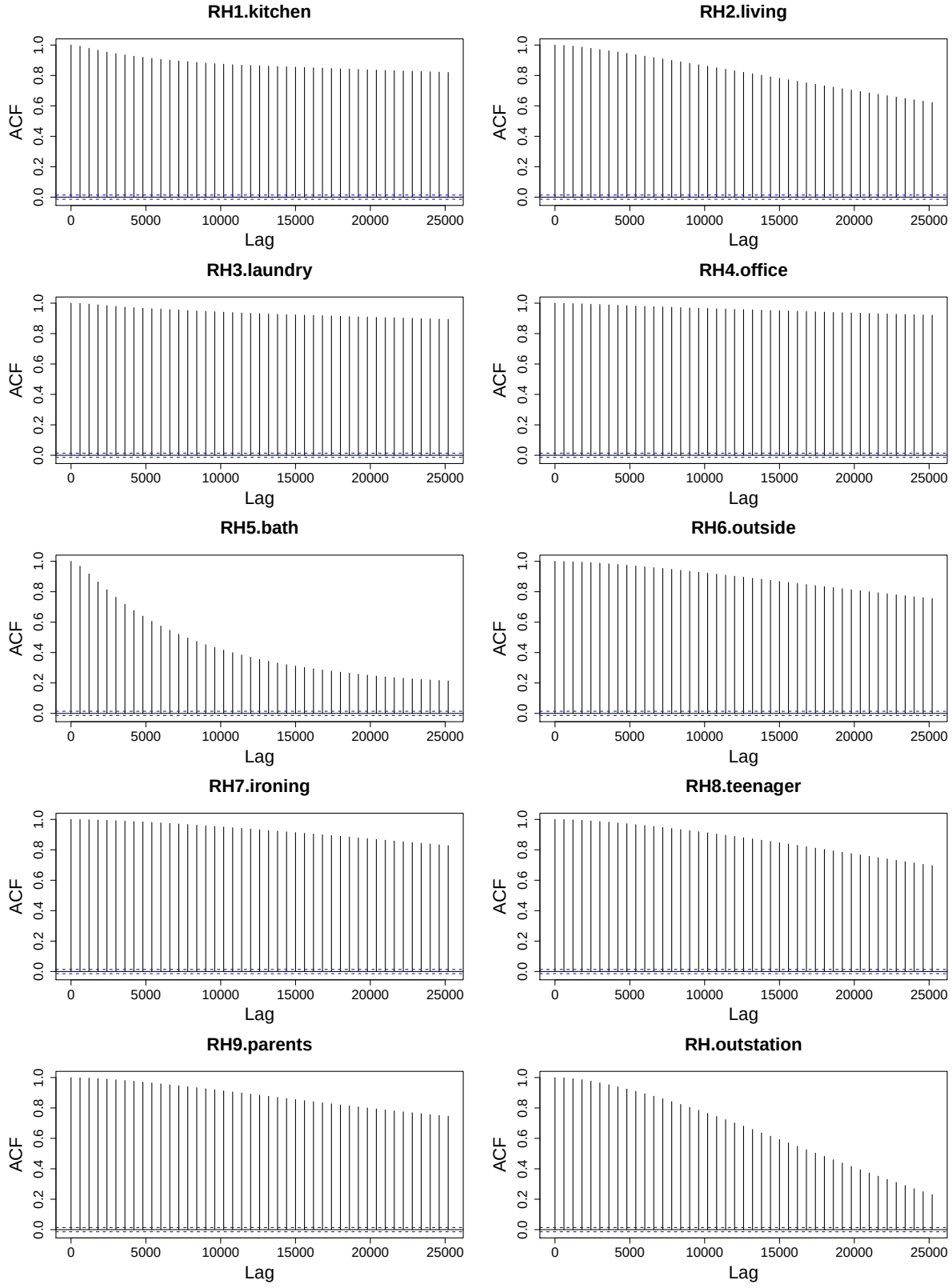



Figure 6.15: Correlogram. The autocorrelation plots of the given area humidity series by lags. The dashed lines around zero showing the statistically significance level ($\alpha = 0.05$).

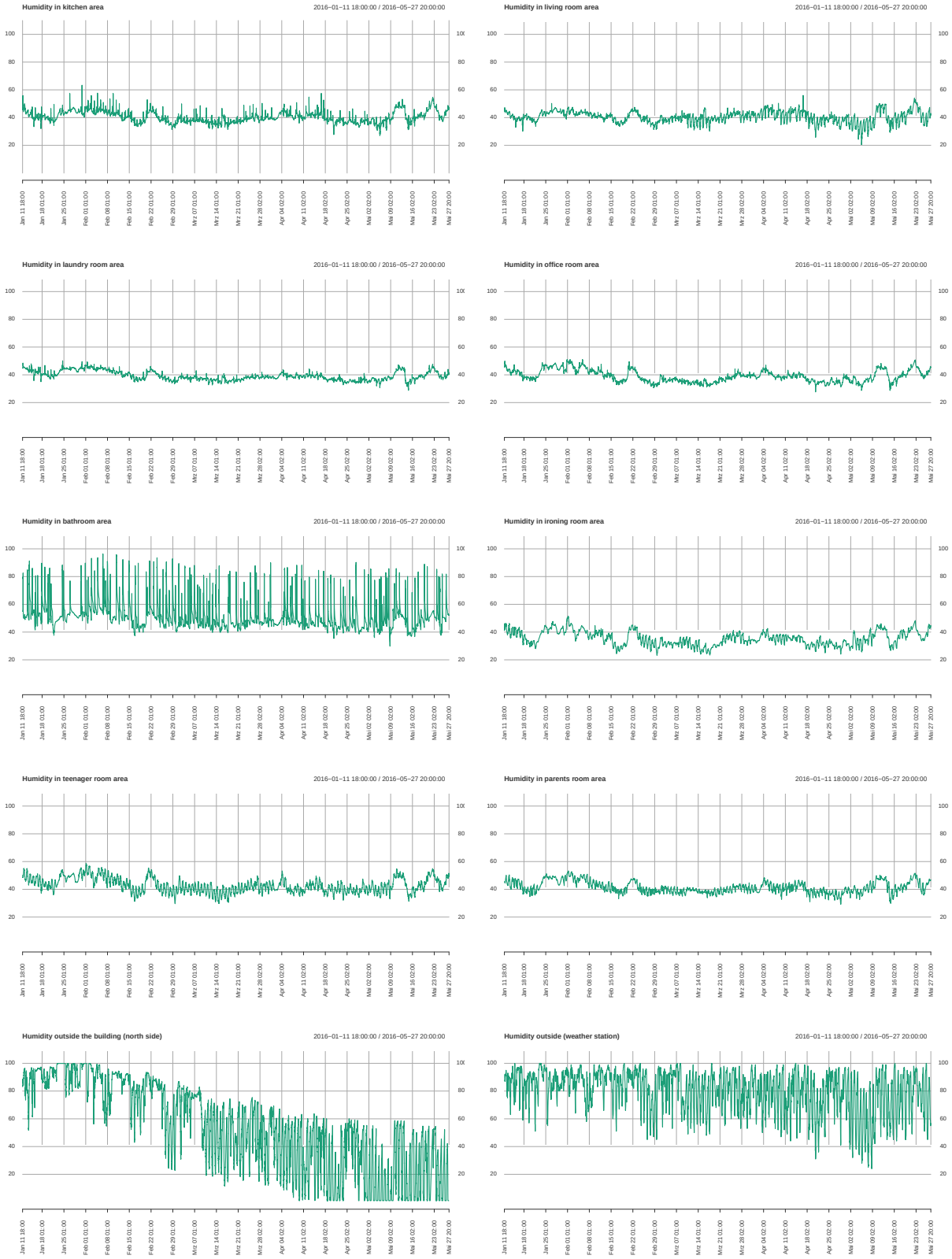



Figure 6.16: Time series of all the humidity variable in % over the entire period.

## 6.2 Pairwise exploration

The Figure 6.17 and 6.18 show bivariate scatter-plots of response `lapp` versus the ten temperature and humidity, respectively, with a locally weighted fitted line to detect linear and non-linear relationships.

The relationship of

$$\begin{aligned}\texttt{lapp}_i \sim \texttt{Tj}_i, \quad & j = \texttt{1.kitchen}, \ldots, \texttt{9.parents}, \texttt{.outstation}, \\ & i = 1, ..., 19735,\end{aligned}$$

show small positive influences (cf. Figure 6.17) which are approximately linear. but there are some non-linear curves looking at the red lines in the scatter-plots, but mostly at the limits. This can happen due to the small density at the boundaries. It explains the rising slope at the tails comparing to the center of the fitted line with a higher density of data points. The same we can observe at the beginning of the red fitted line.

The following table verifies our results. We use the Pearson correlation coefficient $\rho_{X,Y} = \frac{Cov(X,Y)}{\sqrt{Var(X)}\sqrt{Var(Y)}}$ or equivalently for a sample $\hat{\rho}_{x,y} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}}$, we defined in (2.2).

$$\begin{aligned}\hat{\rho}_{\texttt{lapp,T1}} &= 0.161, & \hat{\rho}_{\texttt{lapp,T6}} &= 0.197, \\ \hat{\rho}_{\texttt{lapp,T2}} &= 0.215, & \hat{\rho}_{\texttt{lapp,T7}} &= 0.110, \\ \hat{\rho}_{\texttt{lapp,T3}} &= 0.167, & \hat{\rho}_{\texttt{lapp,T8}} &= 0.153, \\ \hat{\rho}_{\texttt{lapp,T4}} &= 0.132, & \hat{\rho}_{\texttt{lapp,T9}} &= 0.093, \\ \hat{\rho}_{\texttt{lapp,T5}} &= 0.110, & \hat{\rho}_{\texttt{lapp,Tout}} &= 0.176.\end{aligned}$$

Contrary to Figure 6.17, in the scatter-plots of Figure 6.18 we can see small negative approximately linear dependencies in the relationship

$$\begin{aligned}\texttt{lapp}_i \sim \texttt{RHj}_i, \quad & j = \texttt{1.kitchen}, \ldots, \texttt{9.parents}, \texttt{.outstation}, \\ & i = 1, ..., 19735,\end{aligned}$$

except for the kitchen area and bathroom. This supports our belief that humidity increases with the occupancy of the kitchen, such as using appliances for cooking, and the appliances use in the bathroom.

The other observations of an small negative correlation between `lapp` and the humidities are also logically, due to the increasing temperature which leads to an greater capacity of humidity in the air and that results in a decreasing percentage of humidity. The conclusion is made due to the psychometric chart, see Figure A.1 in the Appendix.

On the other hand, looking at the sample correlations, we see only correlations for $\hat{\rho}_{\texttt{lapp,RH6}}$, $\hat{\rho}_{\texttt{lapp,RH8}}$, $\hat{\rho}_{\texttt{lapp,RH9}}$ and $\hat{\rho}_{\texttt{lapp,RHout}}$, since these results score correlations of $|\hat{\rho}_{x,y}| > 0.1$. The rest is approximately uncorrelated, which is supported virtually in Figure 6.18.

$$\begin{aligned}\hat{\rho}_{\texttt{lapp,RH1}} &= 0.084, & \hat{\rho}_{\texttt{lapp,RH6}} &= -0.174, \\ \hat{\rho}_{\texttt{lapp,RH2}} &= -0.094, & \hat{\rho}_{\texttt{lapp,RH7}} &= -0.096, \\ \hat{\rho}_{\texttt{lapp,RH3}} &= -0.006, & \hat{\rho}_{\texttt{lapp,RH8}} &= -0.165, \\ \hat{\rho}_{\texttt{lapp,RH4}} &= -0.007 & \hat{\rho}_{\texttt{lapp,RH9}} &= -0.116, \\ \hat{\rho}_{\texttt{lapp,RH5}} &= 0.024 & \hat{\rho}_{\texttt{lapp,RHout}} &= -0.226\end{aligned}$$

Note that we have the same non-linear limit behavior as we had for the relations between `lapp` and the temperature covariates.

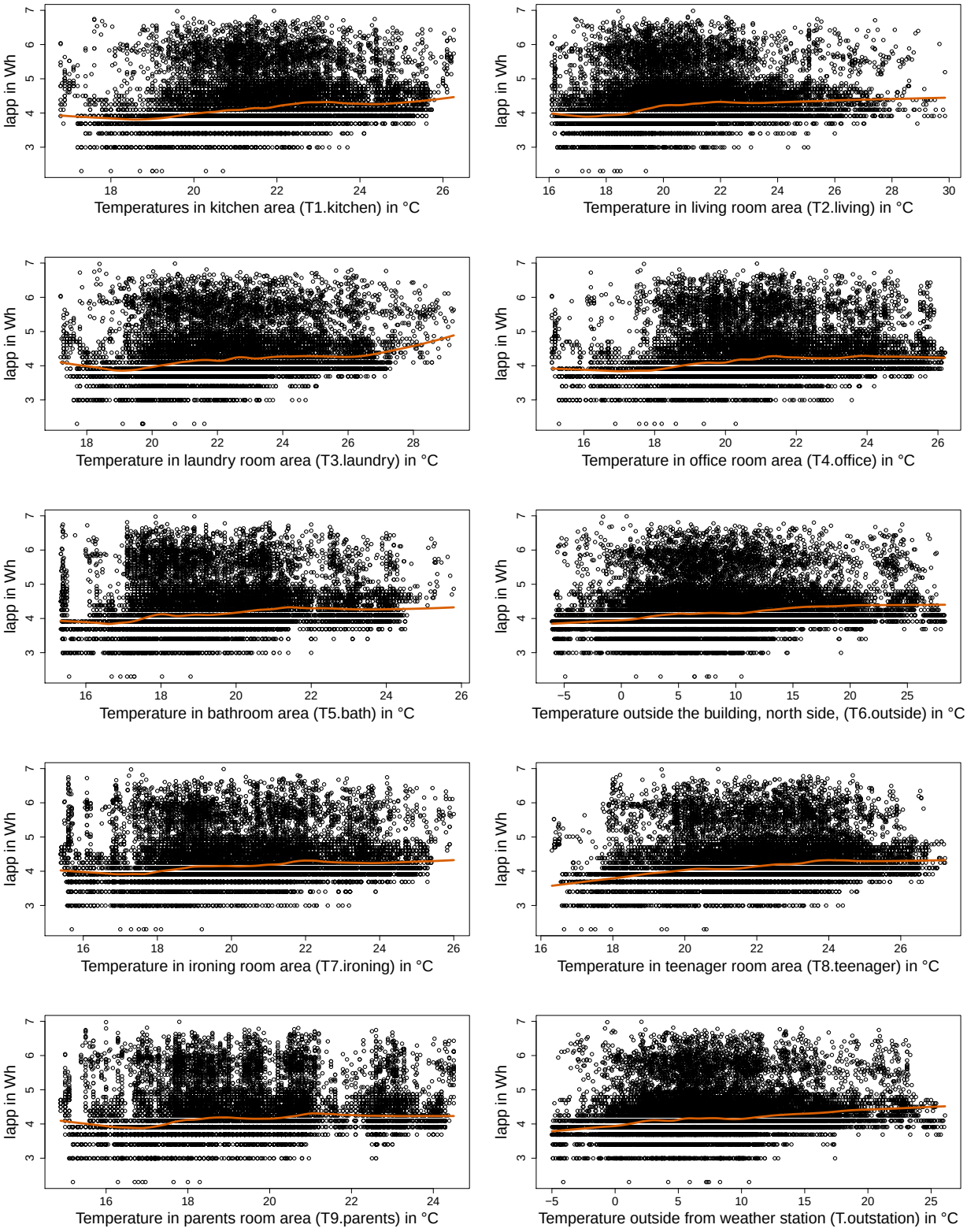



Figure 6.17: Pairwise scatter-plots. Relationship of `lapp` $\sim$ `Tj` with `Tj` being the temperature variables for $j =$ `1.kitchen`, .., `9.parents`, `.outstation`. The red fitted line is created by using the method of *lowess* with the smoothing parameter $f \equiv 0,2$.

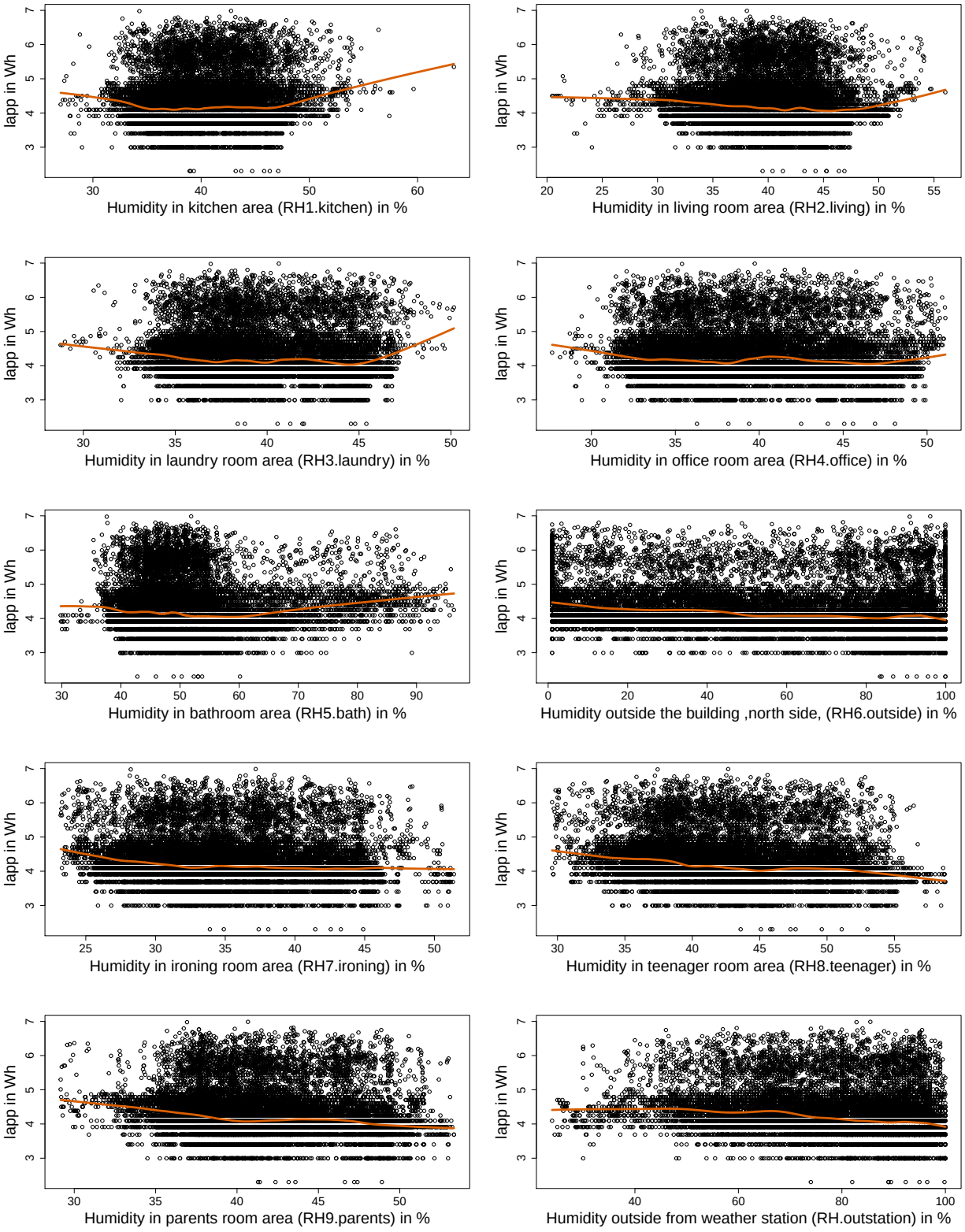



Figure 6.18: Pairwise scatter-plots. Relationship of `lapp` $\sim$ `RHj` with `RHj` being the humidity variables for $j =$ `1.kitchen`, .., `9.parents`, `.outstation`. The red fitted line is created by using the method of *lowess* with the smoothing parameter $f \equiv 0,2$.

## 6.3 Analyzing pattern over time

Finally, we want to study the behavior of the energy use depending on the time. The goal is to detect pattern based on the months, weekdays and hours so that we can include the time component in a suitable form in our model.

### 6.3.1 Box-plots

#### Monthly pattern

Observing the months separately, the lowest median we have for the month January whereas the other months have the same median values. So we have similar pattern on monthly bases, (cf. Figure 6.19).

| Month | *Median* | *Variance* |
|---|---|---|
| January | 3.91 | 0.634 |
| February | 4.09 | 0.456 |
| March | 4.09 | 0.415 |
| April | 4.09 | 0.383 |
| May | 4.09 | 0.303 |

$$\Rightarrow median(\texttt{lapp}_{January}) < median(\texttt{lapp}_{February}) = median(\texttt{lapp}_{March}) = median(\texttt{lapp}_{April}) = median(\texttt{lapp}_{May})$$

| Month | Quantiles | | | | |
|---|---|---|---|---|---|
| | 0.00 | 0.25 | 0.50 | 0.75 | 1.00 |
| January | 2.30 | 3.69 | 3.91 | 4.50 | 6.98 |
| February | 2.30 | 3.91 | 4.09 | 4.61 | 6.65 |
| March | 3.00 | 3.91 | 4.09 | 4.61 | 6.78 |
| April | 3.00 | 3.91 | 4.09 | 4.61 | 6.80 |
| May | 3.00 | 3.91 | 4.09 | 4.61 | 6.75 |

Table 6.1: Quantiles of the monthly pattern using `lapp`. Corresponding to the Figure 6.19.

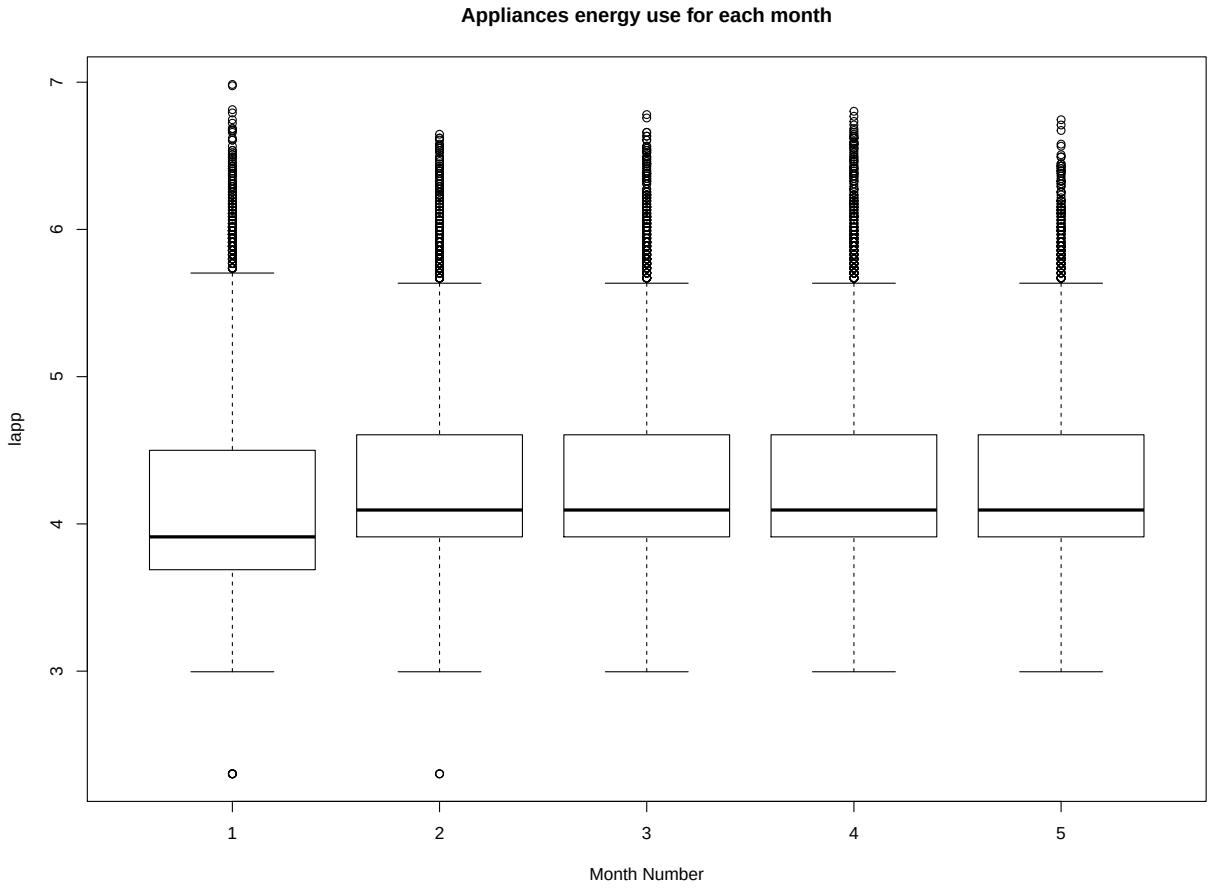



Figure 6.19: Box-plot for monthly pattern using `lapp`. The month number represents January, February, March, April and May, respectively.

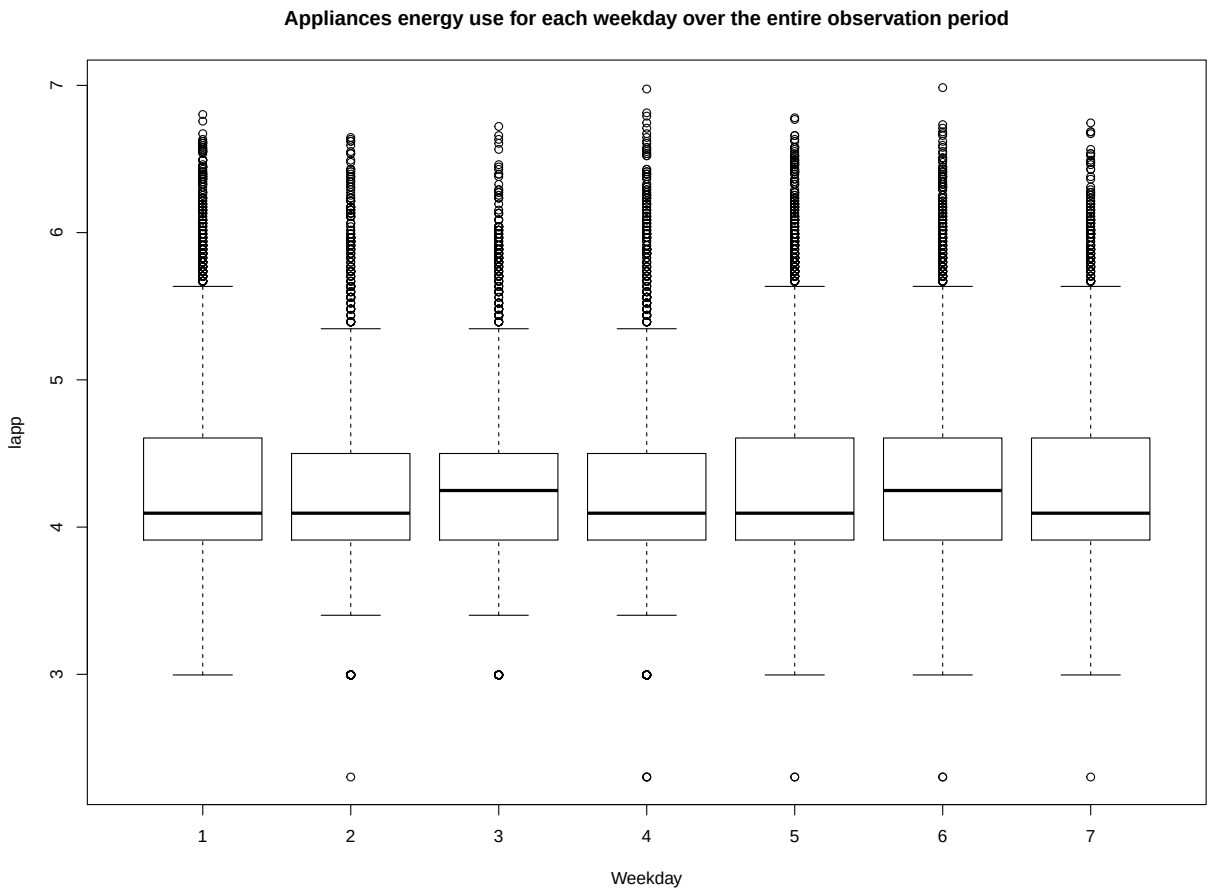



Figure 6.20: Box-plot for daily pattern using `lapp`. The day number represents the weekday. 1 = Monday, 2 = Tuesday, 3 = Wednesday, 4 = Thursday, 5 = Friday, 6 = Saturday, 7 = Sunday.

### Week-daily pattern

The Figure 6.20 shows a bigger time effects based on weekdays compared to the months. But there are no great differences in the week-daily pattern, but looking more into it, median values vary between two values. The highest median is observed for Wednesdays and Saturdays.

| Weekday | *Median* | *Variance* |
|---|---|---|
| Monday | 4.09 | 0.518 |
| Tuesday | 4.09 | 0.378 |
| Wednesday | 4.25 | 0.337 |
| Thursday | 4.09 | 0.393 |
| Friday | 4.09 | 0.538 |
| Saturday | 4.25 | 0.468 |
| Sunday | 4.09 | 0.364 |

$$\begin{aligned} \Rightarrow median(\texttt{lapp}_{Monday}) &= median(\texttt{lapp}_{Tuesday}) \\ &= median(\texttt{lapp}_{Thursday}) \\ &= median(\texttt{lapp}_{Friday}) \\ &= median(\texttt{lapp}_{Sunday}) \\ &< median(\texttt{lapp}_{Wednesday}) \\ &= median(\texttt{lapp}_{Saturday}) \end{aligned}$$

| Weekday | Quantiles | | | | |
|---|---|---|---|---|---|
| | 0.00 | 0.25 | 0.50 | 0.75 | 1.00 |
| Monday | 3.00 | 3.91 | 4.09 | 4.61 | 6.80 |
| Tuesday | 2.30 | 3.91 | 4.09 | 4.50 | 6.65 |
| Wednesday | 3.00 | 3.91 | 4.25 | 4.50 | 6.72 |
| Thursday | 2.30 | 3.91 | 4.09 | 4.50 | 6.98 |
| Friday | 2.30 | 3.91 | 4.09 | 4.61 | 6.78 |
| Saturday | 2.30 | 3.91 | 4.25 | 4.61 | 6.98 |
| Sunday | 2.30 | 3.91 | 4.09 | 4.61 | 6.75 |

Table 6.2: Quantiles of the week-daily pattern using `lapp`. Corresponding to the Figure 6.20.

### Hourly pattern

Finally examine the `lapp` depending on the hours, we have the greatest time effect. In Figure 6.21 we see a first increase of energy consumption in the midday hours, i.e. at hour 10, to the first peak during the early afternoon hours, i.e. at hour 14 and 15. The second and highest peak is in the evening hours 20 and 21. So we can say that the residents intensify the use of their equipment in the twentieth hour.

| Hour | *Median* | *Variance* |
|---|---|---|
| 1 | 3.91 | 0.1125 |
| 2 | 3.91 | 0.1049 |
| 3 | 3.91 | 0.0688 |
| 4 | 3.91 | 0.0639 |
| 5 | 3.91 | 0.0746 |
| 6 | 3.91 | 0.0647 |
| 7 | 3.91 | 0.0615 |
| 8 | 3.91 | 0.1459 |
| 9 | 3.91 | 0.2234 |
| 10 | 4.09 | 0.4652 |
| 11 | 4.25 | 0.5204 |
| 12 | 4.25 | 0.5365 |
| 13 | 4.25 | 0.6354 |
| 14 | 4.38 | 0.5496 |
| 15 | 4.38 | 0.5145 |
| 16 | 4.25 | 0.4900 |
| 17 | 4.25 | 0.4111 |
| 18 | 4.38 | 0.3716 |
| 19 | 4.50 | 0.4940 |
| 20 | 4.79 | 0.6096 |
| 21 | 4.79 | 0.3933 |
| 22 | 4.70 | 0.2284 |
| 23 | 4.61 | 0.1883 |
| 24 | 4.25 | 0.1727 |

$$
\begin{aligned}
\Rightarrow median(\texttt{lapp}_{hour=1}) &= median(\texttt{lapp}_{hour=2}) \\
&= \dots \\
&= median(\texttt{lapp}_{hour=9}) \\
&< median(\texttt{lapp}_{hour=10}) \\
&< median(\texttt{lapp}_{hour=11}) \\
&= median(\texttt{lapp}_{hour=12}) \\
&= median(\texttt{lapp}_{hour=13}) \\
&= median(\texttt{lapp}_{hour=16}) \\
&= median(\texttt{lapp}_{hour=17}) \\
&= median(\texttt{lapp}_{hour=24}) \\
&< median(\texttt{lapp}_{hour=14}) \\
&= median(\texttt{lapp}_{hour=15}) \\
&= median(\texttt{lapp}_{hour=18}) \\
&< median(\texttt{lapp}_{hour=19}) \\
&< median(\texttt{lapp}_{hour=23}) \\
&< median(\texttt{lapp}_{hour=22}) \\
&< median(\texttt{lapp}_{hour=20}) \\
&= median(\texttt{lapp}_{hour=21})
\end{aligned}
$$

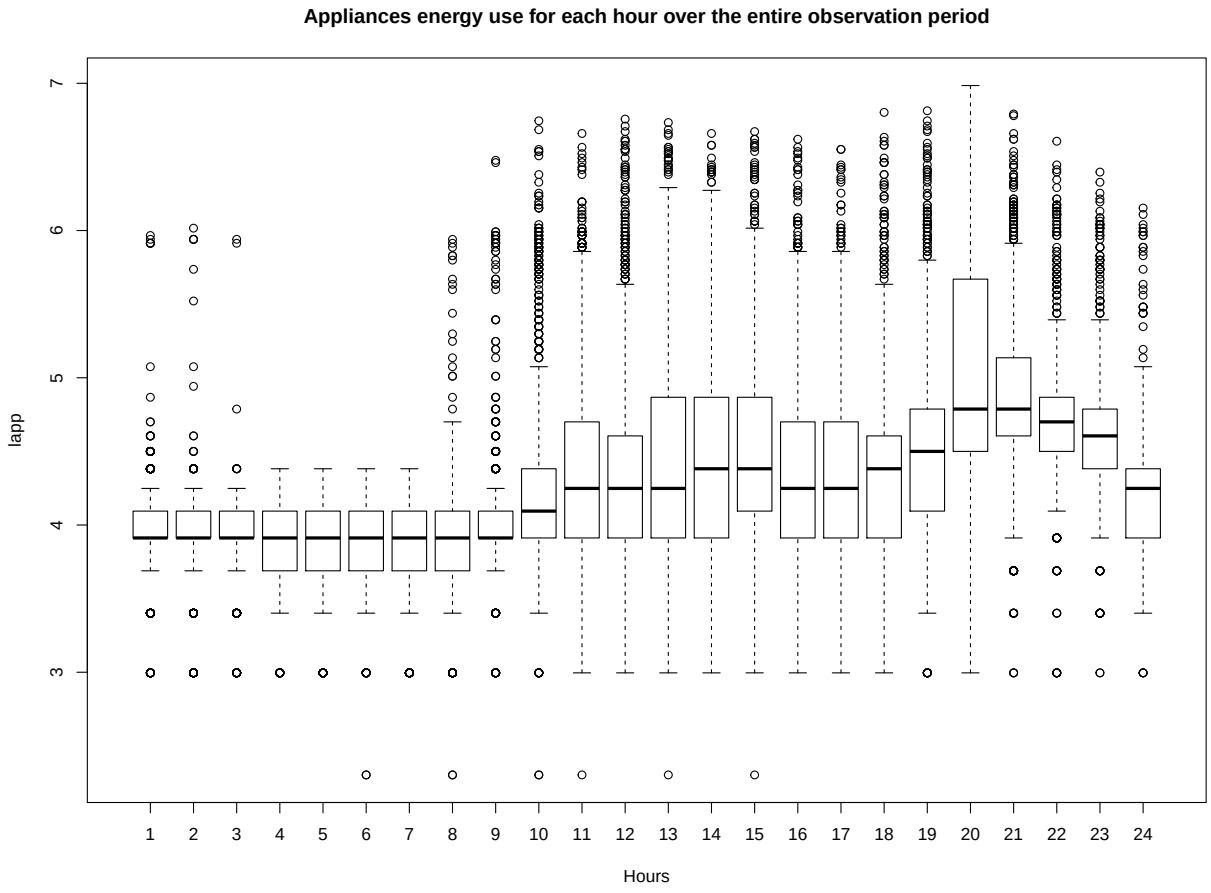



Figure 6.21: Box-plot for hourly pattern using `lapp`. The day number represents the hours, from the first hour to the 24-th hour.

| Hour | Quantiles | | | | |
|---|---|---|---|---|---|
| | 0.00 | 0.25 | 0.50 | 0.75 | 1.00 |
| 1 | 3.00 | 3.91 | 3.91 | 4.09 | 5.97 |
| 2 | 3.00 | 3.91 | 3.91 | 4.09 | 6.02 |
| 3 | 3.00 | 3.91 | 3.91 | 4.09 | 5.94 |
| 4 | 3.00 | 3.69 | 3.91 | 4.09 | 4.38 |
| 5 | 3.00 | 3.69 | 3.91 | 4.09 | 4.38 |
| 6 | 2.30 | 3.69 | 3.91 | 4.09 | 4.38 |
| 7 | 3.00 | 3.69 | 3.91 | 4.09 | 4.38 |
| 8 | 2.30 | 3.69 | 3.91 | 4.09 | 5.94 |
| 9 | 3.00 | 3.91 | 3.91 | 4.09 | 6.48 |
| 10 | 2.30 | 3.91 | 4.09 | 4.38 | 6.75 |
| 11 | 2.30 | 3.91 | 4.25 | 4.70 | 6.66 |
| 12 | 3.00 | 3.91 | 4.25 | 4.61 | 6.76 |
| 13 | 2.30 | 3.91 | 4.25 | 4.87 | 6.73 |
| 14 | 3.00 | 3.91 | 4.38 | 4.87 | 6.66 |
| 15 | 2.30 | 4.09 | 4.38 | 4.87 | 6.67 |
| 16 | 3.00 | 3.91 | 4.25 | 4.70 | 6.62 |
| 17 | 3.00 | 3.91 | 4.25 | 4.70 | 6.55 |
| 18 | 3.00 | 3.91 | 4.38 | 4.61 | 6.80 |
| 19 | 3.00 | 4.09 | 4.50 | 4.79 | 6.81 |
| 20 | 3.00 | 4.50 | 4.79 | 5.67 | 6.98 |
| 21 | 3.00 | 4.61 | 4.79 | 5.14 | 6.79 |
| 22 | 3.00 | 4.50 | 4.70 | 4.87 | 6.61 |
| 23 | 3.00 | 4.38 | 4.61 | 4.79 | 6.40 |
| 24 | 3.00 | 3.91 | 4.25 | 4.38 | 6.15 |

Table 6.3: Quantiles of the hourly pattern using `lapp`. Corresponding to the Figure 6.21.

### Comparing time pattern

In conclusion the weekdays and the hours have to be taken into account regarding the regression model, since we detected some pattern which should not be ignored. Connect the median lines in the box-plot based on the hour (c.f. Figure 6.21), yield a polynomial of an approximate third or fourth degree. The same is obtained when drawing a line though the median values of the box-plot based on the weekdays in Figure 6.20. On the other hand, it is more promising to integrate the weekdays as factors. Not only we have just seven unique elements, also the comparison is easier with factors.

Based on that, we will study the time effects for the hours and weekdays in the following chapter.

# Chapter 7

# Linear regression models (LM's) for energy consumption within a house

Finally we come to the ability of statistical methods to provide a mathematical equation that reflects the complexity of the relation between the response variable `lapp` and our set of explanatory variables of temperatures and humidities. Moreover, with the regression model we have tools for interpreting results, checking for the significant predictors, assessing their relative importance to the model and displaying the corresponding graphics to our analysis. We also include time effects by defining appropriate covariates. This type of model might better explain the real dynamics of the relationships and effects.

First, we use the multiple linear models for the explanatory analysis and predictive challenging tasks to characterize relevant factors that impact the appliance energy uses.

Furthermore we do not take the variables `Pressure`, `Windspeed`, `Visibility`, and `Tdewpoint` into account, which we introduced in Table 5.2. We want to focus on the impact of temperatures and humidity in the house. Additionally, putting the other covariates in a linear regression with our response variable `lapp` results in a total variability explanation of just 1%.

## 7.1 Setting the time effects

In the last chapter we already analyzed pattern over time and concluded that there exists some time effects, that is the hourly and week-daily effects. To work on our data set we converted it into a xts-object, which makes it easy to handle the time series, extract time points and intervals and review the periodicity of the observations. (c.f. Ryan and Ulrich (2011))

To get a deeper overview on the periodicity, the Table 7.1 is presented with about 6 observations available per hour, which means 144 observations available for each day.

| data set from 2016-01-11 18:00:00 to 2016-05-27 20:00:00 with/containing: | | | |
|---|---|---|---|
| periodicity | total number of days | observations per days | observations per hour |
| 10 minute | 138 | 144 | 6 |

Table 7.1: The table contains a the periodicity of our energy use data set.

With the provided tools of an xts-object, we add three columns of different frequencies to our data set, that is the covariates

(i) `month` $= (\{1, \ldots, 5\}^i)_{i=1,\ldots 19735}$,

(ii) `weekday` $= (\{1, \ldots, 7\}^i)_{i=1,\ldots 19735}$ and

(iii) `hour` $= (\{1, \ldots, 24\}^i)_{i=1,\ldots 19735}$

of the corresponding observations.

### 7.1.1 Covariate of the weekday effect

For the weekday pattern, we have decided to include the weekdays as factors in our regression model to provide a better comparability between the weekdays. Hence, we set the following new covariates for our regression, i.e. a vector with values between 1 and 7:

```
weekday.Monday = as.factor(1),      weekday.Friday = as.factor(5),
weekday.Tuesday = as.factor(2),     weekday.Saturday = as.factor(6),
weekday.Wednesday = as.factor(3),   weekday.Sunday = as.factor(7),
weekday.Thursday = as.factor(4),
```

There are about 2730 to 2800 observations available for each weekday (cf. Table 7.2)

| 1: Monday | 2: Tuesday | 3: Wednesday | 4: Thursday | 5: Friday | 6: Saturday | 7: Sunday |
|---|---|---|---|---|---|---|
| 2772 | 2880 | 2880 | 2880 | 2857 | 2736 | 2730 |

Table 7.2: The data set provides a number of available observations per weekday.

### 7.1.2 Covariate of the hour effect

Next, we take care of the hourly pattern. Reviewing the analyzes of the pattern over time, we detected a polynomial curve connecting the median lines in the box-plot in Figure 6.21. Analogously to the weekday covariate, we set the hourly effect as a numerical vector with values between 1 and 24.

The covariate

$$\texttt{hour} = \texttt{as.numeric(h)}, \qquad h = 1, \ldots, 24.$$

will later be applied to the polynomial function

$$poly(\texttt{hour}, d),$$

with $d$ as the degree of the polynomial function.

Since we assume a non-linear effect for the variable `hour`, it can be appropriately modeled using (orthogonal) polynomials of degree three or around three.

For our polynomial, the data set provides in total 822 observations for each hour (c.f. Table 7.3).

| hour | number of observations | hour | number of observations |
|---|---|---|---|
| 1 | 822 | 13 | 822 |
| 2 | 822 | 14 | 822 |
| 3 | 816 | 15 | 822 |
| 4 | 822 | 16 | 822 |
| 5 | 822 | 17 | 822 |
| 6 | 822 | 18 | 822 |
| 7 | 822 | 19 | 828 |
| 8 | 822 | 20 | 828 |
| 9 | 822 | 21 | 823 |
| 10 | 822 | 22 | 822 |
| 11 | 822 | 23 | 822 |
| 12 | 822 | 24 | 822 |

Table 7.3: The table showing available observation per hour.

To answer the question whether we should use the third or fourth degree polynomial, we will set the following regressions:

$$\texttt{lapp} \sim \texttt{poly(hour, 3)} \quad \texttt{versus} \quad \texttt{lapp} \sim \texttt{poly(hour, 4)},$$

Check the resulting regression models, we decide to continue with the third degree, i.e. $d = 3$, due to the significance of the three polynomial covariates. The reason to drop the case of the fourth degree polynomial is that we detected a non-significant fourth polynomial covariate. Additionally, it does not reach a higher variability explanation power.

$$R^2_{adj}(\texttt{lapp} \sim \texttt{poly(hour, 3)}) = 0.2272 = R^2_{adj}(\texttt{lapp} \sim \texttt{poly(hour, 4)})$$

In the following section we go deeper into details and the regression model.

## 7.2 Main effect models

We begin fitting our linear regression models with the focus on the main effects for every independent variable, so that we can examine the effect of one covariate on the response variable, while fixing the effect of any other independent variables, e.g. by averaging over the levels of all other variables.

A step-wise selection is used to check the significance of the covariates and the performance of the model. From model class 1 to 7 we are doing a forward selection. (c.f. Table 7.4) After reaching Model 7, we perform a backward selection from which we yield the two reduced models.

| Model Class | Main effects | | | | number of parameters | $R^2_{adj}$ |
|---|---|---|---|---|---|---|
| | temperature | humidity | weekday | hours | | |
| 1 | ✓ | – | – | – | 11 | 0.119 |
| 2 | – | ✓ | – | – | 11 | 0.173 |
| 3 | – | – | ✓ | – | 7 | 0.0071 |
| 4 | – | – | – | ✓ | 4 | 0.227 |
| 5 | ✓ | ✓ | – | – | 21 | 0.238 |
| 6 | – | – | ✓ | ✓ | 10 | 0.234 |
| 7 | ✓ | ✓ | ✓ | ✓ | 30 | 0.328 |

Table 7.4: Model settings and adjusted $R^2$ ($R^2_{adj}$) for each model studied in Section 7.2.1.

### 7.2.1 Original model formulations

In the step-wise procedure, we analyze the model classes by including the covariates presented in Table 7.4. In the following models we use simple multiple linear regression given by

$$\hat{Y}_i = \hat{\beta}_0 + \sum_{j=1}^{p} \hat{\beta}_j x_{ij}, \quad i = 1, \ldots, 19735 \tag{7.1}$$

where $p$ is the amount of used predictors concerning the model class.

**Model 1:**

In our first model we only allow for linear temperature effects on `lapp`.

$$\begin{aligned}\widehat{\texttt{lapp}}_i = \hat{\beta}_0 \; &+\hat{\beta}_1 \texttt{T1.kitchen}_i + \hat{\beta}_2 \texttt{T2.living}_i + \hat{\beta}_3 \texttt{T3.laundry}_i + \hat{\beta}_4 \texttt{T4.office}_i \\ &+\hat{\beta}_5 \texttt{T5.bath}_i + \hat{\beta}_6 \texttt{T6.outside}_i + \hat{\beta}_7 \texttt{T7.ironing}_i + \hat{\beta}_8 \texttt{T8.teenager}_i \\ &+\hat{\beta}_9 \texttt{T9.parents}_i + \hat{\beta}_{10} \texttt{T.outstation}_i\end{aligned}$$

According to the summary of Model 1, we have:

- All covariates are at least significant at the 5% level
- 12% of the total variability is explained by the regression

**Model 2:**

The second model allowing only linear humidity effects on `lapp`.

$$\begin{aligned}\widehat{\texttt{lapp}}_i = \hat{\beta}_0 \quad &+\hat{\beta}_1\texttt{RH1.kitchen}_i + \hat{\beta}_2\texttt{RH2.living}_i + \hat{\beta}_3\texttt{RH3.laundry}_i + \hat{\beta}_4\texttt{RH4.office}_i \\ &+\hat{\beta}_5\texttt{RH5.bath}_i + \hat{\beta}_6\texttt{RH6.outside}_i + \hat{\beta}_7\texttt{RH7.ironing}_i + \hat{\beta}_8\texttt{RH8.teenager}_i \\ &+\hat{\beta}_9\texttt{RH9.parents}_i + \hat{\beta}_{10}\texttt{RH.outstation}_i\end{aligned}$$

According to the summary of Model 2, we have:

- `RH.outstation` is non-significant at the 5% level, not even at 10% level
- all the other covariates are highly significant
- 17% of the total variability is explained by the regression

**Model 3:**

For the weekday effect we note a non-significant coefficient for the sixth factor `Saturday`. But here, only 0.71% of the total variability is explained by the regression.

$$\begin{aligned}\widehat{\texttt{lapp}}_i = \hat{\beta}_0 \quad &+\hat{\beta}_1\texttt{weekday.Tuesday}_i + \hat{\beta}_2\texttt{weekday.Wednesday}_i + \hat{\beta}_3\texttt{weekday.Thursday}_i \\ &+\hat{\beta}_4\texttt{weekday.Friday}_i + \hat{\beta}_5\texttt{weekday.Saturday}_i + \hat{\beta}_6\texttt{weekday.Sunday}_i\end{aligned}$$

Note that the covariate `weekday.Monday` represents the reference and therefore is not included. So we can argue from the fact of a non-significant factor Saturday, that the covariates `weekday.Saturday` behaves like `weekday.Monday` regarding the fitted response variable.

**Model 4:**

Comparing the indicator $R^2_{adj}$ and the significant coefficients for the different polynomial regression models, we select the intra-day effect with the polynomial regression of order three. Review, even through the polynomial of order four has the same $R^2_{adj}$ as for the third order, `poly(hour, 4)4` is not significant any more. And looking at the first and second polynomial order, we have $R^2_{adj}\left(\text{Model4}_{poly(\texttt{hour},i)_{i=1,2}}\right) \approx 0.19 < 0.23 \approx R^2_{adj}\left(\text{Model4}_{poly(\texttt{hour},3)}\right)$ (and a better $F$-statistic). Therefore we choose a third order polynomial for the hour effect.

The polynomial of order three regression scaled on the hourly effects explains 23% of the total variability and has a significance level smaller than 1% for all covariate effects.

$$\widehat{\texttt{lapp}}_i = \hat{\beta}_0 \quad +\hat{\beta}_1\texttt{poly(hour, 3)1}_i + \hat{\beta}_2\texttt{poly(hour, 3)2}_i + \hat{\beta}_3\texttt{poly(hour, 3)3}_i$$

**Model 5:**

- `T5.bath` and `RH.outstation` are non-significant at the 5% and 10% level, all the other covariates are highly significant (`RH7.ironing` only at 1.13% level)
- 24% of the total variability is explained by the regression

$$\begin{aligned}\widehat{\texttt{lapp}}_i = \hat{\beta}_0 \quad & +\hat{\beta}_1\texttt{T1.kitchen}_i + \hat{\beta}_2\texttt{T2.living}_i + \hat{\beta}_3\texttt{T3.laundry}_i + \hat{\beta}_4\texttt{T4.office}_i \\ & +\hat{\beta}_5\texttt{T5.bath}_i + \hat{\beta}_6\texttt{T6.outside}_i + \hat{\beta}_7\texttt{T7.ironing}_i + \hat{\beta}_8\texttt{T8.teenager}_i \\ & +\hat{\beta}_9\texttt{T9.parents}_i + \hat{\beta}_{10}\texttt{T.outstation}_i \\ & +\hat{\beta}_{11}\texttt{RH1.kitchen}_i + \hat{\beta}_{12}\texttt{RH2.living}_i + \hat{\beta}_{13}\texttt{RH3.laundry}_i + \hat{\beta}_{14}\texttt{RH4.office}_i \\ & +\hat{\beta}_{15}\texttt{RH5.bath}_i + \hat{\beta}_{16}\texttt{RH6.outside}_i + \hat{\beta}_{17}\texttt{RH7.ironing}_i + \hat{\beta}_{18}\texttt{RH8.teenager}_i \\ & +\hat{\beta}_{19}\texttt{RH9.parents}_i + \hat{\beta}_{20}\texttt{RH.outstation}_i\end{aligned}$$

**Model 6:**

Putting the two time effects, weekday and hours, together as a time effect, we have the same problem with the non-significant factor for `weekday.Saturday`, like we had in the Model 3. All the other covariates are highly significant. Furthermore, 23% of the total variability is explained by the regression

$$\begin{aligned}\widehat{\texttt{lapp}}_i = \hat{\beta}_0 \quad & +\hat{\beta}_1\texttt{weekday.Tuesday}_i + \hat{\beta}_2\texttt{weekday.Wednesday}_i + \hat{\beta}_3\texttt{weekday.Thursday}_i \\ & +\hat{\beta}_4\texttt{weekday.Friday}_i + \hat{\beta}_5\texttt{weekday.Saturday}_i + \hat{\beta}_6\texttt{weekday.Sunday}_i \\ & +\hat{\beta}_7\texttt{poly(hour, 3)1}_i + \hat{\beta}_8\texttt{poly(hour, 3)2}_i + \hat{\beta}_9\texttt{poly(hour, 3)3}_i\end{aligned}$$

**Model 7: The full main model**

Finally we include all the covariates in the model to see the effect on the response variable `lapp`. Taking a look at the full main model summary in Table 7.5, we conclude:

- Only covariates `RH4.office`, `RH7.ironing` and fifth weekday factor `Friday` is non-significant at the 10% level
- 33% of the total variability is explained by the regression

$$\begin{aligned}\widehat{\texttt{lapp}}_i = \hat{\beta}_0 \quad & +\textstyle\sum_{j=1}^{10}\hat{\beta}_j\texttt{Tj}_i + \sum_{j=11}^{20}\hat{\beta}_j\texttt{RH(j-10)}_i \\ & +\textstyle\sum_{j=21}^{26}\hat{\beta}_j\texttt{weekday.(j-19)}_i + \sum_{j=27}^{29}\hat{\beta}_j\texttt{poly(hour, 3)(j-26)}_i,\end{aligned}$$

for $j =$ `1.kitchen`, ..., `9.parents`, `.outstation` and $i = 1, \ldots, 19735$.

If we write out the formula of the model, it has the form

$$
\begin{aligned}
\widehat{\texttt{lapp}}_i = \hat{\beta}_0 \quad & + \hat{\beta}_1 \texttt{T1.kitchen}_i + \hat{\beta}_2 \texttt{T2.living}_i + \hat{\beta}_3 \texttt{T3.laundry}_i + \hat{\beta}_4 \texttt{T4.office}_i \\
& + \hat{\beta}_5 \texttt{T5.bath}_i + \hat{\beta}_6 \texttt{T6.outside}_i + \hat{\beta}_7 \texttt{T7.ironing}_i + \hat{\beta}_8 \texttt{T8.teenager}_i \\
& + \hat{\beta}_9 \texttt{T9.parents}_i + \hat{\beta}_{10} \texttt{T.outstation}_i \\
& + \hat{\beta}_{11} \texttt{RH1.kitchen}_i + \hat{\beta}_{12} \texttt{RH2.living}_i + \hat{\beta}_{13} \texttt{RH3.laundry}_i + \hat{\beta}_{14} \texttt{RH4.office}_i \\
& + \hat{\beta}_{15} \texttt{RH5.bath}_i + \hat{\beta}_{16} \texttt{RH6.outside}_i + \hat{\beta}_{17} \texttt{RH7.ironing}_i + \hat{\beta}_{18} \texttt{RH8.teenager}_i \\
& + \hat{\beta}_{19} \texttt{RH9.parents}_i + \hat{\beta}_{20} \texttt{RH.outstation}_i \\
& + \hat{\beta}_{21} \texttt{weekday.Tuesday}_i + \hat{\beta}_{22} \texttt{weekday.Wednesday}_i + \hat{\beta}_{23} \texttt{weekday.Thursday}_i \\
& + \hat{\beta}_{24} \texttt{weekday.Friday}_i + \hat{\beta}_{25} \texttt{weekday.Saturday}_i + \hat{\beta}_{26} \texttt{weekday.Sunday}_i \\
& + \hat{\beta}_{27} \texttt{poly(hour, 3)1}_i + \hat{\beta}_{28} \texttt{poly(hour, 3)2}_i + \hat{\beta}_{29} \texttt{poly(hour, 3)3}_i,
\end{aligned}
$$

$i = 1, \ldots, 19735$.

### 7.2.2 Comparing $R^2_{adj}$ of the energy consumption models

Working with the full main model (Model 7) is the best option, not only because of the highest $R^2_{adj}$ comparing to the other models (cf. Table 7.4 with $R^2_{adj}(\text{Model 7}) = 0.328$), but also the adjusted coefficient of determination exceeds the threshold of $R^2_{adj} = 0.30$.

Now the question arises, if we can obtain a better model by reducing the full main model (Model 7). From that point we examine the non-significant covariates and doing an backward selection, which will be illustrated in the next step.

| | Estimate | Std. Error | t value | Pr(>\|t\|) | |
|---|---|---|---|---|---|
| (Intercept) | 2.0645 | 0.1066 | 19.37 | $< 2e^{-16}$ | *** |
| T1.kitchen | 0.0572 | 0.0113 | 5.05 | $4e^{-07}$ | *** |
| T2.living | -0.1329 | 0.0094 | -14.21 | $< 2e^{-16}$ | *** |
| T3.laundry | 0.1381 | 0.0062 | 22.29 | $< 2e^{-16}$ | *** |
| T4.office | 0.0425 | 0.0058 | 7.30 | $3e^{-13}$ | *** |
| T5.bath | 0.0116 | 0.0067 | 1.73 | 0.0830 | . |
| T6.outside | -0.0082 | 0.0040 | -2.05 | 0.0408 | * |
| T7.ironing | -0.0291 | 0.0076 | -3.83 | 0.0001 | *** |
| T8.teenager | 0.1192 | 0.0058 | 20.69 | $< 2e^{-16}$ | *** |
| T9.parents | -0.1647 | 0.0107 | -15.39 | $< 2e^{-16}$ | *** |
| T.outstation | 0.0136 | 0.0045 | 3.01 | 0.0026 | ** |
| RH1.kitchen | 0.0777 | 0.0040 | 19.40 | $< 2e^{-16}$ | *** |
| RH2.living | -0.0783 | 0.0044 | -17.77 | $< 2e^{-16}$ | *** |
| RH3.laundry | 0.0419 | 0.0039 | 10.61 | $< 2e^{-16}$ | *** |
| RH4.office | 0.0047 | 0.0037 | 1.27 | 0.2035 | |
| RH5.bath | 0.0042 | 0.0005 | 8.18 | $3e^{-16}$ | *** |
| RH6.outside | -0.0016 | 0.0004 | -4.13 | $4e^{-05}$ | *** |
| RH7.ironing | -0.0037 | 0.0024 | -1.52 | 0.1275 | |
| RH8.teenager | -0.0202 | 0.0024 | -8.54 | $< 2e^{-16}$ | *** |
| RH9.parents | -0.0194 | 0.0025 | -7.63 | $2e^{-14}$ | *** |
| RH.outstation | 0.0073 | 0.0007 | 10.59 | $< 2e^{-16}$ | *** |
| weekday.Tuesday | -0.1377 | 0.0147 | -9.35 | $< 2e^{-16}$ | *** |
| weekday.Wednesday | -0.0415 | 0.0146 | -2.83 | 0.0046 | ** |
| weekday.Thursday | -0.1043 | 0.0146 | -7.14 | $9e^{-13}$ | *** |
| weekday.Friday | -0.0011 | 0.0149 | -0.07 | 0.9434 | |
| weekday.Saturday | 0.0576 | 0.0149 | 3.86 | 0.0001 | *** |
| weekday.Sunday | -0.0542 | 0.0148 | -3.67 | 0.0002 | *** |
| poly(hour, 3)1 | 32.2239 | 0.8925 | 36.11 | $< 2e^{-16}$ | *** |
| poly(hour, 3)2 | -20.4014 | 0.8409 | -24.26 | $< 2e^{-16}$ | *** |
| poly(hour, 3)3 | -14.2197 | 0.7427 | -19.15 | $< 2e^{-16}$ | *** |

Table 7.5: Summary of full main model (Model 7): $\widehat{\texttt{lapp}} = \hat{\boldsymbol{\beta}}\mathbf{X}$, where $\mathbf{X} = (\mathbf{1}, \texttt{Tj}, \texttt{RHj}, \texttt{weekday.d}, \texttt{poly(hour, 3).c})_{j=\texttt{1.kitchen},\ldots,\texttt{.outstation};d=2,\ldots7;c=1,2,3} \in \mathbb{R}^{19735x30}$ is the design matrix with all the predictors. We yield the performances $R^2_{adj} = 0.328$, $F$-statistic = 333 on 29 and 19705 DF, $p$-value $< 2e - 16$.

### 7.2.3 Reduced models

Regarding possible interaction between the variables, it is sufficient to start the reduction from the full main Model 7 (cf. Table 7.5), where we have a look at all main effects. Have a look at the effects first, we detect some non-significant covariates.

- non-significant at 5% level:

  T5.bath

  RH4.office

  RH7.ironing

  weekday.Friday

- non-significant at 10% level:

  RH4.office

  RH7.ironing

  weekday.Friday

For our purpose it is suffices to focus only on the non-significant covariates at a 10% level.

**Reduced Model 1:**

It seems that Friday has similar pattern like its reference covariate `weekday.Monday`, since `weekday.Friday` is highly non-significant. Nevertheless, we leave all the weekday variables for now in the model as they are of the class factor.

Reducing in a stepwise fashion, we remove the covariate with the highest p-value that is `RH4.office`, then afterwards when necessary RH7.ironing. (c.f. Table 7.5)

$$p(\texttt{RH4.office}) = 0.2035 > 0.1275 = p(\texttt{RH7.ironing})$$

After removing `RH4.office`, `RH7.ironing` has an even higher $p$-value. Reducing again, we obtain the reduced model summarized in Table 7.6.

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | 2.10 | 0.10 | 20.31 | 0.00 |
| T1.kitchen | 0.06 | 0.01 | 5.06 | 0.00 |
| T2.living | -0.13 | 0.01 | -14.72 | 0.00 |
| T3.laundry | 0.14 | 0.01 | 22.37 | 0.00 |
| T4.office | 0.04 | 0.01 | 7.36 | 0.00 |
| T5.bath | 0.01 | 0.01 | 1.80 | 0.07 |
| T6.outside | -0.01 | 0.00 | -2.06 | 0.04 |
| T7.ironing | -0.03 | 0.01 | -4.87 | 0.00 |
| T8.teenager | 0.12 | 0.01 | 21.86 | 0.00 |
| T9.parents | -0.16 | 0.01 | -15.44 | 0.00 |
| T.outstation | 0.01 | 0.00 | 2.97 | 0.00 |
| RH1.kitchen | 0.08 | 0.00 | 19.52 | 0.00 |
| RH2.living | -0.08 | 0.00 | -18.55 | 0.00 |
| RH3.laundry | 0.04 | 0.00 | 10.97 | 0.00 |
| RH5.bath | 0.00 | 0.00 | 8.17 | 0.00 |
| RH6.outside | -0.00 | 0.00 | -4.12 | 0.00 |
| RH8.teenager | -0.02 | 0.00 | -10.05 | 0.00 |
| RH9.parents | -0.02 | 0.00 | -7.69 | 0.00 |
| RH.outstation | 0.01 | 0.00 | 10.77 | 0.00 |
| weekday.Tuesday | -0.14 | 0.01 | -9.28 | 0.00 |
| weekday.Wednesday | -0.04 | 0.01 | -2.85 | 0.00 |
| weekday.Thursday | -0.10 | 0.01 | -7.13 | 0.00 |
| weekday.Friday | -0.00 | 0.01 | -0.11 | 0.91 |
| weekday.Saturday | 0.06 | 0.01 | 3.93 | 0.00 |
| weekday.Sunday | -0.06 | 0.01 | -3.77 | 0.00 |
| poly(hour, 3)1 | 32.48 | 0.87 | 37.26 | 0.00 |
| poly(hour, 3)2 | -20.27 | 0.84 | -24.23 | 0.00 |
| poly(hour, 3)3 | -14.43 | 0.73 | -19.86 | 0.00 |

Table 7.6: Summary of Reduced Model 1: From the full main model (Model 7), we removed non-significant `RH4.office` and `RH7.ironing` and obtained this reduced model. We have the performances $R^2_{adj} = 0.3279$, $F$-statistic $= 357.6$ on 27 and 19707 DF, $p$-value $< 2e{-}16$.

Now all covariates are significant at a 10% level. But we still have the non-significant factor `weekday.Friday` with an increased $p$-value. At 5 % level `T5.bath` would not be significant anymore, however bathroom is an important room for appliances use as we have a few electrical devices here. It is sufficient to set $\alpha = 0.10$.

With the reduction, we are not reaching a higher adjusted coefficient of determination, i.e. $R^2_{adj}$(Reduced Model 1) $= 0.3279$ comparing to $R^2_{adj}$(Model 7) $= 0.328$.

**Reduced Model 2:**

Trying to reduce factor Friday first, since it has the highest p-value in the full main model 7.5, that is

$$p(\texttt{weekday.Friday}) = 0.94.$$

Removing the factor of weekday Friday, we obtain a model with non-significant weekday factor of Monday with

$$p(\texttt{weekday.Monday}) = 0.94.$$

Regarding the full main model (Model 7) we have highly related weekdays Monday and Friday, maybe due to the same day daily routine, e.g. home office.

Continue with reducing the factor Monday, we are having a model with still non-significant covariates `RH4.office` and `RH7.ironing` with the same p-value as in the full main model (Model 7). So proceeding with the same backward selection as we did for the Reduced Model 1, we obtain the following significant Reduced Model 2 in Table 7.7.

To summarize the path to Reduced Model 2, we started with the full main model (Model 7) and removed the factor `weekday.Friday` first, then continued with the reduction of `weekday.Monday`, the covariates `RH4.office` and `RH7.ironing`, by the highest $p$-value respectively.

Again, with the reduction, we are not reaching a higher adjusted coefficient of determination, i.e. $R^2_{adj}$(Reduced Model 2) $= 0.328 = R^2_{adj}$(Model 7).

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | 2.0937 | 0.1022 | 20.48 | 0.0000 |
| T1.kitchen | 0.0572 | 0.0113 | 5.06 | 0.0000 |
| T2.living | -0.1317 | 0.0089 | -14.73 | 0.0000 |
| T3.laundry | 0.1381 | 0.0062 | 22.40 | 0.0000 |
| T4.office | 0.0403 | 0.0055 | 7.37 | 0.0000 |
| T5.bath | 0.0121 | 0.0066 | 1.82 | 0.0681 |
| T6.outside | -0.0083 | 0.0040 | -2.06 | 0.0395 |
| T7.ironing | -0.0339 | 0.0069 | -4.88 | 0.0000 |
| T8.teenager | 0.1210 | 0.0055 | 22.17 | 0.0000 |
| T9.parents | -0.1615 | 0.0104 | -15.46 | 0.0000 |
| T.outstation | 0.0133 | 0.0045 | 2.97 | 0.0030 |
| RH1.kitchen | 0.0781 | 0.0040 | 19.53 | 0.0000 |
| RH2.living | -0.0777 | 0.0042 | -18.56 | 0.0000 |
| RH3.laundry | 0.0424 | 0.0038 | 11.02 | 0.0000 |
| RH5.bath | 0.0041 | 0.0005 | 8.17 | 0.0000 |
| RH6.outside | -0.0016 | 0.0004 | -4.13 | 0.0000 |
| RH8.teenager | -0.0211 | 0.0021 | -10.11 | 0.0000 |
| RH9.parents | -0.0189 | 0.0025 | -7.69 | 0.0000 |
| RH.outstation | 0.0072 | 0.0007 | 10.78 | 0.0000 |
| weekday.Tuesday | -0.1355 | 0.0127 | -10.66 | 0.0000 |
| weekday.Wednesday | -0.0410 | 0.0128 | -3.21 | 0.0013 |
| weekday.Thursday | -0.1034 | 0.0127 | -8.15 | 0.0000 |
| weekday.Saturday | 0.0594 | 0.0128 | 4.62 | 0.0000 |
| weekday.Sunday | -0.0549 | 0.0129 | -4.24 | 0.0000 |
| poly(hour, 3)1 | 32.4729 | 0.8705 | 37.31 | 0.0000 |
| poly(hour, 3)2 | -20.2766 | 0.8365 | -24.24 | 0.0000 |
| poly(hour, 3)3 | -14.4245 | 0.7256 | -19.88 | 0.0000 |

Table 7.7: Summary of Reduced Model 2: From the full main model (Model 7), we removed non-significant `weekday.Friday`, `weekday.Monday`, `RH4.office` and `RH7.ironing` and obtained this Reduced Model 2. We have the performances $R^2_{adj} = 0.328$, $F$-statistic = 371.4 on 26 and 19708 DF, $p$-value: $< 2e - 16$.

## 7.3 Interaction models

When the effect of the temperature and time or humidity and time is not additive, that is the effect of the two cause variables temperature or humidity on the appliances also depends on the state of the time variable, to be exact, depends on the hour of the day. We already observed strong hourly pattern concerning the response variable `lapp`. But now we want to couple the hourly pattern with the other covariates, temperatures and then humidities, to detect interactions between them.

### 7.3.1 Interaction between temperatures and hours

#### Correlation - Scatter-plots

Our goal is to check if all the areas have the same slope. A look at the difference between the maximal and minimal slope tells us that it is not higher than one.

$$\left| max_h(\hat{\rho}_{((\texttt{lapp},\texttt{Tj})|h)}) - min_h(\hat{\rho}_{((\texttt{lapp},\texttt{Tj})|h)}) \right| < 1$$

For our response variable `lapp`, we can perform residual plots separated hourly-wise, that is 24 plots for all the hours, to detect the linear dependencies and obtain the correlations with formula in (2.2). An example with one temperature covariate, i.e. $\hat{\rho}_{((\texttt{lapp},\texttt{T1.kitchen})|h)}$ for $h = 1, \ldots, 24$, is provided in the Appendix in Figure A.2 and A.3.

- $\hat{\rho}_{((\texttt{lapp},\texttt{T1.kitchen})|h)}$ for $h = 1, \ldots, 24$ in Figure A.2 and Figure A.3, which is the hourly-wise correlation between `lapp` and kitchen temperature:

  → there are positive slopes,

  except for the hours ten and twelve, where we have a small descending line

  → highest slopes for hours one to eight,

  afterwards we have a slightly ascending slope

  → rising and higher slopes are reached again after hour 16

- $\hat{\rho}_{((\texttt{lapp},\texttt{T2.living})|h)}$ for $h = 1, \ldots, 24$:

  → until hour 12 we have ascending line,

  where the strongest slopes observed for hour one to nine

  → small descending lines from hour 13 to 15

  → slightly positive slopes until hour 23 and the strongest in hour 24

- $\hat{\rho}_{((\texttt{lapp},\texttt{T3.laundry})|h)}$ for $h = 1, \ldots, 24$:

  → almost the same pattern as for the kitchen temperature

- $\hat{\rho}_{((\texttt{lapp},\texttt{T4.office})|h)}$ for $h = 1, \ldots, 24$:

  → almost the same pattern as for the living room temperature.

- $\hat{\rho}_{((\texttt{lapp},\texttt{T5.bath})|h)}$ for $h = 1, \ldots, 24$:

  $\rightarrow$ almost the same pattern as for the kitchen temperature,

  where we have stronger positive slopes in the morning hours and between 16-19 and 23-24.

- $\hat{\rho}_{((\texttt{lapp},\texttt{T6.outside})|h)}$ for $h = 1, \ldots, 24$:

  $\rightarrow$ higher positive slopes between hour 1 and 8

  $\rightarrow$ small positive slopes from 9 to 24,

  where it turns to a descending line in hour 12, 15, 20, 22 and 23.

Summarizing all the correlations in Figure 7.1 gives a good overview of the pattern and dependencies. We see weak, mostly positive, correlation values for all the indoor and outdoor temperatures which means that we have mostly positive linear dependent variables temperatures and hours. The highest correlation we see in the morning hours where the maximum values are reached.

Whereas values near zero showing non-correlated variables temperatures and time which is the case in early afternoon hours.

On the one hand, we can see a clearly pattern through the day. On the other hand we have different slopes, consequently the lines with the points of hours are intersecting one another and are not identically parallel. This mean there are interactions due to uneven changing correlations.

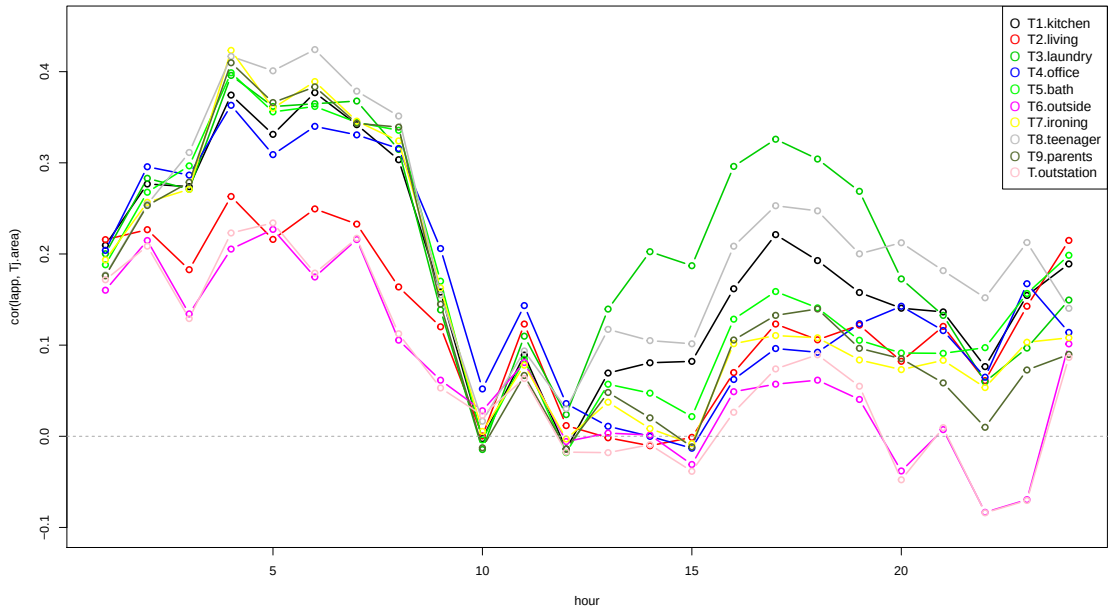



Figure 7.1: Plot is showing the estimated correlation $\hat{\rho}_{((\texttt{lapp},\texttt{Tj})|h)}$ between `lapp` and all the temperature in an area, i.e. `Tj` with $j =$ `1.kitchen`, $\ldots$, `9.parents`, `.outstation` at hour $h$, with $h = 1, \ldots, 24$. Here we have mainly positive correlations and the weakest for outside temperatures.

**Regression allowing for interaction between temperatures and polynomial hourly effect**

In Summary (c.f. Table 7.8) we have significant interaction terms, since there is no case where all three polynomial interaction terms with the coupled temperature are non-significant. For example, the regression $\texttt{lapp}_i \sim \texttt{T9.parents}_i * \texttt{poly(hour, 3)}_i$, $i = 1, \ldots, 19735$, has two non-significant interactions between the parents room temperature and the second and third polynomial coefficient at a significance level of 0.1, but the interaction with the first polynomial coefficient is highly significant which makes the overall interaction between the parents room temperature and the polynomial hour effect significant.

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | 3.2449 | 0.0577 | 56.24 | 0.0000 |
| T1.kitchen | 0.0486 | 0.0027 | 18.32 | 0.0000 |
| poly(hour, 3)1 | 39.9875 | 8.4978 | 4.71 | 0.0000 |
| poly(hour, 3)2 | -45.0675 | 8.2873 | -5.44 | 0.0000 |
| poly(hour, 3)3 | 9.6581 | 8.2148 | 1.18 | 0.2397 |
| T1.kitchen:poly(hour, 3)1 | -0.0478 | 0.3848 | -0.12 | 0.9011 |
| T1.kitchen:poly(hour, 3)2 | 1.6324 | 0.3776 | 4.32 | 0.0000 |
| T1.kitchen:poly(hour, 3)3 | -1.1672 | 0.3742 | -3.12 | 0.0018 |

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | 3.7327 | 0.0436 | 85.56 | 0.0000 |
| T2.living | 0.0283 | 0.0021 | 13.16 | 0.0000 |
| poly(hour, 3)1 | 46.1698 | 6.8290 | 6.76 | 0.0000 |
| poly(hour, 3)2 | -59.5758 | 6.2015 | -9.61 | 0.0000 |
| poly(hour, 3)3 | -38.0029 | 6.3705 | -5.97 | 0.0000 |
| T2.living:poly(hour, 3)1 | -0.4281 | 0.3334 | -1.28 | 0.1991 |
| T2.living:poly(hour, 3)2 | 2.5229 | 0.3011 | 8.38 | 0.0000 |
| T2.living:poly(hour, 3)3 | 1.1038 | 0.3119 | 3.54 | 0.0004 |

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | 3.2264 | 0.0452 | 71.34 | 0.0000 |
| T3.laundry | 0.0484 | 0.0020 | 23.91 | 0.0000 |
| poly(hour, 3)1 | 24.8080 | 6.4827 | 3.83 | 0.0001 |
| poly(hour, 3)2 | 7.4483 | 6.4862 | 1.15 | 0.2508 |
| poly(hour, 3)3 | 27.0408 | 6.4560 | 4.19 | 0.0000 |
| T3.laundry:poly(hour, 3)1 | 0.6622 | 0.2891 | 2.29 | 0.0220 |
| T3.laundry:poly(hour, 3)2 | -0.6699 | 0.2903 | -2.31 | 0.0210 |
| T3.laundry:poly(hour, 3)3 | -1.9448 | 0.2883 | -6.75 | 0.0000 |

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | 3.6523 | 0.0422 | 86.46 | 0.0000 |
| T4.office | 0.0313 | 0.0020 | 15.51 | 0.0000 |
| poly(hour, 3)1 | 59.2344 | 6.0455 | 9.80 | 0.0000 |
| poly(hour, 3)2 | -27.1185 | 6.0678 | -4.47 | 0.0000 |
| poly(hour, 3)3 | -27.6381 | 6.0134 | -4.60 | 0.0000 |
| T4.office:poly(hour, 3)1 | -0.9441 | 0.2867 | -3.29 | 0.0010 |
| T4.office:poly(hour, 3)2 | 0.9239 | 0.2878 | 3.21 | 0.0013 |
| T4.office:poly(hour, 3)3 | 0.5634 | 0.2859 | 1.97 | 0.0488 |

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | 3.5671 | 0.0440 | 81.06 | 0.0000 |
| data.T5 | 0.0376 | 0.0022 | 16.79 | 0.0000 |
| poly(hour, 3)1 | 53.5842 | 6.2860 | 8.52 | 0.0000 |
| poly(hour, 3)2 | -22.2245 | 6.3071 | -3.52 | 0.0004 |
| poly(hour, 3)3 | -13.1016 | 6.2993 | -2.08 | 0.0376 |
| data.T5:poly(hour, 3)1 | -0.7033 | 0.3162 | -2.22 | 0.0261 |
| data.T5:poly(hour, 3)2 | 0.7137 | 0.3178 | 2.25 | 0.0248 |
| data.T5:poly(hour, 3)3 | -0.1776 | 0.3181 | -0.56 | 0.5767 |

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | 4.2885 | 0.0071 | 601.81 | 0.0000 |
| T6.outside | 0.0043 | 0.0008 | 5.75 | 0.0000 |
| poly(hour, 3)1 | 44.4450 | 0.9472 | 46.92 | 0.0000 |
| poly(hour, 3)2 | -10.6284 | 0.9628 | -11.04 | 0.0000 |
| poly(hour, 3)3 | -22.0363 | 0.9660 | -22.81 | 0.0000 |
| T6.outside:poly(hour, 3)1 | -0.6791 | 0.1058 | -6.42 | 0.0000 |
| T6.outside:poly(hour, 3)2 | 0.4523 | 0.1039 | 4.35 | 0.0000 |
| T6.outside:poly(hour, 3)3 | 0.7364 | 0.1046 | 7.04 | 0.0000 |

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | 3.7915 | 0.0396 | 95.65 | 0.0000 |
| T7.ironing | 0.0253 | 0.0019 | 13.01 | 0.0000 |
| poly(hour, 3)1 | 61.8153 | 5.5778 | 11.08 | 0.0000 |
| poly(hour, 3)2 | -28.4755 | 5.5948 | -5.09 | 0.0000 |
| poly(hour, 3)3 | -29.4390 | 5.5843 | -5.27 | 0.0000 |
| T7.ironing:poly(hour, 3)1 | -1.0881 | 0.2728 | -3.99 | 0.0001 |
| T7.ironing:poly(hour, 3)2 | 1.0457 | 0.2752 | 3.80 | 0.0001 |
| T7.ironing:poly(hour, 3)3 | 0.6459 | 0.2734 | 2.36 | 0.0181 |

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | 3.2421 | 0.0469 | 69.17 | 0.0000 |
| T8.teenager | 0.0481 | 0.0021 | 22.68 | 0.0000 |
| poly(hour, 3)1 | 21.2354 | 6.6176 | 3.21 | 0.0013 |
| poly(hour, 3)2 | -18.1352 | 6.5171 | -2.78 | 0.0054 |
| poly(hour, 3)3 | 15.6390 | 6.5407 | 2.39 | 0.0168 |
| T8.teenager:poly(hour, 3)1 | 0.8226 | 0.2947 | 2.79 | 0.0052 |
| T8.teenager:poly(hour, 3)2 | 0.3846 | 0.2929 | 1.31 | 0.1892 |
| T8.teenager:poly(hour, 3)3 | -1.4412 | 0.2931 | -4.92 | 0.0000 |

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | 3.7623 | 0.0398 | 94.64 | 0.0000 |
| T9.parents | 0.0278 | 0.0020 | 13.70 | 0.0000 |
| poly(hour, 3)1 | 62.3657 | 5.6071 | 11.12 | 0.0000 |
| poly(hour, 3)2 | -15.5625 | 5.5968 | -2.78 | 0.0054 |
| poly(hour, 3)3 | -22.5552 | 5.6212 | -4.01 | 0.0001 |
| T9.parents:poly(hour, 3)1 | -1.1486 | 0.2869 | -4.00 | 0.0001 |
| T9.parents:poly(hour, 3)2 | 0.4506 | 0.2863 | 1.57 | 0.1156 |
| T9.parents:poly(hour, 3)3 | 0.3057 | 0.2874 | 1.06 | 0.2874 |

Table 7.8: Summary of the interaction models. The fitted model $\widehat{\texttt{lapp}} = \hat{\boldsymbol{\beta}}\boldsymbol{X}$, where $\boldsymbol{X} = (\mathbf{1}, \texttt{Tj}, \texttt{poly(hour, 3)1}, \texttt{poly(hour, 3)2}, \texttt{poly(hour, 3)3}, \texttt{Tj} \times \texttt{poly(hour, 3)1}, \texttt{Tj} \times \texttt{poly(hour, 3)2}, \texttt{Tj} \times \texttt{poly(hour, 3)3})$, whereby Tj represents in each case another given area temperature. The corresponding $R^2_{adj}$ for the significant model will be summarized in a following Table 7.9.

As for the nine models

$$\texttt{lapp}_i \sim \texttt{Tj}_i * \texttt{poly(hour, 3)}_i, \quad i = 1, \ldots, 19735.$$

for $j =$ 1.kitchen, 2.living, 3.laundry, 4.office, 5.bath, 6.outside, 7.ironing, 8.teenager, 9.parents, all the interaction terms are significant, we also examine all

temperatures coupled polynomial hourly effect with the model for $i = 1, \ldots, 19735$:

$$\begin{aligned} \texttt{lapp}_i \sim \; ( \; & \texttt{T1.kitchen}_i + \texttt{T2.living}_i + \texttt{T5.laundry}_i + \texttt{T4.office}_i + \texttt{T5.bath}_i \\ & + \texttt{T6.outside}_i + \texttt{T7.ironing}_i + \texttt{T8.teenager}_i + \texttt{T9.parents}_i + \texttt{T.outstation}_i) \\ * \; & (\texttt{poly(hour, 3)}_i). \end{aligned}$$

At a significance level of 0.1, we accept all the interaction terms like the regressions in Table 7.8, since we do not have non-significant room temperatures coupled with the polynomial hour effect. Like we have seen before, the interaction with all three polynomial coefficients have to be non-significant so that the interaction is non-significant.

In addition, we prepare interaction plots, given in Figure A.6, of the regressions

$$\texttt{lapp}_i \sim \texttt{Tj}_i * \texttt{poly(hour, 3)}_i, \quad i = 1, \ldots, 19735,$$

for $j =$ `1.kitchen`, `2.living`, `3.laundry`, `4.office`, `5.bath`, `6.outside`, `7.ironing`, `8.teenager`, `9.parents`, `.outstation`, to visualize the partitioned hourly pattern with their 95 % confidence interval.

### Comparing $R^2_{adj}$ of main and interaction effect models

| Interaction effects | |
|---|---|
| Model | $R^2_{adj}$ |
| `T1.kitchen` * `poly(hour,3)` | 0.241 |
| `T2.living` * `poly(hour,3)` | 0.235 |
| `T3.laundry` * `poly(hour,3)` | 0.252 |
| `T4.office` * `poly(hour,3)` | 0.237 |
| `T5.bath` * `poly(hour,3)` | 0.238 |
| `T6.outside` * `poly(hour,3)` | 0.232 |
| `T7.ironing` * `poly(hour,3)` | 0.235 |
| `T8.teenager` * `poly(hour,3)` | 0.248 |
| `T9.parents` * `poly(hour,3)` | 0.235 |
| `T.outstation` * `poly(hour,3)` | 0.231 |
| $(\sum_{j=1}^{10} \texttt{Tj}) * \texttt{poly(hour,3)}$ | 0.341 |

Table 7.9: $R^2_{adj}$ for the interaction models between temperatures and polynomial hourly effect.

| Main effects | |
|---|---|
| Model | $R^2_{adj}$ |
| `T1.kitchen` + `poly(hour,3)` | 0.240 |
| `T2.living` + `poly(hour,3)` | 0.231 |
| `T3.laundry` + `poly(hour,3)` | 0.250 |
| `T4.office` + `poly(hour,3)` | 0.236 |
| `T5.bath` + `poly(hour,3)` | 0.238 |
| `T6.outside` + `poly(hour,3)` | 0.228 |
| `T7.ironing` + `poly(hour,3)` | 0.234 |
| `T8.teenager` + `poly(hour,3)` | 0.247 |
| `T9.parents` + `poly(hour,3)` | 0.234 |
| `T.outstation` + `poly(hour,3)` | 0.228 |
| $(\sum_{j=1}^{10} \texttt{Tj}) + \texttt{poly(hour,3)}$ | 0.296 |

Table 7.10: $R^2_{adj}$ for the main models for temperatures and polynomial hourly effect, corresponding to Table 7.9.

In the Table 7.9 and 7.10 we see the small improvement using the interaction between temperatures and the hours. For the interaction model using all the temperatures, the greatest improvement in the adjusted coefficient of determination is observed.

### 7.3.2 Interaction between humidities and hours

#### Correlation - Scatter-plots

As for the temperature we look at the hourly-wise correlation plots between `lapp` and humidities, e.g. visualized in Figure A.4 and A.5. The slopes for humidities are even flatter, since we observe correlation values between [-0,15; 0,15]. Here patterns are harder to detect, but looking closely we have similarities, but with the greatest slope differences in the morning hours. For bathroom, office and laundry room the figures are showing negative whereas for kitchen and laundry room positive slopes are observed in the morning hours. As for the outside humidity, i.e. `RH6.outside`, all the lines are descending, except for the hour between 13 and 15.

Summarizing all the hourly-wise correlations obtained form the residuals, in Figure 7.2, we see very fluctuated correlations with positive and negative correlated variables. Overall the highest correlations are in the early afternoon, see the maximum and minimum peaks. So due to the different slopes, i.e. uneven changing correlations, we have interactions between the humidities and the hours.

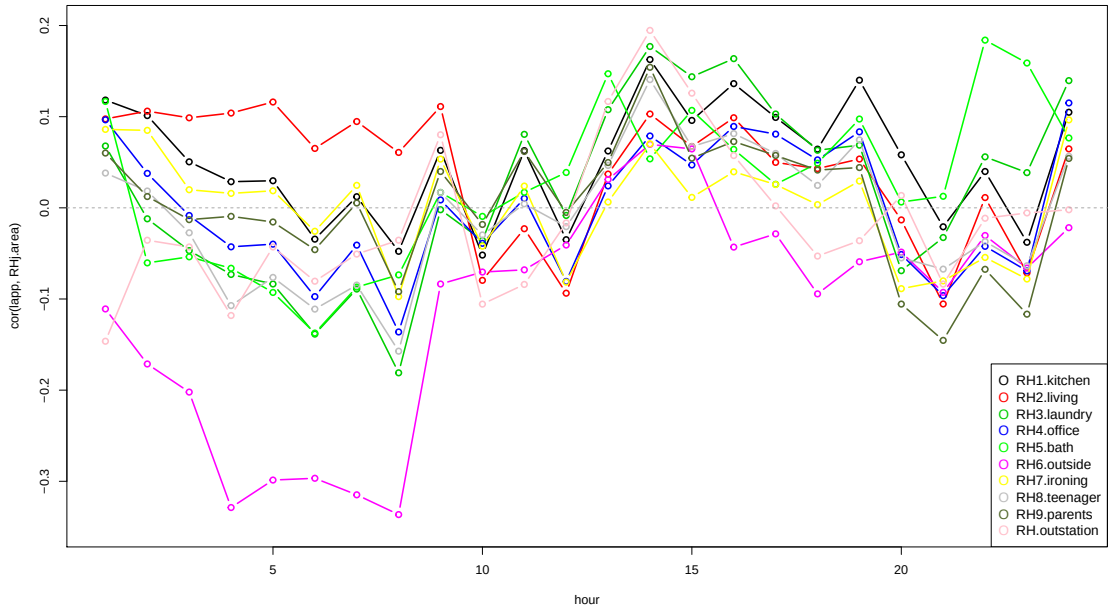



Figure 7.2: Plot is showing the estimated correlation $\hat{\rho}_{((\texttt{lapp},\texttt{RHj})|h)}$ between `lapp` and all the humidities in an area, i.e. `RHj` with $j = \texttt{1.kitchen}, \ldots, \texttt{9.parents}, \texttt{.outstation}$ at hour $h$ with $h = 1, \ldots, 24$.

**Regression allowing for interaction between humidities and polynomial hourly effect**

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | 3.9077 | 0.0426 | 91.78 | 0.0000 |
| RH1.kitchen | 0.0098 | 0.0011 | 9.32 | 0.0000 |
| poly(hour, 3)1 | 20.7990 | 5.9834 | 3.48 | 0.0005 |
| poly(hour, 3)2 | 6.6421 | 6.0864 | 1.09 | 0.2752 |
| poly(hour, 3)3 | 13.5707 | 6.0111 | 2.26 | 0.0240 |
| RH1.kitchen:poly(hour, 3)1 | 0.4752 | 0.1485 | 3.20 | 0.0014 |
| RH1.kitchen:poly(hour, 3)2 | -0.3373 | 0.1507 | -2.24 | 0.0252 |
| RH1.kitchen:poly(hour, 3)3 | -0.7559 | 0.1497 | -5.05 | 0.0000 |

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | 4.0808 | 0.0455 | 89.61 | 0.0000 |
| RH2.living | 0.0056 | 0.0011 | 5.07 | 0.0000 |
| poly(hour, 3)1 | 48.0184 | 6.6233 | 7.25 | 0.0000 |
| poly(hour, 3)2 | 9.4643 | 6.2834 | 1.51 | 0.1320 |
| poly(hour, 3)3 | 18.8422 | 6.4637 | 2.92 | 0.0036 |
| RH2.living:poly(hour, 3)1 | -0.1770 | 0.1622 | -1.09 | 0.2753 |
| RH2.living:poly(hour, 3)2 | -0.4213 | 0.1542 | -2.73 | 0.0063 |
| RH2.living:poly(hour, 3)3 | -0.9031 | 0.1586 | -5.69 | 0.0000 |

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | 4.0905 | 0.0502 | 81.50 | 0.0000 |
| RH3.laundry | 0.0055 | 0.0013 | 4.32 | 0.0000 |
| poly(hour, 3)1 | 7.1360 | 6.9884 | 1.02 | 0.3072 |
| poly(hour, 3)2 | 16.7481 | 6.9996 | 2.39 | 0.0167 |
| poly(hour, 3)3 | 21.8549 | 7.0384 | 3.11 | 0.0019 |
| RH3.laundry:poly(hour, 3)1 | 0.8428 | 0.1771 | 4.76 | 0.0000 |
| RH3.laundry:poly(hour, 3)2 | -0.6076 | 0.1776 | -3.42 | 0.0006 |
| RH3.laundry:poly(hour, 3)3 | -0.9867 | 0.1786 | -5.52 | 0.0000 |

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | 4.2472 | 0.0373 | 113.82 | 0.0000 |
| RH4.office | 0.0015 | 0.0010 | 1.54 | 0.1234 |
| poly(hour, 3)1 | 33.7654 | 5.1565 | 6.55 | 0.0000 |
| poly(hour, 3)2 | -1.1897 | 5.2425 | -0.23 | 0.8205 |
| poly(hour, 3)3 | 5.9885 | 5.1838 | 1.16 | 0.2480 |
| RH4.office:poly(hour, 3)1 | 0.1639 | 0.1311 | 1.25 | 0.2113 |
| RH4.office:poly(hour, 3)2 | -0.1500 | 0.1332 | -1.13 | 0.2599 |
| RH4.office:poly(hour, 3)3 | -0.5819 | 0.1321 | -4.41 | 0.0000 |

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | 4.1333 | 0.0265 | 156.17 | 0.0000 |
| RH5.bath | 0.0035 | 0.0005 | 6.68 | 0.0000 |
| poly(hour, 3)1 | 18.0408 | 3.5163 | 5.13 | 0.0000 |
| poly(hour, 3)2 | 3.3530 | 3.1705 | 1.06 | 0.2903 |
| poly(hour, 3)3 | 8.4375 | 3.6204 | 2.33 | 0.0198 |
| RH5.bath:poly(hour, 3)1 | 0.4466 | 0.0676 | 6.61 | 0.0000 |
| RH5.bath:poly(hour, 3)2 | -0.2153 | 0.0597 | -3.61 | 0.0003 |
| RH5.bath:poly(hour, 3)3 | -0.5133 | 0.0703 | -7.30 | 0.0000 |

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | 4.3886 | 0.0095 | 461.48 | 0.0000 |
| RH6.outside | -0.0013 | 0.0001 | -9.18 | 0.0000 |
| poly(hour, 3)1 | 32.9257 | 1.2991 | 25.34 | 0.0000 |
| poly(hour, 3)2 | -1.8856 | 1.2460 | -1.51 | 0.1302 |
| poly(hour, 3)3 | -9.1099 | 1.2949 | -7.04 | 0.0000 |
| RH6.outside:poly(hour, 3)1 | 0.1073 | 0.0203 | 5.29 | 0.0000 |
| RH6.outside:poly(hour, 3)2 | -0.0983 | 0.0196 | -5.03 | 0.0000 |
| RH6.outside:poly(hour, 3)3 | -0.1251 | 0.0202 | -6.19 | 0.0000 |

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | 4.2988 | 0.0298 | 144.34 | 0.0000 |
| RH7.ironing | 0.0001 | 0.0008 | 0.17 | 0.8648 |
| poly(hour, 3)1 | 46.8920 | 4.0923 | 11.46 | 0.0000 |
| poly(hour, 3)2 | -1.4773 | 4.1461 | -0.36 | 0.7216 |
| poly(hour, 3)3 | -5.1810 | 4.0893 | -1.27 | 0.2052 |
| RH7.ironing:poly(hour, 3)1 | -0.1941 | 0.1143 | -1.70 | 0.0896 |
| RH7.ironing:poly(hour, 3)2 | -0.1619 | 0.1158 | -1.40 | 0.1621 |
| RH7.ironing:poly(hour, 3)3 | -0.3272 | 0.1143 | -2.86 | 0.0042 |

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | 4.2886 | 0.0370 | 115.86 | 0.0000 |
| RH8.teenager | 0.0006 | 0.0009 | 0.65 | 0.5125 |
| poly(hour, 3)1 | 29.2453 | 5.1293 | 5.70 | 0.0000 |
| poly(hour, 3)2 | 15.6118 | 5.0963 | 3.06 | 0.0022 |
| poly(hour, 3)3 | 18.2670 | 4.9931 | 3.66 | 0.0003 |
| RH8.teenager:poly(hour, 3)1 | 0.2439 | 0.1180 | 2.07 | 0.0388 |
| RH8.teenager:poly(hour, 3)2 | -0.5430 | 0.1182 | -4.59 | 0.0000 |
| RH8.teenager:poly(hour, 3)3 | -0.8274 | 0.1155 | -7.17 | 0.0000 |

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | 4.2845 | 0.0437 | 97.94 | 0.0000 |
| RH9.parents | 0.0004 | 0.0010 | 0.38 | 0.7051 |
| poly(hour, 3)1 | 45.9151 | 5.9671 | 7.69 | 0.0000 |
| poly(hour, 3)2 | 18.7373 | 5.8200 | 3.22 | 0.0013 |
| poly(hour, 3)3 | -6.5935 | 5.8759 | -1.12 | 0.2618 |
| RH9.parents:poly(hour, 3)1 | -0.1512 | 0.1444 | -1.05 | 0.2948 |
| RH9.parents:poly(hour, 3)2 | -0.6322 | 0.1411 | -4.48 | 0.0000 |
| RH9.parents:poly(hour, 3)3 | -0.2565 | 0.1414 | -1.81 | 0.0698 |

Table 7.11: Summary of the interaction models. The fitted model $\widehat{\texttt{lapp}} = \hat{\boldsymbol{\beta}}\boldsymbol{X}$, where $\boldsymbol{X} = (\mathbf{1}, \texttt{RHj}, \texttt{poly(hour, 3)1}, \texttt{poly(hour, 3)2}, \texttt{poly(hour, 3)3}, \texttt{RHj} \times \texttt{poly(hour, 3)1}, \texttt{RHj} \times \texttt{poly(hour, 3)2}, \texttt{RHj} \times \texttt{poly(hour, 3)3})$, whereby RHj represents in each case another given area humidities. The corresponding $R^2_{adj}$ for the significant model will be summarized in a following table.

We can summarize from Tables 7.11 that all the interactions are significant using the nine models

$$\texttt{lapp}_i \sim \texttt{RHj}_i * \texttt{poly(hour, 3)}_i, \quad i = 1, \ldots, 19735.$$

for $j$ = 1.kitchen, 2.living, 3.laundry, 4.office, 5.bath, 6.outside, 7.ironing, 8.teenager, 9.parents.

Observing that a few interaction terms are non-significant, but since the non-significance does not involve all the polynomial coefficients, we have an overall significant interaction. For example the interaction term RH9.parents : poly(hour, 3)1 is non-significant, but as for the other polynomial coefficient RH9.parents : poly(hour, 3)2, RH9.parents : poly(hour, 3)2 we reach significance. That means in conclusion that there is an interaction between the covariates RH9.parents and poly(hour, 3). (review in theoretical background chapter - interactions)

We will come to the same result by looking at the interaction between factor weekday and polynomial hour.

So it indicates that the model needs all the interaction terms with the polynomial transformed hour covariate.

Now, coupling all the humidities with the polynomial hourly effect with regression for $i = 1, \ldots, 19735$:

$$\begin{aligned} \texttt{lapp}_i \sim \; & ( \; \texttt{RH1.kitchen}_i + \texttt{RH2.living}_i + \texttt{RH5.laundry}_i + \texttt{RH4.office}_i + \texttt{RH5.bath}_i \\ & + \texttt{RH6.outside}_i + \texttt{RH7.ironing}_i + \texttt{RH8.teenager}_i + \texttt{RH9.parents}_i \\ & + \texttt{RH.outstation}_i) \\ & * \; (\texttt{poly(hour, 3)}_i). \end{aligned}$$

With the same arguments as before, we can say that all the interactions in the regression with all the humidities are significant at a significance level of 0.1 or even 0.05.

In addition, we prepare interaction plots, given in Figure A.7, of the regressions

$$\texttt{lapp}_i \sim \texttt{RHj}_i * \texttt{poly(hour, 3)}_i, \quad i = 1, \ldots, 19735,$$

for $j$ = `1.kitchen`, `2.living`, `3.laundry`, `4.office`, `5.bath`, `6.outside`, `7.ironing`, `8.teenager`, `9.parents`, `.outstation`, to visualize the partitioned hourly pattern with their 95 % confidence interval.

**Comparing $R^2_{adj}$ of main and interaction effect models**

| Interaction effects | |
|---|---|
| Model | $R^2_{adj}$ |
| `RH1.kitchen` * `poly(hour,3)` | 0.233 |
| `RH2.living` * `poly(hour,3)` | 0.230 |
| `RH3.laundry` * `poly(hour,3)` | 0.231 |
| `RH4.office` * `poly(hour,3)` | 0.228 |
| `RH5.bath` * `poly(hour,3)` | 0.233 |
| `RH6.outside` * `poly(hour,3)` | 0.233 |
| `RH7.ironing` * `poly(hour,3)` | 0.228 |
| `RH8.teenager` * `poly(hour,3)` | 0.230 |
| `RH9.parents` * `poly(hour,3)` | 0.228 |
| `RH.outstation` * `poly(hour,3)` | 0.235 |
| $(\sum_{j=1}^{10}$ `RHj`$)$ * `poly(hour,3)` | 0.287 |

Table 7.12: $R^2_{adj}$ for the interaction models between humidities and polynomial hourly effect.

| Main effects | |
|---|---|
| Model | $R^2_{adj}$ |
| `RH1.kitchen` + `poly(hour,3)` | 0.232 |
| `RH2.living` + `poly(hour,3)` | 0.229 |
| `RH3.laundry` + `poly(hour,3)` | 0.228 |
| `RH4.office` + `poly(hour,3)` | 0.227 |
| `RH5.bath` + `poly(hour,3)` | 0.229 |
| `RH6.outside` + `poly(hour,3)` | 0.229 |
| `RH7.ironing` + `poly(hour,3)` | 0.227 |
| `RH8.teenager` + `poly(hour,3)` | 0.227 |
| `RH9.parents` + `poly(hour,3)` | 0.227 |
| `RH.outstation` + `poly(hour,3)` | 0.228 |
| $(\sum_{j=10}^{10}$ `RHj`$)$ + `poly(hour,3)` | 0.257 |

Table 7.13: $R^2_{adj}$ for the main models for humidities and polynomial hourly effect, corresponding to Table 7.12.

In the Table 7.12 and 7.13 we also have just small improvements of the adjusted coefficients of determination using the interaction between humidity and the hours. For the interaction model using all the humidities, the greatest improvement in the $R^2_{adj}$ is observed.

### 7.3.3 Interaction between weekdays and hours

In the box-plot, c.f. Figure 7.3, we see almost the same pattern during the day, but with different variations, see the box lengths which include 50 % of the data. Furthermore we observe a higher variation at weekday Monday in the afternoon hours, that can be due to changing workflow pattern of the occupants or other factors. Nearly the same variation we notice for the weekday Friday. This supports the non-significance of the character Friday in the full main model (Model 7) as Monday is the reference. Overall the we can say that in the afternoon hours we have the highest appliances energy use by the occupants which spreads very much in the nineteenth to twenty-first hour.

Examine the interaction between the hours per weekday in Figure 7.4. We have almost identical parallel lines in the morning hours, intersecting lines in the afternoon and between hour 20 and 24 which means there exists interactions. On the one hand, we can observe intersecting lines but on the other hand they are of the same quadratic shape. We can conclude that there are clear less interaction in the morning and evening hours and interaction in the afternoon hours.

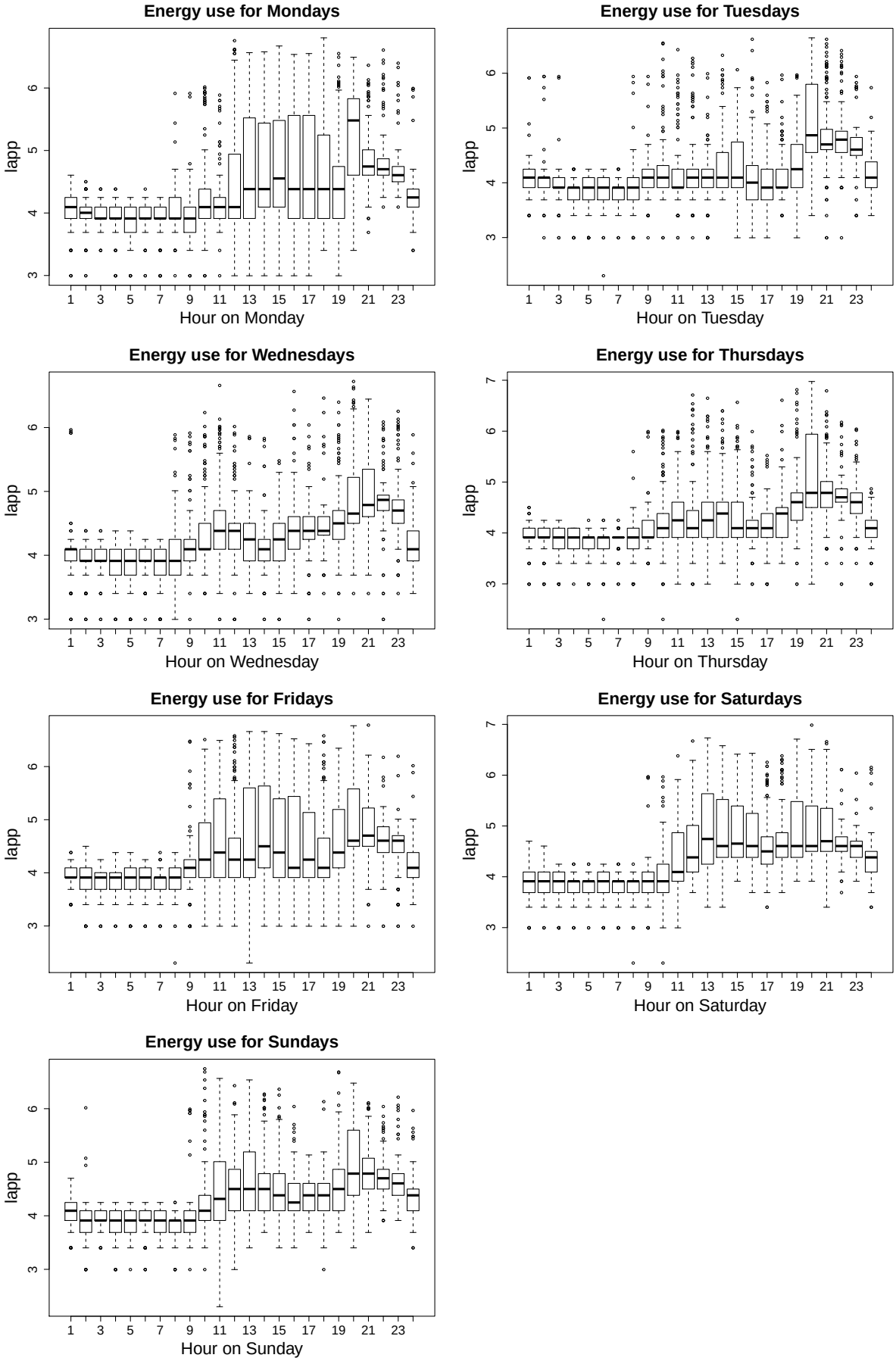



Figure 7.3: Box-plots. Appliances energy use for all weekdays relative to hours. Weekdays over the entire observation period are summarized here.

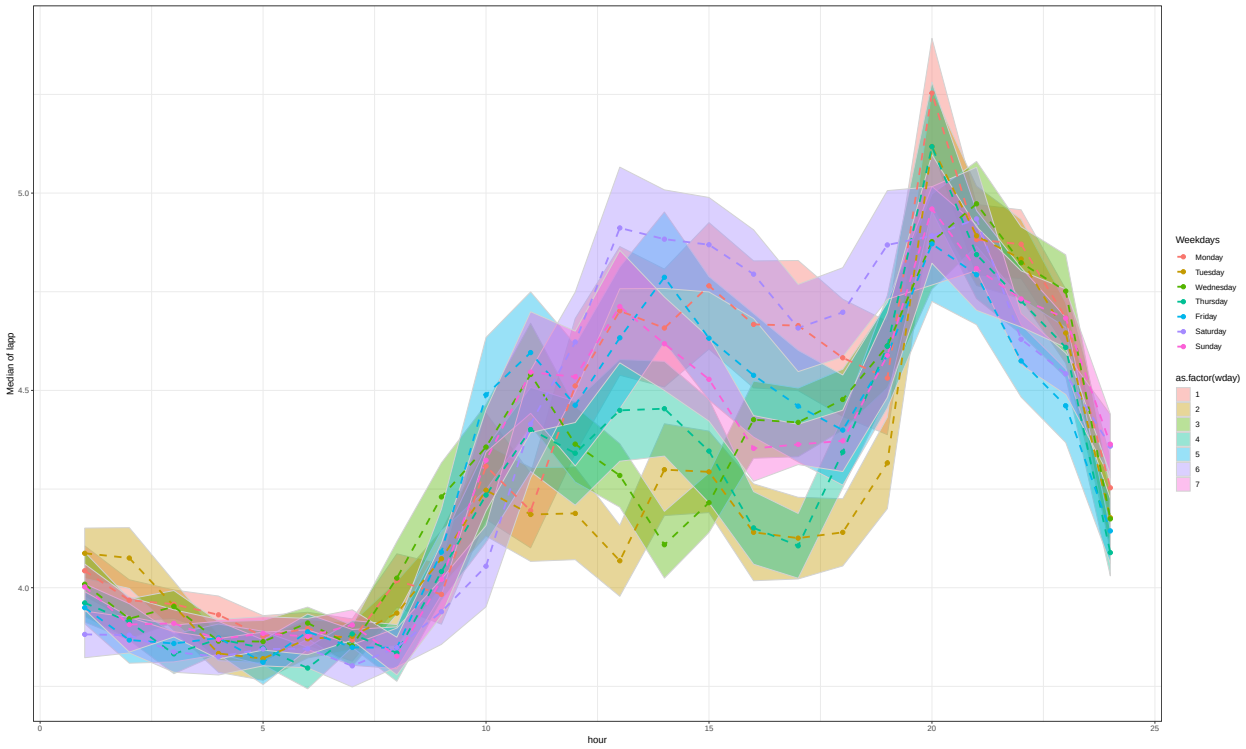



Figure 7.4: Plot of the confidence intervals for hourly mean values of `lapp` with level 0.95. The $\overline{\texttt{lapp}}_h$ and their confidence shadow is classified with different colors by weekdays.

**Regression allowing for interaction between factor weekday and polynomial hourly effect**

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | 4.3788 | 0.0108 | 405.14 | 0.0000 |
| poly(hour, 3)1 | 44.2960 | 1.5110 | 29.32 | 0.0000 |
| poly(hour, 3)2 | -8.7483 | 1.5157 | -5.77 | 0.0000 |
| poly(hour, 3)3 | -23.1180 | 1.5161 | -15.25 | 0.0000 |
| weekday.Tuesday | -0.1619 | 0.0151 | -10.70 | 0.0000 |
| weekday.Wednesday | -0.0854 | 0.0151 | -5.64 | 0.0000 |
| weekday.Thursday | -0.1365 | 0.0151 | -9.02 | 0.0000 |
| weekday.Friday | -0.0677 | 0.0152 | -4.46 | 0.0000 |
| weekday.Saturday | -0.0101 | 0.0153 | -0.66 | 0.5087 |
| weekday.Sunday | -0.0581 | 0.0153 | -3.79 | 0.0002 |
| poly(hour, 3)1:weekday.Tuesday | -10.6658 | 2.1215 | -5.03 | 0.0000 |
| poly(hour, 3)2:weekday.Tuesday | 16.7471 | 2.1246 | 7.88 | 0.0000 |
| poly(hour, 3)3:weekday.Tuesday | 12.0955 | 2.1253 | 5.69 | 0.0000 |
| poly(hour, 3)1:weekday.Wednesday | -5.9573 | 2.1215 | -2.81 | 0.0050 |
| poly(hour, 3)2:weekday.Wednesday | 7.1011 | 2.1246 | 3.34 | 0.0008 |
| poly(hour, 3)3:weekday.Wednesday | 11.8700 | 2.1253 | 5.58 | 0.0000 |
| poly(hour, 3)1:weekday.Thursday | -6.2979 | 2.1215 | -2.97 | 0.0030 |
| poly(hour, 3)2:weekday.Thursday | 5.1126 | 2.1246 | 2.41 | 0.0161 |
| poly(hour, 3)3:weekday.Thursday | 8.7334 | 2.1253 | 4.11 | 0.0000 |
| poly(hour, 3)1:weekday.Friday | -7.4167 | 2.1308 | -3.48 | 0.0005 |
| poly(hour, 3)2:weekday.Friday | -9.2724 | 2.1329 | -4.35 | 0.0000 |
| poly(hour, 3)3:weekday.Friday | 5.3033 | 2.1314 | 2.49 | 0.0128 |
| poly(hour, 3)1:weekday.Saturday | 5.0501 | 2.1488 | 2.35 | 0.0188 |
| poly(hour, 3)2:weekday.Saturday | -9.4597 | 2.1519 | -4.40 | 0.0000 |
| poly(hour, 3)3:weekday.Saturday | -3.4610 | 2.1526 | -1.61 | 0.1079 |
| poly(hour, 3)1:weekday.Sunday | -3.9890 | 2.1511 | -1.85 | 0.0637 |
| poly(hour, 3)2:weekday.Sunday | 1.0883 | 2.1531 | 0.51 | 0.6132 |
| poly(hour, 3)3:weekday.Sunday | 10.0374 | 2.1527 | 4.66 | 0.0000 |

Table 7.14: Summary of the significant interaction model `lapp` $\sim$ `weekday` $*$ `poly(hour, 3)`, with $R^2_{adj} = 0.2485$ and $F$-statistic $= 242.7$ on 27 and 19707 DF.

### Comparing $R^2_{adj}$ of the main and interaction effect model

Considering just the influence of the weekdays and the polynomial hourly effect on our response variable `lapp`, we can also observe improvement adding the interaction term, see the Table 7.15 and 7.16, which gives us the corresponding adjusted coefficient of determinations.

| Model: Interaction effect | $R^2_{adj}$ |
|---|---|
| `lapp` $\sim$ `weekday` $*$ `poly(hour,3)` | 0.249 |

Table 7.15: $R^2_{adj}$ for the interaction model between the factor weekday and polynomial hourly effect.

| Model: Main effect | $R^2_{adj}$ |
|---|---|
| `lapp` $\sim$ `weekday` $+$ `poly(hour,3)` | 0.234 |

Table 7.16: $R^2_{adj}$ for the main model between the factor weekday and polynomial hourly effect.

**F-test - Testing for interacting time effects**

At this position, we want to analyze if an interaction between `poly(hour, d)` and `weekday` is needed. To do so, we will perform three $F$-tests to evaluate the form of interaction.

The three tests will have the specification of the covariates polynomial degree $d$.

(i) $\underline{d = 1}$: The interaction with the weekdays has a linear form for the hour.

(ii) $\underline{d = 2}$: The interaction with the weekdays has a quadratic form for the hour.

(iii) $\underline{d = 3}$: The interaction with the weekdays has a cubic form for the hour.

We already excluded the fourth degree due to non-significance, see Section 7.1.

With help of the ANOVA-Table from R-package `stats`, we will determine the F-tests. ANOVA is the analysis of variance, we already discussed in the section 3.5.1. For a deeper inside on this topic with good examples, we refer to Christensen (2018).

The test hypothesis $H_0$ are stated as follows:

- Case 1: There is no difference in the means of `lapp` grouped by `poly(hour, d)`$_{d=1,2,3}$.
- Case 2: There is no difference in means of `lapp` grouped by factor `weekday`.
- Case 3: There is no interaction between `poly(hour, d)`$_{d=1,2,3}$ and `weekday`.

Whereas the alternative hypothesis $H_1$ for Case 1 and 2 is that the means are not equal, and for the Case 3 that there is an interaction between `poly(hour, d)`$_{d=1,2,3}$ and `weekday`.

As we observed in Figure 7.4, there are clearly some intersections which means interaction. So we just want to verify the interaction and look which interaction form fits the data the best.

Before we start note that we assume that the observations within each cell are normally distributed and have equal variances. We will check these assumptions after fitting ANOVA.

We already visualized the data by box-plots (see Figure 7.3) and line plots with confidence intervals (see Figure 7.4) to identify group differences by weekdays. The box-plots reveal the energy consumption conditioned on the combinations of the levels of `weekday` and `hour`. Whereas the interaction plot illustrate possible interactions which are visible through the intersections, especially in the afternoon hours. The two-way interaction plot (Figure 7.4) visualize the mean with a confidence interval level of 0.95 of the response for

two-way combinations of `weekday` and `hour`, where `hour` are plotted on the x-axis and the factor `weekday` as seven different weekday lines.

**1. F-test of Model** `lapp` $\sim$ `weekday` $*$ `poly(hour, 3)`, **($d = 3$):**

Finally we compute a F-test with help of the analysis of variance. Is our response `lapp` depending on `weekday` and `hour`? And in particular which interaction form is needed?

First we review just the main effect in Table 7.17. From the ANOVA table we can conclude that both our covariates `weekday` and `poly(hour, 3)` are statistically significant. `poly(hour, 3)` is the even more significant covariate due to the high $F$-value. (These results would lead us to believe that changing or removing the weekday and the hour would impact significantly the mean of response `lapp`.)

| | Df | Sum Sq | Mean Sq | F value | Pr(>F) |
|---|---|---|---|---|---|
| `weekday` | 6 | 62.80 | 10.47 | 31.73 | 0.0000 |
| `poly(hour, 3)` | 3 | 1930.38 | 643.46 | 1950.55 | 0.0000 |
| Residuals | 19725 | 6507.01 | 0.33 | | |

Table 7.17: Summary of the ANOVA model with main effects. The model been used is $\widehat{\texttt{lapp}}_i = \hat{\beta}_0 + \sum_{j=1}^{6} \hat{\beta}_j \texttt{weekday.(j+1)}_i + \hat{\beta}_7(\texttt{poly(hour, 3)1})_i + \hat{\beta}_8(\texttt{poly(hour, 3)2})_i + \hat{\beta}_9(\texttt{poly(hour, 3)3})_i$. The output includes the $F$-value and the corresponding $p$-value of the test.

After inspecting the additive models with the assumption of two independent variables, we add the interaction term since we conjecture that the two covariates might interact. By including the synergistic effect, i.e. using the model

$$\begin{aligned}\widehat{\texttt{lapp}}_i = \;& \hat{\beta}_0 + \textstyle\sum_{j=1}^{6} \hat{\beta}_j \texttt{weekday.(j+1)}_i \\ & + \hat{\beta}_7(\texttt{poly(hour, 3)1})_i + \hat{\beta}_8(\texttt{poly(hour, 3)2})_i + \hat{\beta}_9(\texttt{poly(hour, 3)3})_i \\ & + \textstyle\sum_{j=10}^{15} \hat{\beta}_j (\texttt{poly(hour, 3)1})_i \times \texttt{weekday.(j-8)}_i \\ & + \textstyle\sum_{j=16}^{21} \hat{\beta}_j (\texttt{poly(hour, 3)2})_i \times \texttt{weekday.(j-14)}_i \\ & + \textstyle\sum_{j=22}^{27} \hat{\beta}_j (\texttt{poly(hour, 3)3})_i \times \texttt{weekday.(j-20)}, \quad i = 1, \ldots, 19735\end{aligned}$$

we get the results in Table 7.18. It can be seen that the two main effects are still statistically significant. The added interaction effect is also statistically significant, which means that we should use the interaction model.

Based on the $p$-values and a significance level of 0.05, we can conclude from the results of ANOVA Table 7.17 and 7.18 that we can reject the null hypothesis. Due to significant $p$-values smaller than $2e-16$, the levels of both `weekday` and `poly(hour, 3)` are associated with significant different energy consumption, with the response variable `lapp`. Further the $p$-value for the interaction between `weekday` and `poly(hour, 3)` are also highly significant, i.e. p-value $< 2e-16$, which indicates the relationship between the hour and the energy consumption depends on the weekday.

| | Df | Sum Sq | Mean Sq | F value | Pr(>F) |
|---|---|---|---|---|---|
| `weekday` | 6 | 62.80 | 10.47 | 32.34 | 0.0000 |
| `poly(hour, 3)` | 3 | 1930.38 | 643.46 | 1987.88 | 0.0000 |
| `weekday:poly(hour, 3)` | 18 | 128.00 | 7.11 | 21.97 | 0.0000 |
| Residuals | 19707 | 6379.01 | 0.32 | | |

Table 7.18: Summary of the ANOVA model with interaction effect. The model been used is `lapp` $\sim$ `weekday` + `poly(hour, 3)` + `weekday` : `poly(hour, 3)`. The output includes the $F$-value and the corresponding $p$-value of the test. Residual standard error is given by 0.5689.

Multiple pairwise-comparison between the means of groups

Due to the significant ANOVA test, we will reject the null hypothesis and conclude different group means. But that arises the question, which pairs of groups differ from the others. To get a closer look, we determine if the mean difference between specific pairs of group are statistically significant. With the help of a pairwise t-test, we get the following result:

| | Monday | Tuesday | Wednesday | Thursday | Friday | Saturday |
|---|---|---|---|---|---|---|
| Tuesday | $< 2e-16$ | - | - | - | - | - |
| Wednesday | $7.7e-07$ | $2.1e-05$ | - | - | - | - |
| Thursday | $4.4e-15$ | 0.16490 | 0.00423 | - | - | - |
| Friday | $5.4e-05$ | $2.4e-07$ | 0.36943 | 0.00015 | - | - |
| Saturday | 0.42882 | $< 2e-16$ | $3.6e-05$ | $2.6e-12$ | 0.00126 | - |
| Sunday | 0.00080 | $8.8e-09$ | 0.13101 | $1.5e-05$ | 0.50899 | 0.01057 |

Table 7.19: Pairwise comparison using t-tests with pooled standard deviation (i.e. the corresponding formula is $SD_{pooled} = \sqrt{\frac{\sum_{g=1}^{s}(n_g-1)SD_g^2}{\sum_{g=1}^{s} n_g - s}}$, where $SD_g$ is the standard deviation and $n_g$ the sample size for group $g$). The response `lapp` and the weekdays being used. For the multiple comparison the $p$-value adjustment method "BH" (c.f. Benjamini and Hochberg (1995)) is used which is shown in this table.

Note that we only performed the test for the covariate `weekday`. This is because of simplicity and the `poly(hour, 3)` had highly significant differences in the ANOVA tables. Now validate the results in Table 7.19, it can be seen that almost all pairwise comparisons are significant with an adjusted p-value $< 0.05$. The highest significant difference in energy consumption is between Tuesday and Monday ($p \approx 0.00$), as well as between Saturday and Tuesday ($p \approx 0.00$). The exception is for the pairs (Saturday, Monday), (Thursday, Tuesday) and all paired combination of weekdays (Wednesday, Friday, Sunday), i.e. (Friday, Wednesday), (Sunday, Wednesday), (Sunday, Friday), where there is no significant difference as the corresponding p-values in the Table 7.19 are higher than the level of 0.05.

**2. Testing interaction form, specified by different $d$, ($d = 1, 2$):**

Before we check the model assumption, we want to decide on the interaction form as we already discussed above. For $d = 3$ we already did our analysis of variance on the interaction model. So there is still to determine the ANOVA for the cases $d = 1$ and $d = 2$.

Case d = 1:

Since the main effect model is already highly significant for both covariates, we directly focus on the interaction model. Here we also have highly significant differences which can be seen in the Table 7.20. For this Table we use the model

$$\begin{aligned}\widehat{\texttt{lapp}}_i &= \hat{\beta}_0 + \sum_{j=1}^{6} \hat{\beta}_j \texttt{weekday.(j+1)}_i \\ &\quad + \hat{\beta}_7 \texttt{(poly(hour, 1)1)}_i \\ &\quad + \sum_{j=8}^{13} \hat{\beta}_j \texttt{(poly(hour, 1)1)}_i \times \texttt{weekday.(j-6)}_i \\ &= \hat{\beta}_0 + \sum_{j=1}^{6} \hat{\beta}_j \texttt{weekday.(j+1)}_i + \hat{\beta}_7 \texttt{hour}_i \\ &\quad + \sum_{j=8}^{13} \hat{\beta}_j \texttt{hour}_i \times \texttt{weekday.(j-6)}_i\end{aligned}$$

| | Df | Sum Sq | Mean Sq | F value | Pr(>F) |
|---|---|---|---|---|---|
| weekday | 6 | 62.80 | 10.47 | 30.31 | 0.0000 |
| poly(hour, 1) | 1 | 1604.17 | 1604.17 | 4644.88 | 0.0000 |
| weekday:poly(hour, 1) | 6 | 22.34 | 3.72 | 10.78 | 0.0000 |
| Residuals | 19721 | 6810.89 | 0.35 | | |

Table 7.20: Analysis of variance table. Model with interaction effect $\texttt{lapp} \sim \texttt{weekday} + \texttt{poly(hour, 1)} + \texttt{weekday} : \texttt{poly(hour, 1)}$. Residual standard error is 0.5877.

Case d = 2:

Also for this case we have significant main effects, so we proceed with the interaction model:

$$\begin{aligned}\widehat{\texttt{lapp}}_i = \; & \hat{\beta}_0 + \sum_{j=1}^{6} \hat{\beta}_j \texttt{weekday.(j+1)}_i \\ & + \hat{\beta}_7 \texttt{(poly(hour, 2)1)}_i + \hat{\beta}_8 \texttt{(poly(hour, 2)2)}_i \\ & + \sum_{j=9}^{14} \hat{\beta}_j \texttt{(poly(hour, 2)1)}_i \times \texttt{weekday.(j-7)}_i \\ & + \sum_{j=15}^{20} \hat{\beta}_j \texttt{(poly(hour, 2)2)}_i \times \texttt{weekday.(j-13)}_i\end{aligned}$$

In the Table 7.21 we have again highly significant differences, that is we reject the null hypothesis.

| | Df | Sum Sq | Mean Sq | F value | Pr(>F) |
|---|---|---|---|---|---|
| weekday | 6 | 62.80 | 10.47 | 30.85 | 0.0000 |
| poly(hour, 2) | 2 | 1653.24 | 826.62 | 2436.59 | 0.0000 |
| weekday:poly(hour, 2) | 12 | 96.11 | 8.01 | 23.61 | 0.0000 |
| Residuals | 19714 | 6688.05 | 0.34 | | |

Table 7.21: Analysis of variance table. Model with interaction effect $\texttt{lapp} \sim \texttt{weekday} + \texttt{poly(hour, 2)} + \texttt{weekday} : \texttt{poly(hour, 2)}$. Residual standard error is 0.5825.

**3. Conclusion for the interaction effect:**

All polynomial degrees, i.e. $d = 1, 2, 3$, are highly significant which means that they indicate a significant interactions for the `weekday` and `poly(hour, d)`$_{d=1,2,3}$. To decide which degree we should use for the interaction, we consider the $F$-values. A large $F$-value indicates more difference between the groups than within the groups and therefore a greater effect on the response `lapp`. Case $d = 1$ has the lowest $F$-value with 10.78, which leads to a decision between $d = 2$ and $d = 3$. Even though the $F$-value of the interaction term for the polynomial degree two is higher with $23.61 > 21.97$, the sum squares and the degree of freedom speaks for the degree of three.

We proceed with the interaction term between `weekday` and `poly(hour, 3)`.

**4. Check ANOVA assumptions:**

ANOVA assumes that the data are normally distributed and the variance across groups are homogeneous. Now exploring the assumption with help of some diagnostic plots.

Check the homogeneity of variance assumption

The residuals versus fits plot is used to check the homogeneity of variances. In the Figure 7.5, there is not a significant evident relationships between residuals and fitted values, i.e. the mean of each groups, which is sufficient. For Case 3 (Figure 7.5) it looks more random than for Case 2 (Figure 7.6). So, we can assume the homogeneity of variances for Case 3.

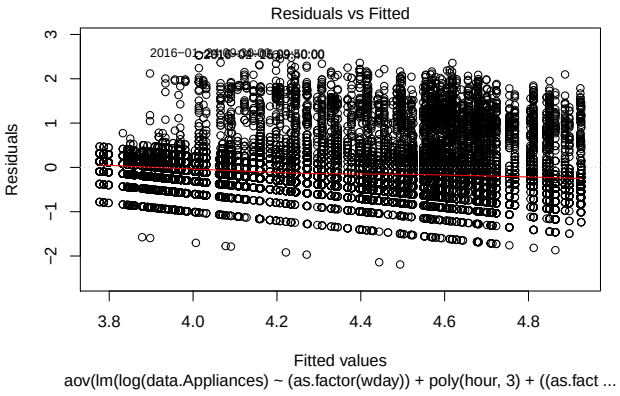



Figure 7.5: Residuals versus fits. The plot checks the homogeneity of variances for the model `lapp` $\sim$ `weekday` $*$ `poly(hour,3)`.

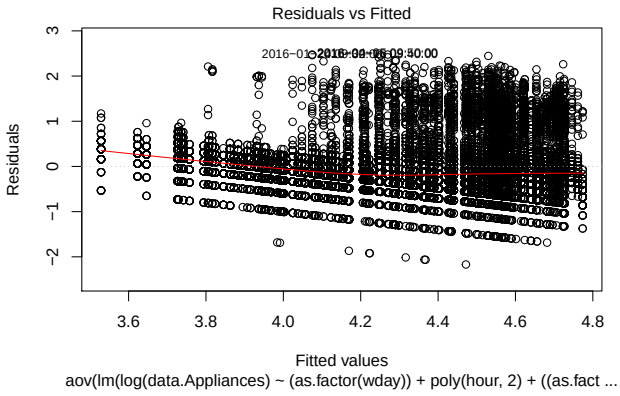



Figure 7.6: Residuals versus fits. The plot checks the homogeneity of variances for the model `lapp` $\sim$ `weekday` $*$ `poly(hour,2)`.

Check the normality assumption

Next, we want to verify the assumption that the residuals are normally distributed. To do this, the quantiles of the residuals are plotted against the quantiles of the normal

distribution. The 45-degree reference line is the optimal case which the normal probability plot of the residuals should be follow approximately.

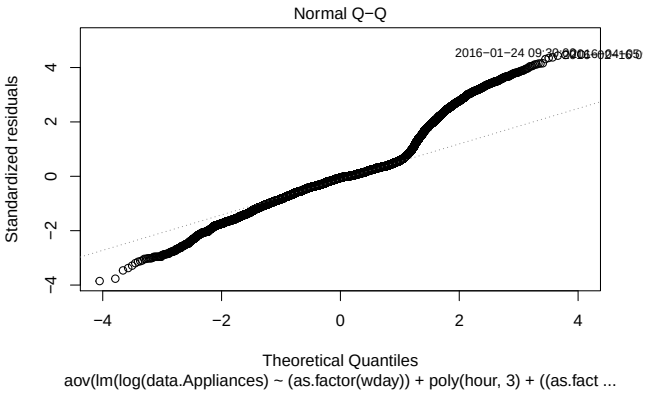



Figure 7.7: Normality plot of the residuals. Figure shows non-normality for the ANOVA model `lapp` $\sim$ `weekday` $*$ `poly(hour,3)`.

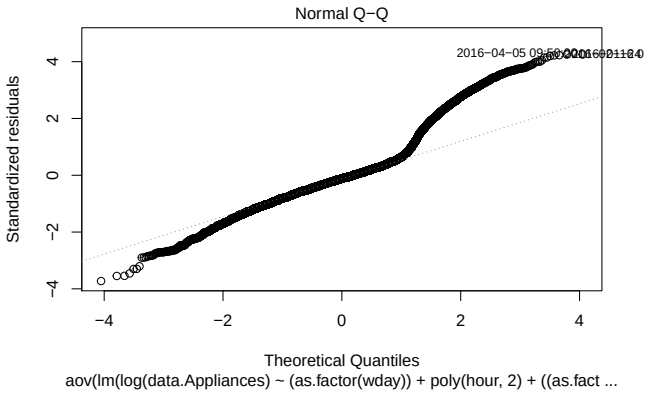



Figure 7.8: Normality plot of the residuals. Figure shows non-normality for the ANOVA model `lapp` $\sim$ `weekday` $*$ `poly(hour,2)`.

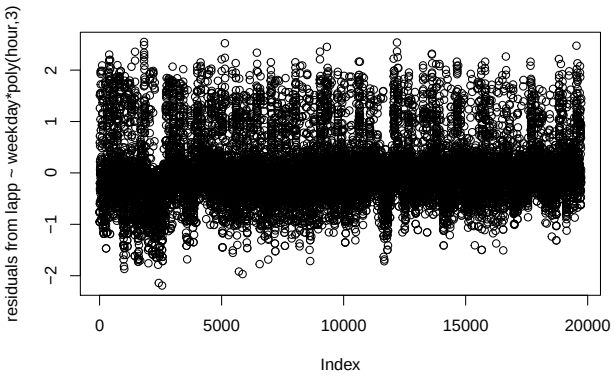



Figure 7.9: Normality plot of the ANOVA residuals. Figure shows the ANOVA model `lapp` $\sim$ `weekday` $*$ `poly(hour,3)`.

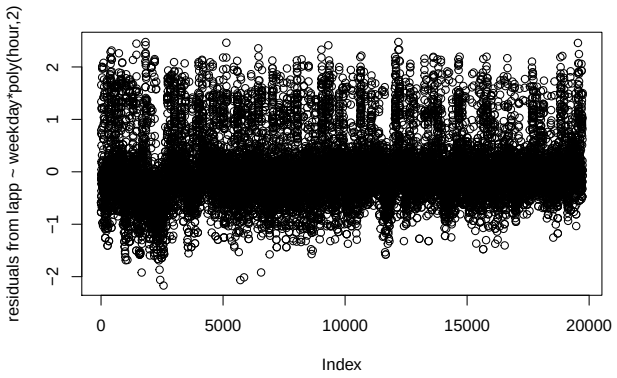



Figure 7.10: Normality plot of the ANOVA residuals. Figure shows the ANOVA model `lapp` $\sim$ `weekday` $*$ `poly(hour,2)`.

If we can assume normality we additionally perform a plot where we extract the residuals from the corresponding ANOVA model Case 3 in Table 7.18 and Case 2 in Table 7.21. The resulting figures for Case 3 are the Figure 7.7, 7.9 and Case 2 the Figure 7.8, 7.10 showing indication whether normality is violated or not. Unfortunately the can see in Figure 7.7 and 7.8 that the normality assumption is violated due to outliers. However, from these normality plots, we can not detect big differences between the cases. But since our final interaction model (LMinter) reaches an improved and satisfying $R^2_{adj}$, we use this transformation either way and continue with this transformed hourly effect.

These assumptions strengthen and supports the decision to continue working with the third degree polynomial as the interaction form, i.e. we use the model

$$\begin{aligned}\widehat{\texttt{lapp}}_i = \;& \hat{\beta}_0 + \textstyle\sum_{j=1}^{6} \hat{\beta}_j \texttt{weekday.(j+1)}_i \\ & + \hat{\beta}_7(\texttt{poly(hour, 3)1})_i + \hat{\beta}_8(\texttt{poly(hour, 3)2})_i + \hat{\beta}_9(\texttt{poly(hour, 3)3})_i \\ & + \textstyle\sum_{j=10}^{15} \hat{\beta}_j(\texttt{poly(hour, 3)1})_i \times \texttt{weekday.(j-8)}_i \\ & + \textstyle\sum_{j=16}^{21} \hat{\beta}_j(\texttt{poly(hour, 3)2})_i \times \texttt{weekday.(j-14)}_i \\ & + \textstyle\sum_{j=22}^{27} \hat{\beta}_j(\texttt{poly(hour, 3)3})_i \times \texttt{weekday.(j-20)}, \quad i = 1, \ldots, 19735.\end{aligned}$$

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 2 | 0.0428 | - | - | - | - | - | - | - | - | - | - | - |
| 3 | 0.0013 | 0.2463 | - | - | - | - | - | - | - | - | - | - |
| 4 | $8.4e-06$ | 0.0173 | 0.2404 | - | - | - | - | - | - | - | - | - |
| 5 | $3.9e-07$ | 0.0027 | 0.0732 | 0.5586 | - | - | - | - | - | - | - | - |
| 6 | $1.3e-05$ | 0.0221 | 0.2770 | 0.9282 | 0.4994 | - | - | - | - | - | - | - |
| 7 | $4.3e-06$ | 0.0116 | 0.1877 | 0.8913 | 0.6525 | 0.8259 | - | - | - | - | - | - |
| 8 | 0.0023 | 0.3252 | 0.8659 | 0.1754 | 0.0481 | 0.2067 | 0.1344 | - | - | - | - | - |
| 9 | 0.0217 | $1.2e-05$ | $2.6e-08$ | $1.0e-11$ | $1.2e-13$ | $2.0e-11$ | $3.8e-12$ | $6.9e-08$ | - | - | - | - |
| 10 | $\star\star\star$ | $\star\star\star$ | $\star\star\star$ | $\star\star\star$ | $\star\star\star$ | $\star\star\star$ | $\star\star\star$ | $\star\star\star$ | $\star\star\star$ | - | - | - |
| 11 | $\star\star\star$ | $\star\star\star$ | $\star\star\star$ | $\star\star\star$ | $\star\star\star$ | $\star\star\star$ | $\star\star\star$ | $\star\star\star$ | $\star\star\star$ | $1.1e-05$ | - | - |
| 12 | $\star\star\star$ | $\star\star\star$ | $\star\star\star$ | $\star\star\star$ | $\star\star\star$ | $\star\star\star$ | $\star\star\star$ | $\star\star\star$ | $\star\star\star$ | $4.8e-07$ | 0.5540 | - |
| 13 | $\star\star\star$ | $\star\star\star$ | $\star\star\star$ | $\star\star\star$ | $\star\star\star$ | $\star\star\star$ | $\star\star\star$ | $\star\star\star$ | $\star\star\star$ | $\star\star\star$ | $1.7e-05$ | 0.0002 |
| 14 | $\star\star\star$ | $\star\star\star$ | $\star\star\star$ | $\star\star\star$ | $\star\star\star$ | $\star\star\star$ | $\star\star\star$ | $\star\star\star$ | $\star\star\star$ | $\star\star\star$ | $4.3e-06$ | $7.3e-05$ |
| 15 | $\star\star\star$ | $\star\star\star$ | $\star\star\star$ | $\star\star\star$ | $\star\star\star$ | $\star\star\star$ | $\star\star\star$ | $\star\star\star$ | $\star\star\star$ | $\star\star\star$ | 0.0002 | 0.0018 |
| 16 | $\star\star\star$ | $\star\star\star$ | $\star\star\star$ | $\star\star\star$ | $\star\star\star$ | $\star\star\star$ | $\star\star\star$ | $\star\star\star$ | $\star\star\star$ | $1.6e-07$ | 0.4250 | 0.8464 |
| 17 | $\star\star\star$ | $\star\star\star$ | $\star\star\star$ | $\star\star\star$ | $\star\star\star$ | $\star\star\star$ | $\star\star\star$ | $\star\star\star$ | $\star\star\star$ | 0.0001 | 0.5848 | 0.2454 |
| 18 | $\star\star\star$ | $\star\star\star$ | $\star\star\star$ | $\star\star\star$ | $\star\star\star$ | $\star\star\star$ | $\star\star\star$ | $\star\star\star$ | $\star\star\star$ | $6.3e-07$ | 0.5848 | 0.9575 |
| 19 | $\star\star\star$ | $\star\star\star$ | $\star\star\star$ | $\star\star\star$ | $\star\star\star$ | $\star\star\star$ | $\star\star\star$ | $\star\star\star$ | $\star\star\star$ | $\star\star\star$ | $3.4e-10$ | $1.6e-08$ |
| 20 | $\star\star\star$ | $\star\star\star$ | $\star\star\star$ | $\star\star\star$ | $\star\star\star$ | $\star\star\star$ | $\star\star\star$ | $\star\star\star$ | $\star\star\star$ | $\star\star\star$ | $\star\star\star$ | $\star\star\star$ |
| 21 | $\star\star\star$ | $\star\star\star$ | $\star\star\star$ | $\star\star\star$ | $\star\star\star$ | $\star\star\star$ | $\star\star\star$ | $\star\star\star$ | $\star\star\star$ | $\star\star\star$ | $\star\star\star$ | $\star\star\star$ |
| 22 | $\star\star\star$ | $\star\star\star$ | $\star\star\star$ | $\star\star\star$ | $\star\star\star$ | $\star\star\star$ | $\star\star\star$ | $\star\star\star$ | $\star\star\star$ | $\star\star\star$ | $\star\star\star$ | $\star\star\star$ |
| 23 | $\star\star\star$ | $\star\star\star$ | $\star\star\star$ | $\star\star\star$ | $\star\star\star$ | $\star\star\star$ | $\star\star\star$ | $\star\star\star$ | $\star\star\star$ | $\star\star\star$ | $1.2e-14$ | $1.3e-12$ |
| 24 | $\star\star\star$ | $\star\star\star$ | $\star\star\star$ | $\star\star\star$ | $\star\star\star$ | $\star\star\star$ | $\star\star\star$ | $\star\star\star$ | $2.6e-09$ | 0.0169 | $7.6e-12$ | $7.8e-14$ |

| | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 2 | - | - | - | - | - | - | - | - | - | - | - |
| 3 | - | - | - | - | - | - | - | - | - | - | - |
| 4 | - | - | - | - | - | - | - | - | - | - | - |
| 5 | - | - | - | - | - | - | - | - | - | - | - |
| 6 | - | - | - | - | - | - | - | - | - | - | - |
| 7 | - | - | - | - | - | - | - | - | - | - | - |
| 8 | - | - | - | - | - | - | - | - | - | - | - |
| 9 | - | - | - | - | - | - | - | - | - | - | - |
| 10 | - | - | - | - | - | - | - | - | - | - | - |
| 11 | - | - | - | - | - | - | - | - | - | - | - |
| 12 | - | - | - | - | - | - | - | - | - | - | - |
| 13 | - | - | - | - | - | - | - | - | - | - | - |
| 14 | 0.7895 | - | - | - | - | - | - | - | - | - | - |
| 15 | 0.6054 | 0.4250 | - | - | - | - | - | - | - | - | - |
| 16 | 0.0005 | 0.0002 | 0.0035 | - | - | - - | - | - | - | - | - |
| 17 | $1.1e-06$ | $2.3e-07$ | $1.5e-05$ | 0.1724 | - | - | - | - | - | - | - |
| 18 | 0.0002 | $5.9e-05$ | 0.0015 | 0.8104 | 0.2662 | - | - | - | - | - | - |
| 19 | 0.0548 | 0.1043 | 0.0137 | $5.4e-08$ | $7.0e-12$ | $1.2e-08$ | - | - | - | - | - |
| 20 | $\star\star\star$ | $\star\star\star$ | $\star\star\star$ | $\star\star\star$ | $\star\star\star$ | $\star\star\star$ | $\star\star\star$ | - | - | - | - |
| 21 | $\star\star\star$ | $\star\star\star$ | $\star\star\star$ | $\star\star\star$ | $\star\star\star$ | $\star\star\star$ | $\star\star\star$ | $7.0e-07$ | - | - | - |
| 22 | $2.9e-14$ | $2.6e-13$ | $3.7e-16$ | $\star\star\star$ | $\star\star\star$ | $\star\star\star$ | $1.6e-08$ | $\star\star\star$ | $2.1e-06$ | - | - |
| 23 | 0.0008 | 0.0021 | $9.1e-05$ | $5.9e-12$ | $\star\star\star$ | $9.1e-13$ | 0.1609 | $\star\star\star$ | $\star\star\star$ | $2.8e-05$ | - |
| 24 | $\star\star\star$ | $\star\star\star$ | $\star\star\star$ | $1.6e-14$ | $3.7e-10$ | $1.2e-13$ | $\star\star\star$ | $\star\star\star$ | $\star\star\star$ | $\star\star\star$ | $\star\star\star$ |

Table 7.22: Pairwise comparison using $t$-test with pooled standard deviation. The response variable `lapp` and the covariate `hour` being used. For $p$-value$< 0.05$, there is a significant difference between the paired hours, whereas for $p$-value$> 0.05$ there is no significant difference between the paired hours. $\star\star\star$ denotes the $p$-values $< 2e-16$.

### 7.3.4 Interaction model - Analyzing the interaction term

#### The full interaction model

Now summarizing all the results from the above interaction analysis and forming a model using all the covariates interacting with the hourly time effect, i.e. `poly(hour, 3)`.

We set up the full interaction model (LMinter) as follows:

$$\begin{aligned}\widehat{\texttt{lapp}}_i = \hat{\beta}_0 \quad &+\big( \quad \hat{\beta}_1\texttt{T1.kitchen}_i + \hat{\beta}_2\texttt{T2.living}_i + \hat{\beta}_3\texttt{T3.laundry}_i + \hat{\beta}_4\texttt{T4.office}_i \\ &\quad +\hat{\beta}_5\texttt{T5.bath}_i + \hat{\beta}_6\texttt{T6.outside}_i + \hat{\beta}_7\texttt{T7.ironing}_i + \hat{\beta}_8\texttt{T8.teenager}_i \\ &\quad +\hat{\beta}_9\texttt{T9.parents}_i + \hat{\beta}_{10}\texttt{T.outstation}_i \\ &\quad +\hat{\beta}_{11}\texttt{RH1.kitchen}_i + \hat{\beta}_{12}\texttt{RH2.living}_i + \hat{\beta}_{13}\texttt{RH3.laundry}_i + \hat{\beta}_{14}\texttt{RH4.office}_i \\ &\quad +\hat{\beta}_{15}\texttt{RH5.bath}_i + \hat{\beta}_{16}\texttt{RH6.outside}_i + \hat{\beta}_{17}\texttt{RH7.ironing}_i + \hat{\beta}_{18}\texttt{RH8.teenager}_i \\ &\quad +\hat{\beta}_{19}\texttt{RH9.parents}_i + \hat{\beta}_{20}\texttt{RH.outstation}_i \\ &\quad +\hat{\beta}_{21}\texttt{weekday.Tuesday}_i + \hat{\beta}_{22}\texttt{weekday.Wednesday}_i + \hat{\beta}_{23}\texttt{weekday.Thursday}_i \\ &\quad +\hat{\beta}_{24}\texttt{weekday.Friday}_i + \hat{\beta}_{25}\texttt{weekday.Saturday}_i + \hat{\beta}_{26}\texttt{weekday.Sunday}_i \quad \big) \\ &*\big( \quad \hat{\beta}_{27}\texttt{poly(hour, 3)1}_i + \hat{\beta}_{28}\texttt{poly(hour, 3)2}_i + \hat{\beta}_{29}\texttt{poly(hour, 3)3}_i \quad \big),\end{aligned}$$

for $i = 1, \ldots, 19735$ with its summary in Table 7.23. Defining the formula with the operator $*$ as follows: $a * b = a + b + ab$, with $a, b \in \mathbb{R}$. By including the interaction effect we achieve an adjusted coefficient of determination of $R^2_{adj} = 0.393$.

#### The reduced interaction model

We have be cautious when we eliminate non-significant covariates. The main effects `T5.bath`, `RH4.office` and `RH8.teenager` are non-significant at the 10% level, but the interaction terms including `RH4.office` and `RH8.teenager` are definitely significant. This means we can not remove these main effects, since we can not use it for the interaction terms anymore. Analyze all the interaction terms, there is just one non-significant interaction to eliminate. The interaction term including `T5.bath` is in all three polynomial component non-significant. Since `T5.bath` is also non-significant in its main effect, we not only exclude the interaction with it but also the covariate itself. Thus we obtain the following summary of the reduced interaction model (RedInterModel) in Table 7.24.

Note that we only reduced the interaction term `T5.bath`×`poly(hour,3)` which showed no improvement in setting the interaction in contrast with the main effect `T5.bath` + `poly(hour,3)`, see Table 7.9 and 7.10, that both showed an adjusted coefficient of determination of 0.238. Moreover, with setting the reduced interaction model (RedInterModel) we do not improve the coefficient of determination as it is again $R^2_{adj} = 0.393$.

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | 2.34 | 0.11 | 20.92 | 0.00 |
| T1.kitchen | -0.03 | 0.01 | -1.78 | 0.07 |
| T2.living | -0.06 | 0.01 | -4.45 | 0.00 |
| T3.laundry | 0.10 | 0.01 | 14.87 | 0.00 |
| T4.office | 0.05 | 0.01 | 8.38 | 0.00 |
| T5.bath | 0.01 | 0.01 | 1.13 | 0.26 |
| T6.outside | -0.01 | 0.00 | -2.15 | 0.03 |
| T7.ironing | -0.02 | 0.01 | -2.37 | 0.02 |
| T8.teenager | 0.13 | 0.01 | 19.68 | 0.00 |
| T9.parents | -0.13 | 0.01 | -11.28 | 0.00 |
| T.outstation | 0.01 | 0.00 | 2.95 | 0.00 |
| RH1.kitchen | 0.04 | 0.01 | 5.83 | 0.00 |
| RH2.living | -0.05 | 0.01 | -9.46 | 0.00 |
| RH3.laundry | 0.04 | 0.00 | 10.51 | 0.00 |
| RH4.office | -0.00 | 0.00 | -0.70 | 0.48 |
| RH5.bath | 0.00 | 0.00 | 6.41 | 0.00 |
| RH6.outside | -0.00 | 0.00 | -6.35 | 0.00 |
| RH7.ironing | -0.01 | 0.00 | -4.99 | 0.00 |
| RH8.teenager | 0.00 | 0.00 | 0.15 | 0.88 |
| RH9.parents | -0.01 | 0.00 | -3.69 | 0.00 |
| RH.outstation | 0.01 | 0.00 | 8.46 | 0.00 |
| weekday.Tuesday | -0.10 | 0.01 | -6.71 | 0.00 |
| weekday.Wednesday | -0.05 | 0.01 | -3.51 | 0.00 |
| weekday.Thursday | -0.10 | 0.01 | -6.88 | 0.00 |
| weekday.Friday | -0.01 | 0.01 | -0.35 | 0.72 |
| weekday.Saturday | 0.01 | 0.02 | 0.88 | 0.38 |
| weekday.Sunday | -0.07 | 0.01 | -4.86 | 0.00 |
| poly(hour, 3)1 | -95.61 | 15.63 | -6.12 | 0.00 |
| poly(hour, 3)2 | 25.42 | 14.92 | 1.70 | 0.09 |
| poly(hour, 3)3 | 167.00 | 15.06 | 11.09 | 0.00 |
| T1.kitchen:poly(hour, 3)1 | 2.13 | 2.24 | 0.95 | 0.34 |
| T1.kitchen:poly(hour, 3)2 | -3.85 | 1.91 | -2.02 | 0.04 |
| T1.kitchen:poly(hour, 3)3 | -17.81 | 2.01 | -8.85 | 0.00 |
| T2.living:poly(hour, 3)1 | -5.43 | 2.15 | -2.52 | 0.01 |
| T2.living:poly(hour, 3)2 | 14.70 | 1.72 | 8.53 | 0.00 |
| T2.living:poly(hour, 3)3 | 21.28 | 1.86 | 11.42 | 0.00 |
| T3.laundry:poly(hour, 3)1 | 5.61 | 0.94 | 5.99 | 0.00 |
| T3.laundry:poly(hour, 3)2 | -7.04 | 0.91 | -7.75 | 0.00 |
| T3.laundry:poly(hour, 3)3 | -9.31 | 0.92 | -10.16 | 0.00 |
| T4.office:poly(hour, 3)1 | -0.95 | 0.82 | -1.16 | 0.25 |
| T4.office:poly(hour, 3)2 | -4.22 | 0.78 | -5.42 | 0.00 |
| T4.office:poly(hour, 3)3 | 3.13 | 0.82 | 3.81 | 0.00 |
| T5.bath:poly(hour, 3)1 | 1.06 | 0.95 | 1.11 | 0.27 |
| T5.bath:poly(hour, 3)2 | -0.95 | 0.94 | -1.01 | 0.31 |
| T5.bath:poly(hour, 3)3 | 1.37 | 0.95 | 1.45 | 0.15 |
| T6.outside:poly(hour, 3)1 | -2.03 | 0.59 | -3.48 | 0.00 |
| T6.outside:poly(hour, 3)2 | 2.24 | 0.53 | 4.22 | 0.00 |
| T6.outside:poly(hour, 3)3 | 5.45 | 0.54 | 10.16 | 0.00 |
| T7.ironing:poly(hour, 3)1 | 0.62 | 1.12 | 0.55 | 0.58 |
| T7.ironing:poly(hour, 3)2 | 6.33 | 1.15 | 5.51 | 0.00 |
| T7.ironing:poly(hour, 3)3 | -0.86 | 1.11 | -0.77 | 0.44 |
| T8.teenager:poly(hour, 3)1 | 5.14 | 0.82 | 6.26 | 0.00 |
| T8.teenager:poly(hour, 3)2 | -5.16 | 0.84 | -6.16 | 0.00 |
| T8.teenager:poly(hour, 3)3 | -0.84 | 0.81 | -1.04 | 0.30 |
| T9.parents:poly(hour, 3)1 | -6.97 | 1.55 | -4.49 | 0.00 |
| T9.parents:poly(hour, 3)2 | 1.45 | 1.45 | 1.00 | 0.32 |
| T9.parents:poly(hour, 3)3 | -0.56 | 1.51 | -0.38 | 0.71 |
| T.outstation:poly(hour, 3)1 | 1.94 | 0.68 | 2.87 | 0.00 |
| T.outstation:poly(hour, 3)2 | -2.68 | 0.61 | -4.41 | 0.00 |
| T.outstation:poly(hour, 3)3 | -5.47 | 0.61 | -9.02 | 0.00 |
| poly(hour, 3)1:RH1.kitchen | 3.91 | 0.91 | 4.32 | 0.00 |
| poly(hour, 3)2:RH1.kitchen | -6.18 | 0.79 | -7.79 | 0.00 |
| poly(hour, 3)3:RH1.kitchen | -4.29 | 0.86 | -5.01 | 0.00 |
| poly(hour, 3)1:RH2.living | -4.96 | 0.84 | -5.90 | 0.00 |
| poly(hour, 3)2:RH2.living | 7.37 | 0.75 | 9.79 | 0.00 |
| poly(hour, 3)3:RH2.living | 6.87 | 0.82 | 8.43 | 0.00 |
| poly(hour, 3)1:RH3.laundry | 4.04 | 0.60 | 6.77 | 0.00 |
| poly(hour, 3)2:RH3.laundry | -2.33 | 0.58 | -4.04 | 0.00 |
| poly(hour, 3)3:RH3.laundry | -2.99 | 0.58 | -5.19 | 0.00 |
| poly(hour, 3)1:RH4.office | -0.11 | 0.50 | -0.21 | 0.83 |
| poly(hour, 3)2:RH4.office | 1.64 | 0.51 | 3.25 | 0.00 |
| poly(hour, 3)3:RH4.office | 0.73 | 0.51 | 1.43 | 0.15 |
| poly(hour, 3)1:RH5.bath | 0.01 | 0.08 | 0.17 | 0.87 |
| poly(hour, 3)2:RH5.bath | -0.22 | 0.06 | -3.39 | 0.00 |
| poly(hour, 3)3:RH5.bath | -0.05 | 0.08 | -0.63 | 0.53 |
| poly(hour, 3)1:RH6.outside | -0.29 | 0.06 | -4.69 | 0.00 |
| poly(hour, 3)2:RH6.outside | 0.16 | 0.06 | 2.67 | 0.01 |
| poly(hour, 3)3:RH6.outside | -0.04 | 0.06 | -0.70 | 0.48 |
| poly(hour, 3)1:RH7.ironing | -2.03 | 0.37 | -5.54 | 0.00 |
| poly(hour, 3)2:RH7.ironing | 0.27 | 0.39 | 0.71 | 0.48 |
| poly(hour, 3)3:RH7.ironing | 0.92 | 0.37 | 2.48 | 0.01 |
| poly(hour, 3)1:RH8.teenager | 1.84 | 0.35 | 5.30 | 0.00 |
| poly(hour, 3)2:RH8.teenager | 0.12 | 0.34 | 0.33 | 0.74 |
| poly(hour, 3)3:RH8.teenager | -3.84 | 0.33 | -11.50 | 0.00 |
| poly(hour, 3)1:RH9.parents | -1.68 | 0.38 | -4.39 | 0.00 |
| poly(hour, 3)2:RH9.parents | -0.89 | 0.36 | -2.48 | 0.01 |
| poly(hour, 3)3:RH9.parents | 1.92 | 0.36 | 5.34 | 0.00 |
| poly(hour, 3)1:RH.outstation | 0.62 | 0.11 | 5.88 | 0.00 |
| poly(hour, 3)2:RH.outstation | -0.52 | 0.10 | -5.29 | 0.00 |
| poly(hour, 3)3:RH.outstation | -0.42 | 0.10 | -4.23 | 0.00 |
| poly(hour, 3)1:weekday.Tuesday | -6.01 | 2.09 | -2.87 | 0.00 |
| poly(hour, 3)2:weekday.Tuesday | 14.83 | 2.08 | 7.13 | 0.00 |
| poly(hour, 3)3:weekday.Tuesday | 6.33 | 2.04 | 3.10 | 0.00 |
| poly(hour, 3)1:weekday.Wednesday | -7.47 | 2.08 | -3.59 | 0.00 |
| poly(hour, 3)2:weekday.Wednesday | 4.99 | 2.04 | 2.44 | 0.01 |
| poly(hour, 3)3:weekday.Wednesday | 10.79 | 2.02 | 5.34 | 0.00 |
| poly(hour, 3)1:weekday.Thursday | -3.29 | 2.00 | -1.65 | 0.10 |
| poly(hour, 3)2:weekday.Thursday | 4.25 | 2.00 | 2.12 | 0.03 |
| poly(hour, 3)3:weekday.Thursday | -0.31 | 1.99 | -0.16 | 0.87 |
| poly(hour, 3)1:weekday.Friday | -3.60 | 2.10 | -1.71 | 0.09 |
| poly(hour, 3)2:weekday.Friday | -7.69 | 2.09 | -3.69 | 0.00 |
| poly(hour, 3)3:weekday.Friday | 5.07 | 2.05 | 2.48 | 0.01 |
| poly(hour, 3)1:weekday.Saturday | 0.23 | 2.15 | 0.11 | 0.92 |
| poly(hour, 3)2:weekday.Saturday | -9.67 | 2.12 | -4.56 | 0.00 |
| poly(hour, 3)3:weekday.Saturday | 0.56 | 2.08 | 0.27 | 0.79 |
| poly(hour, 3)1:weekday.Sunday | -5.23 | 2.12 | -2.47 | 0.01 |
| poly(hour, 3)2:weekday.Sunday | 6.49 | 2.10 | 3.09 | 0.00 |
| poly(hour, 3)3:weekday.Sunday | 12.26 | 2.05 | 5.96 | 0.00 |

Table 7.23: Summary of full interaction model (LMinter): `lapp` $\sim$ (`temperatures` + `humidities` + `weekday`) $*$ `poly(hour, 3)`. With $R^2_{adj} = 0.393$, $F$-statistic $= 120.4$ on 107 and 19627 DF, $p$-value $< 2e-16$.

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | 2.3238 | 0.1112 | 20.90 | 0.0000 |
| T1.kitchen | -0.0244 | 0.0147 | -1.66 | 0.0972 |
| T2.living | -0.0592 | 0.0133 | -4.45 | 0.0000 |
| T3.laundry | 0.1032 | 0.0068 | 15.20 | 0.0000 |
| T4.office | 0.0513 | 0.0060 | 8.58 | 0.0000 |
| T6.outside | -0.0090 | 0.0042 | -2.15 | 0.0313 |
| T7.ironing | -0.0203 | 0.0088 | -2.31 | 0.0211 |
| T8.teenager | 0.1286 | 0.0065 | 19.70 | 0.0000 |
| T9.parents | -0.1302 | 0.0115 | -11.34 | 0.0000 |
| T.outstation | 0.0139 | 0.0048 | 2.89 | 0.0039 |
| RH1.kitchen | 0.0353 | 0.0060 | 5.86 | 0.0000 |
| RH2.living | -0.0519 | 0.0055 | -9.44 | 0.0000 |
| RH3.laundry | 0.0440 | 0.0042 | 10.44 | 0.0000 |
| RH4.office | -0.0022 | 0.0039 | -0.56 | 0.5763 |
| RH5.bath | 0.0038 | 0.0006 | 6.78 | 0.0000 |
| RH6.outside | -0.0028 | 0.0004 | -6.36 | 0.0000 |
| RH7.ironing | -0.0147 | 0.0029 | -5.04 | 0.0000 |
| RH8.teenager | 0.0006 | 0.0027 | 0.22 | 0.8239 |
| RH9.parents | -0.0103 | 0.0027 | -3.75 | 0.0002 |
| RH.outstation | 0.0061 | 0.0007 | 8.37 | 0.0000 |
| weekday.Tuesday | -0.0995 | 0.0147 | -6.75 | 0.0000 |
| weekday.Wednesday | -0.0521 | 0.0147 | -3.55 | 0.0004 |
| weekday.Thursday | -0.0992 | 0.0143 | -6.93 | 0.0000 |
| weekday.Friday | -0.0064 | 0.0147 | -0.43 | 0.6640 |
| weekday.Saturday | 0.0136 | 0.0151 | 0.90 | 0.3663 |
| weekday.Sunday | -0.0728 | 0.0148 | -4.91 | 0.0000 |
| poly(hour, 3)1 | -97.4838 | 15.5580 | -6.27 | 0.0000 |
| poly(hour, 3)2 | 27.3026 | 14.8049 | 1.84 | 0.0652 |
| poly(hour, 3)3 | 164.2781 | 14.9439 | 10.99 | 0.0000 |
| T1:poly(hour, 3)1 | 2.2770 | 2.2384 | 1.02 | 0.3091 |
| T1:poly(hour, 3)2 | -4.0939 | 1.8957 | -2.16 | 0.0308 |
| T1:poly(hour, 3)3 | -17.6548 | 2.0087 | -8.79 | 0.0000 |
| T2:poly(hour, 3)1 | -5.1861 | 2.1447 | -2.42 | 0.0156 |
| T2:poly(hour, 3)2 | 14.6738 | 1.7174 | 8.54 | 0.0000 |
| T2:poly(hour, 3)3 | 21.5339 | 1.8555 | 11.61 | 0.0000 |
| T3:poly(hour, 3)1 | 5.7256 | 0.9253 | 6.19 | 0.0000 |
| T3:poly(hour, 3)2 | -7.2081 | 0.9001 | -8.01 | 0.0000 |
| T3:poly(hour, 3)3 | -9.0226 | 0.9055 | -9.96 | 0.0000 |
| T4:poly(hour, 3)1 | -0.9050 | 0.8119 | -1.11 | 0.2650 |
| T4:poly(hour, 3)2 | -4.2525 | 0.7767 | -5.48 | 0.0000 |
| T4:poly(hour, 3)3 | 3.2782 | 0.8156 | 4.02 | 0.0001 |
| T6:poly(hour, 3)1 | -2.1642 | 0.5752 | -3.76 | 0.0002 |
| T6:poly(hour, 3)2 | 2.3231 | 0.5227 | 4.44 | 0.0000 |
| T6:poly(hour, 3)3 | 5.2832 | 0.5294 | 9.98 | 0.0000 |
| T7:poly(hour, 3)1 | 0.8162 | 1.1138 | 0.73 | 0.4637 |
| T7:poly(hour, 3)2 | 6.2716 | 1.1424 | 5.49 | 0.0000 |
| T7:poly(hour, 3)3 | -0.5772 | 1.1016 | -0.52 | 0.6003 |
| T8:poly(hour, 3)1 | 5.1076 | 0.8198 | 6.23 | 0.0000 |
| T8:poly(hour, 3)2 | -5.1771 | 0.8369 | -6.19 | 0.0000 |
| T8:poly(hour, 3)3 | -0.8092 | 0.8124 | -1.00 | 0.3193 |
| T9:poly(hour, 3)1 | -6.6029 | 1.5117 | -4.37 | 0.0000 |
| T9:poly(hour, 3)2 | 1.0702 | 1.4106 | 0.76 | 0.4480 |
| T9:poly(hour, 3)3 | -0.3745 | 1.4578 | -0.26 | 0.7973 |
| T.out:poly(hour, 3)1 | 2.0866 | 0.6699 | 3.11 | 0.0018 |
| T.out:poly(hour, 3)2 | -2.7112 | 0.6017 | -4.51 | 0.0000 |
| T.out:poly(hour, 3)3 | -5.3262 | 0.6021 | -8.85 | 0.0000 |
| poly(hour, 3)1:RH1 | 3.8655 | 0.9042 | 4.28 | 0.0000 |
| poly(hour, 3)2:RH1 | -6.1819 | 0.7922 | -7.80 | 0.0000 |
| poly(hour, 3)3:RH1 | -4.2975 | 0.8531 | -5.04 | 0.0000 |
| poly(hour, 3)1:RH2 | -4.9170 | 0.8403 | -5.85 | 0.0000 |
| poly(hour, 3)2:RH2 | 7.3368 | 0.7509 | 9.77 | 0.0000 |
| poly(hour, 3)3:RH2 | 6.9283 | 0.8128 | 8.52 | 0.0000 |
| poly(hour, 3)1:RH3 | 4.0604 | 0.5951 | 6.82 | 0.0000 |
| poly(hour, 3)2:RH3 | -2.2837 | 0.5764 | -3.96 | 0.0001 |
| poly(hour, 3)3:RH3 | -3.0108 | 0.5763 | -5.22 | 0.0000 |
| poly(hour, 3)1:RH4 | -0.0623 | 0.4993 | -0.12 | 0.9006 |
| poly(hour, 3)2:RH4 | 1.6098 | 0.5025 | 3.20 | 0.0014 |
| poly(hour, 3)3:RH4 | 0.8155 | 0.5115 | 1.59 | 0.1109 |
| poly(hour, 3)1:RH5 | 0.0354 | 0.0747 | 0.47 | 0.6358 |
| poly(hour, 3)2:RH5 | -0.2233 | 0.0612 | -3.65 | 0.0003 |
| poly(hour, 3)3:RH5 | -0.0232 | 0.0762 | -0.30 | 0.7604 |
| poly(hour, 3)1:RH6 | -0.2907 | 0.0628 | -4.63 | 0.0000 |
| poly(hour, 3)2:RH6 | 0.1601 | 0.0583 | 2.75 | 0.0061 |
| poly(hour, 3)3:RH6 | -0.0433 | 0.0590 | -0.73 | 0.4628 |
| poly(hour, 3)1:RH7 | -2.0458 | 0.3663 | -5.58 | 0.0000 |
| poly(hour, 3)2:RH7 | 0.2614 | 0.3850 | 0.68 | 0.4971 |
| poly(hour, 3)3:RH7 | 0.9060 | 0.3717 | 2.44 | 0.0148 |
| poly(hour, 3)1:RH8 | 1.8700 | 0.3468 | 5.39 | 0.0000 |
| poly(hour, 3)2:RH8 | 0.0997 | 0.3421 | 0.29 | 0.7706 |
| poly(hour, 3)3:RH8 | -3.8633 | 0.3322 | -11.63 | 0.0000 |
| poly(hour, 3)1:RH9 | -1.7793 | 0.3786 | -4.70 | 0.0000 |
| poly(hour, 3)2:RH9 | -0.8837 | 0.3520 | -2.51 | 0.0121 |
| poly(hour, 3)3:RH9 | 1.8622 | 0.3553 | 5.24 | 0.0000 |
| poly(hour, 3)1:RH.out | 0.6244 | 0.1057 | 5.91 | 0.0000 |
| poly(hour, 3)2:RH.out | -0.5231 | 0.0987 | -5.30 | 0.0000 |
| poly(hour, 3)3:RH.out | -0.4158 | 0.0993 | -4.19 | 0.0000 |
| poly(hour, 3)1:weekday.Tuesday | -5.9389 | 2.0930 | -2.84 | 0.0046 |
| poly(hour, 3)2:weekday.Tuesday | 14.9919 | 2.0783 | 7.21 | 0.0000 |
| poly(hour, 3)3:weekday.Tuesday | 6.2782 | 2.0392 | 3.08 | 0.0021 |
| poly(hour, 3)1:weekday.Wednesday | -7.5385 | 2.0804 | -3.62 | 0.0003 |
| poly(hour, 3)2:weekday.Wednesday | 5.1233 | 2.0409 | 2.51 | 0.0121 |
| poly(hour, 3)3:weekday.Wednesday | 10.7376 | 2.0184 | 5.32 | 0.0000 |
| poly(hour, 3)1:weekday.Thursday | -3.3827 | 1.9994 | -1.69 | 0.0907 |
| poly(hour, 3)2:weekday.Thursday | 4.4160 | 1.9966 | 2.21 | 0.0270 |
| poly(hour, 3)3:weekday.Thursday | -0.4498 | 1.9822 | -0.23 | 0.8205 |
| poly(hour, 3)1:weekday.Friday | -3.7746 | 2.0903 | -1.81 | 0.0710 |
| poly(hour, 3)2:weekday.Friday | -7.3491 | 2.0755 | -3.54 | 0.0004 |
| poly(hour, 3)3:weekday.Friday | 4.8581 | 2.0417 | 2.38 | 0.0174 |
| poly(hour, 3)1:weekday.Saturday | 0.1261 | 2.1434 | 0.06 | 0.9531 |
| poly(hour, 3)2:weekday.Saturday | -9.5805 | 2.1222 | -4.51 | 0.0000 |
| poly(hour, 3)3:weekday.Saturday | 0.4866 | 2.0801 | 0.23 | 0.8150 |
| poly(hour, 3)1:weekday.Sunday | -5.1403 | 2.1088 | -2.44 | 0.0148 |
| poly(hour, 3)2:weekday.Sunday | 6.9113 | 2.0902 | 3.31 | 0.0009 |
| poly(hour, 3)3:weekday.Sunday | 12.2086 | 2.0515 | 5.95 | 0.0000 |

Table 7.24: Summary of reduced interaction Model (RedInterModel), where we just removed `T5.bath`. With $R^2_{adj} = 0.393$, $F$-statistic $= 125$ on 103 and 19631 DF, and $p$-value $< 2e - 16$.

## 7.4 Comparing main effect model and interaction model

### 7.4.1 Statistics

| | **Main Effect Model** | **Interaction Model** |
|---|---|---|
| Full | $R^2_{adj} = 0.328$<br>$F$-statistic = 333 on 29 and 19705 DF | $R^2_{adj} = 0.393$<br>$F$-statistic = 120 on 107 and 19627 DF |
| Reduced 1 | $R^2_{adj} = 0.328$<br>$F$-statistic = 358 on 27 and 19707 DF | $R^2_{adj} = 0.393$<br>$F$-statistic = 125 on 103 and 19631 DF |
| Reduced 2 | $R^2_{adj} = 0.328$<br>$F$-statistic = 371 on 26 and 19708 DF | - |

Table 7.25: Comparing $R^2$ and $F$-statistic of all the set main and interaction models.

| | df | AIC |
|---|---|---|
| **Main effect model** | | |
| Full (Model 7) | 31 | 31572 |
| Reduced 1 | 29 | 31571 |
| Reduced 2 | 28 | 31569 |
| **Interaction effect model** | | |
| Full (LMinter) | 109 | 29641 |
| Reduced 1 (RedInterModel) | 105 | 29638 |

Table 7.26: Comparing all the created main and interaction models with model selection criterion AIC.

Finally, we can conclude from the above results in Table 7.25 and 7.26 that there are hardly any differences between the full and reduced models. This means, the reduction does not achieve much improvement regarding adjusted coefficient of determination and AIC. Thus we can work with the full models, so for main model prediction we use the full main model (Model 7) and for the interaction model prediction we operate with the full interaction model (LMinter).

## 7.5 Predictions on the energy consumption

After finding out that the full main model (Model 7) and full interaction model (LMinter) are relatively more convincing in terms of their model selection values, we perform and study predictions with these models. For this, we want to explore the relationship between variables `lapp` and `hour` and set some predictions by working with Model 7 first and then LMinter.

### Preparation

While we make some prediction of `lapp` depending on one variable covariate and hours, we set all the other covariates to their medians with respect to the hours.

Since we fix the covariates at their median, we first observe the median of all the temperature depending on the hours (c.f. Figure 7.11). To visualize the medians, we leave out the two outside temperatures, `T6.outside` and `T.outstation`, as they vary similarly between 5 and 10 °C, whereas the other inside temperatures range around 20 and 21 °C. A detailed observed median plot of the inside temperatures is prepared in Figure 7.11.

All the observed medians of the temperature covariates depending on the hours are listed in Table 7.28.

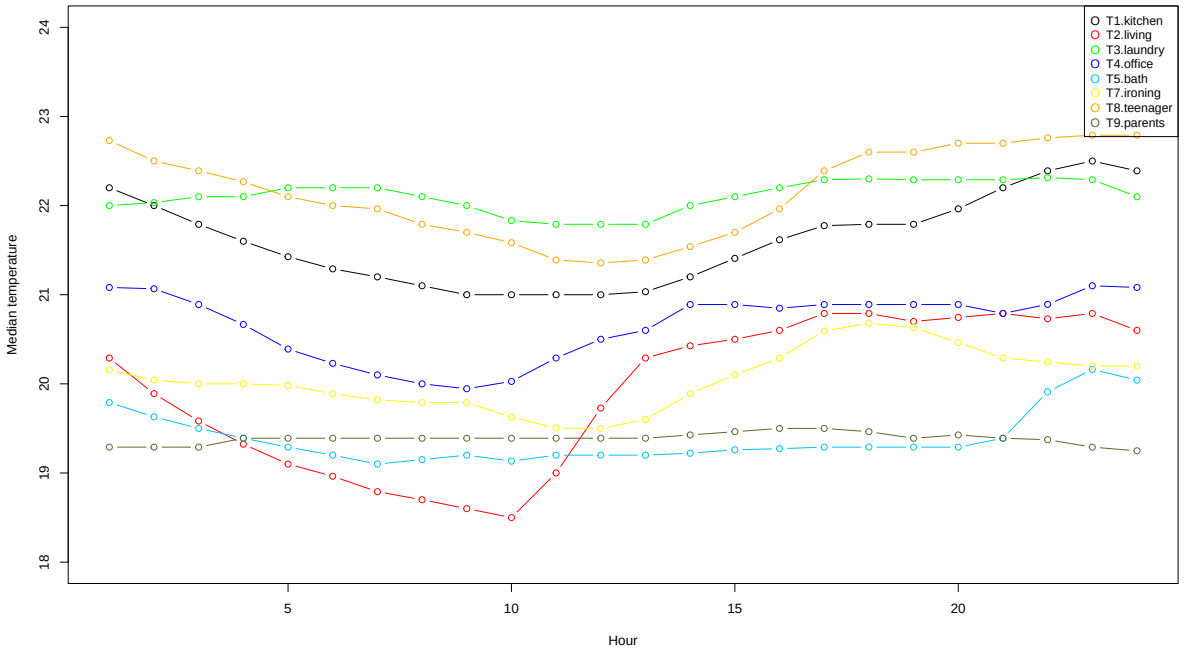



Figure 7.11: Observed median of all temperatures dependent of corresponding `hour`.

Next, we are doing the same for the humidity covariates. The visualization of the observed medians depending on the hours is given in Figure 7.12. The two outside humidities `T6.outside` and `T.outstation` have the almost same curve, while `T6.outside` ranges more around the medians of inside humidities. The observed medians of inside humidities varies approximately around 40 %.

All the observed medians of the humidity covariates depending on the hours are listed in Table 7.29.

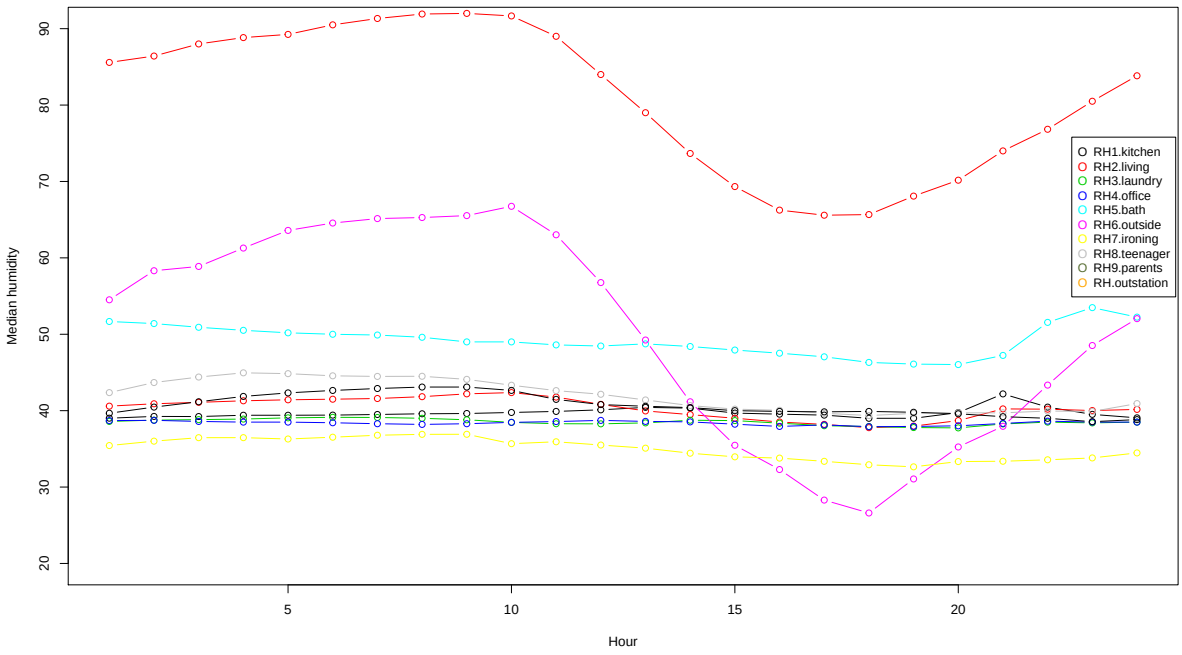



Figure 7.12: Observed median of all humidities dependent of corresponding `hour`.

After fixing the medians of the temperature and humidity covariates, we look at the three fixed polynomial coefficients of the corresponding hours that will used for the predictions, c.f. Table 7.27.

| hour | $poly(\texttt{hour}_i, 1)$ | $poly(\texttt{hour}_i, 2)$ | $poly(\texttt{hour}_i, 3)$ |
|---|---|---|---|
| 1 | -0.012 | 0.014 | -0.015 |
| 2 | -0.011 | 0.010 | -0.007 |
| 3 | -0.0098 | 0.0071 | -0.0011 |
| 4 | -0.0087 | 0.0041 | 0.0032 |
| 5 | -0.0077 | 0.0014 | 0.0062 |
| 6 | -0.0067 | -0.00094 | 0.0079 |
| 7 | -0.0057 | -0.0029 | 0.0085 |
| 8 | -0.0046 | -0.0046 | 0.0082 |
| 9 | -0.0036 | -0.0059 | 0.0071 |
| 10 | -0.0026 | -0.0069 | 0.0055 |
| 11 | -0.0016 | -0.0076 | 0.0035 |
| 12 | -0.00052 | -0.0079 | 0.0012 |
| 13 | 0.00051 | -0.0079 | -0.0012 |
| 14 | 0.0015 | -0.0076 | -0.0035 |
| 15 | 0.0026 | -0.0069 | -0.0055 |
| 16 | 0.0036 | -0.0059 | -0.0071 |
| 17 | 0.0046 | -0.0046 | -0.0082 |
| 18 | 0.0056 | -0.0029 | -0.0085 |
| 19 | 0.0067 | -0.00095 | -0.0078 |
| 20 | 0.0077 | 0.0014 | -0.0062 |
| 21 | 0.0087 | 0.0041 | -0.0032 |
| 22 | 0.0098 | 0.007 | 0.0011 |
| 23 | 0.011 | 0.01 | 0.007 |
| 24 | 0.012 | 0.014 | 0.015 |

Table 7.27: Polynomial Coefficient of covariate `hour`.

| Hrs | T1. kitchen | T2. living | T3. laundry | T4. office | T5. bath | T6. outside | T7. ironing | T8. teenager | T9. parents | T.out-station |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 22.20 | 20.29 | 22.00 | 21.08 | 19.79 | 5.95 | 20.16 | 22.73 | 19.29 | 6.22 |
| 2 | 22.00 | 19.89 | 22.03 | 21.07 | 19.63 | 5.80 | 20.04 | 22.50 | 19.29 | 6.16 |
| 3 | 21.79 | 19.58 | 22.10 | 20.89 | 19.50 | 5.67 | 20.00 | 22.39 | 19.29 | 5.55 |
| 4 | 21.60 | 19.32 | 22.10 | 20.67 | 19.39 | 5.40 | 20.00 | 22.27 | 19.39 | 5.33 |
| 5 | 21.43 | 19.10 | 22.20 | 20.39 | 19.29 | 5.08 | 19.98 | 22.10 | 19.39 | 5.10 |
| 6 | 21.29 | 18.96 | 22.20 | 20.23 | 19.20 | 4.56 | 19.89 | 22.00 | 19.39 | 4.97 |
| 7 | 21.20 | 18.79 | 22.20 | 20.10 | 19.10 | 4.52 | 19.82 | 21.96 | 19.39 | 4.84 |
| 8 | 21.10 | 18.70 | 22.10 | 20.00 | 19.15 | 4.40 | 19.79 | 21.79 | 19.39 | 4.80 |
| 9 | 21.00 | 18.60 | 22.00 | 19.95 | 19.20 | 4.73 | 19.79 | 21.70 | 19.39 | 4.82 |
| 10 | 21.00 | 18.50 | 21.83 | 20.03 | 19.13 | 5.59 | 19.63 | 21.58 | 19.39 | 5.02 |
| 11 | 21.00 | 19.00 | 21.79 | 20.29 | 19.20 | 6.51 | 19.50 | 21.39 | 19.39 | 5.70 |
| 12 | 21.00 | 19.73 | 21.79 | 20.50 | 19.20 | 8.11 | 19.50 | 21.36 | 19.39 | 6.50 |
| 13 | 21.03 | 20.29 | 21.79 | 20.60 | 19.20 | 9.20 | 19.60 | 21.39 | 19.39 | 7.60 |
| 14 | 21.20 | 20.43 | 22.00 | 20.89 | 19.22 | 9.80 | 19.89 | 21.54 | 19.43 | 8.46 |
| 15 | 21.41 | 20.50 | 22.10 | 20.89 | 19.26 | 10.29 | 20.10 | 21.70 | 19.46 | 9.10 |
| 16 | 21.62 | 20.60 | 22.20 | 20.85 | 19.27 | 10.67 | 20.29 | 21.96 | 19.50 | 9.50 |
| 17 | 21.78 | 20.79 | 22.29 | 20.89 | 19.29 | 10.60 | 20.59 | 22.39 | 19.50 | 9.80 |
| 18 | 21.79 | 20.79 | 22.30 | 20.89 | 19.29 | 10.50 | 20.68 | 22.60 | 19.46 | 9.73 |
| 19 | 21.79 | 20.70 | 22.29 | 20.89 | 19.29 | 9.96 | 20.63 | 22.60 | 19.39 | 9.35 |
| 20 | 21.96 | 20.75 | 22.29 | 20.89 | 19.29 | 9.46 | 20.46 | 22.70 | 19.43 | 8.78 |
| 21 | 22.20 | 20.79 | 22.29 | 20.79 | 19.39 | 8.73 | 20.29 | 22.70 | 19.39 | 8.23 |
| 22 | 22.39 | 20.73 | 22.31 | 20.89 | 19.91 | 8.03 | 20.25 | 22.76 | 19.37 | 7.88 |
| 23 | 22.50 | 20.79 | 22.29 | 21.10 | 20.16 | 7.38 | 20.20 | 22.79 | 19.29 | 7.32 |
| 24 | 22.39 | 20.60 | 22.10 | 21.08 | 20.04 | 6.33 | 20.20 | 22.79 | 19.25 | 6.81 |

Table 7.28: Observed median of all temperature variables in °C.

| Hrs | RH1. kitchen | RH2. living | RH3. laundry | RH4. office | RH5. bath | RH6. outside | RH7. ironing | RH8. teenager | RH9. parents | RH.out-station |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 39.02 | 40.59 | 38.59 | 38.70 | 51.67 | 54.51 | 35.43 | 42.36 | 39.66 | 85.58 |
| 2 | 39.25 | 40.90 | 38.76 | 38.73 | 51.40 | 58.32 | 36.01 | 43.68 | 40.47 | 86.42 |
| 3 | 39.23 | 41.09 | 38.83 | 38.59 | 50.92 | 58.88 | 36.46 | 44.40 | 41.16 | 88.00 |
| 4 | 39.40 | 41.29 | 38.90 | 38.50 | 50.52 | 61.28 | 36.47 | 44.95 | 41.86 | 88.83 |
| 5 | 39.40 | 41.42 | 39.06 | 38.50 | 50.19 | 63.59 | 36.29 | 44.85 | 42.33 | 89.25 |
| 6 | 39.42 | 41.50 | 39.13 | 38.42 | 50.00 | 64.56 | 36.53 | 44.56 | 42.65 | 90.50 |
| 7 | 39.50 | 41.59 | 39.11 | 38.29 | 49.90 | 65.14 | 36.78 | 44.48 | 42.90 | 91.33 |
| 8 | 39.59 | 41.83 | 39.00 | 38.20 | 49.611 | 65.28 | 36.90 | 44.50 | 43.09 | 91.92 |
| 9 | 39.62 | 42.20 | 38.81 | 38.29 | 49.00 | 65.53 | 36.91 | 44.10 | 43.09 | 92.00 |
| 10 | 39.76 | 42.36 | 38.47 | 38.47 | 49.00 | 66.75 | 35.68 | 43.33 | 42.65 | 91.67 |
| 11 | 39.90 | 41.78 | 38.29 | 38.58 | 48.61 | 63.02 | 35.923 | 42.62 | 41.48 | 89.00 |
| 12 | 40.11 | 40.82 | 38.28 | 38.73 | 48.46 | 56.77 | 35.50 | 42.15 | 40.80 | 84.00 |
| 13 | 40.43 | 39.96 | 38.40 | 38.59 | 48.74 | 49.24 | 35.09 | 41.40 | 40.58 | 79.00 |
| 14 | 40.33 | 39.48 | 38.75 | 38.53 | 48.40 | 41.16 | 34.43 | 40.67 | 40.40 | 73.67 |
| 15 | 39.68 | 39.01 | 38.70 | 38.23 | 47.93 | 35.47 | 33.96 | 40.20 | 40.00 | 69.33 |
| 16 | 39.54 | 38.52 | 38.36 | 37.94 | 47.52 | 32.30 | 33.79 | 40.04 | 39.90 | 66.25 |
| 17 | 39.42 | 38.20 | 38.03 | 38.13 | 47.05 | 28.30 | 33.37 | 39.66 | 39.85 | 65.58 |
| 18 | 39.00 | 37.78 | 37.93 | 37.90 | 46.31 | 26.61 | 32.92 | 39.36 | 39.90 | 65.67 |
| 19 | 39.00 | 37.98 | 37.79 | 37.93 | 46.10 | 31.06 | 32.65 | 39.64 | 39.79 | 68.08 |
| 20 | 39.75 | 38.70 | 37.74 | 38.02 | 46.03 | 35.25 | 33.34 | 39.72 | 39.59 | 70.17 |
| 21 | 42.19 | 40.23 | 38.26 | 38.33 | 47.23 | 37.95 | 33.38 | 39.71 | 39.21 | 74.00 |
| 22 | 40.46 | 40.19 | 38.50 | 38.63 | 51.55 | 43.35 | 33.57 | 39.95 | 39.00 | 76.83 |
| 23 | 39.51 | 40.01 | 38.40 | 38.50 | 53.49 | 48.53 | 33.81 | 39.90 | 38.56 | 80.50 |
| 24 | 39.06 | 40.16 | 38.50 | 38.50 | 52.25 | 52.05 | 34.47 | 40.93 | 38.83 | 83.83 |

Table 7.29: Observed median of all humidity variables in %.

These values contained in Table 7.28, 7.29 and 7.27 will be used for our predictions in the next sections.

### 7.5.1 Prediction for main effects

Now to examine the hourly pattern, let's compare the appliances energy consumption predictions of some highly relevant area temperatures visualized by 24 slopes, that is a slope for each hours. Do all the rooms have the same intra-day pattern? Does the pattern differ from day to day? To see this we make some prediction (c.f. Figure 7.13 and 7.14).

For these predictions, we use the full main model (Model 7). The prediction input data is set as follows, so that we yield the formula, for $i = 1, \dots 19735$,

$$\begin{aligned}\hat{Y}_i = \hat{\beta}_0 \quad & +\hat{\beta}_1 x_{i1} + \hat{\boldsymbol{\beta}}\tilde{\mathbf{x}}_i + \hat{\beta}_{21}\texttt{poly}(h_i,\texttt{1}) + \hat{\beta}_{22}\texttt{poly}(h_i,\texttt{2}) + \hat{\beta}_{23}\texttt{poly}(h_i,\texttt{3}) \\ & +\hat{\beta}_{24}\delta_i \texttt{weekday}_i,\end{aligned} \tag{7.2}$$

where $\delta_i$ selects the weekday of interest. The covariate of interest $x_{i1}$ will be able to take different values in a reasonable range. In the case of temperatures, we set $x_{i1} \in [min_j(\texttt{Tj}), max_j(\texttt{Tj})]$ with $j =$ `1.kitchen`, `2.living`, `3.laundry`, `4.office`, `5.bath`, `7.ironing`, `8.teenager`, `9.parents` for indoor temperatures and $j =$ `6.outside`, `.outstation` for the outdoor temperatures. Furthermore, $\tilde{\mathbf{x}}$ is the median design matrix of the remaining 19 temperature and humidity covariates that will be set to fixed values that is the observed median for the corresponding hour $h_i$ with $h_i = 1, \dots, 24$ and $i = 1, \dots 19735$, for which we created our Table 7.28 and 7.29. And the final inputs are the corresponding polynomial coefficient of the hour of interest and the weekday we want to predict.

To set more details in our prediction and to capture some pattern and differences between weekdays, we calculate the daily prediction. Can we notice different pattern for the specialized prediction?

The following plots, in Figure 7.13 and 7.14, represents the prediction of the full main model (Model 7) with confidence level 0.95.

Relationship between fitted $(\widehat{\texttt{lapp}}_{i_h})_{i_h=(i\in[19735]|\texttt{hour}=h,\text{with}h=1,\dots,24)}$ and six different areas to see the variation of hourly wise pattern.

For the kitchen temperature we visualize all weekdays to see the differences between the weekdays (c.f. Figure 7.13). As for the other room temperatures, that is living, laundry, office, bathroom and outside, we only give the Monday predictions. Their predictions for the whole week are in the Appendix given in Figure A.8, A.9, A.10, A.11 and A.12.

In advance, we can give the overall impression that all the hours have almost the same slope for one area. They do not have significantly different gradients. But as for the different room, we have different slopes. On the one hand there are good positive effect for the kitchen and laundry room, a low slope for the office and bathroom. It validates our findings about the low significant covariate `T5.bath`. On the other hand as for the living room, there is a highly negative relation with `lapp` throughout the day. And for our outside temperature we have small descending line for all the hours.

The reason for only giving the Monday plots for the five other room temperatures (c.f. Figure 7.14) is that we do not have an interaction between the temperatures and weekdays, as implemented in our full main model (Model 7). Thus we have the same

pattern, i.e. the parallel lines which represents the hours $h = 1, \ldots, 24$, for each weekday. The only difference between the weekdays is that parallel lines are translated or there are small parallel shifts per weekday.

For instance, inspection of Figure 7.13 indicates a little bit more diversified parallel lines for the hourly-wise Monday prediction than for Sunday prediction.

**Interpretation:** For the main effect, we only observe hourly-wise lines which behave the same as we do not allow any interaction. Only the distances between the hourly-wise slopes and their angle are to be interpreted, whereby these are very weak. Since we already discussed these issues above, we focus on the hours. In the afternoon and evening hours we have the highest appliances energy use, as expected. So in the midnight to the morning hours the occupants behave the same and using less energy, whereas in the afternoon and especially in the evening hours the energy consumption reaches its peaks.

Furthermore, we provide the prediction of full main model (Model 7) for the relevant and corresponding humidities in Figure 7.15. We created the visualization exactly analogously to the temperature figures, i.e. allow the variable of interest to take reasonable values and fix the other to their median. The humidity predictions support our findings in the temperature predictions. The presence of the occupants increases the temperature and humidity and thus probably leads to device application. Note that due to thermodynamic reasons that rising temperature also causes a higher humidity absorption and therefore a decreasing humidity percentage (c.f. Figure A.1). The figure also shows the same afternoon behavior like we detected in the temperature predictions.

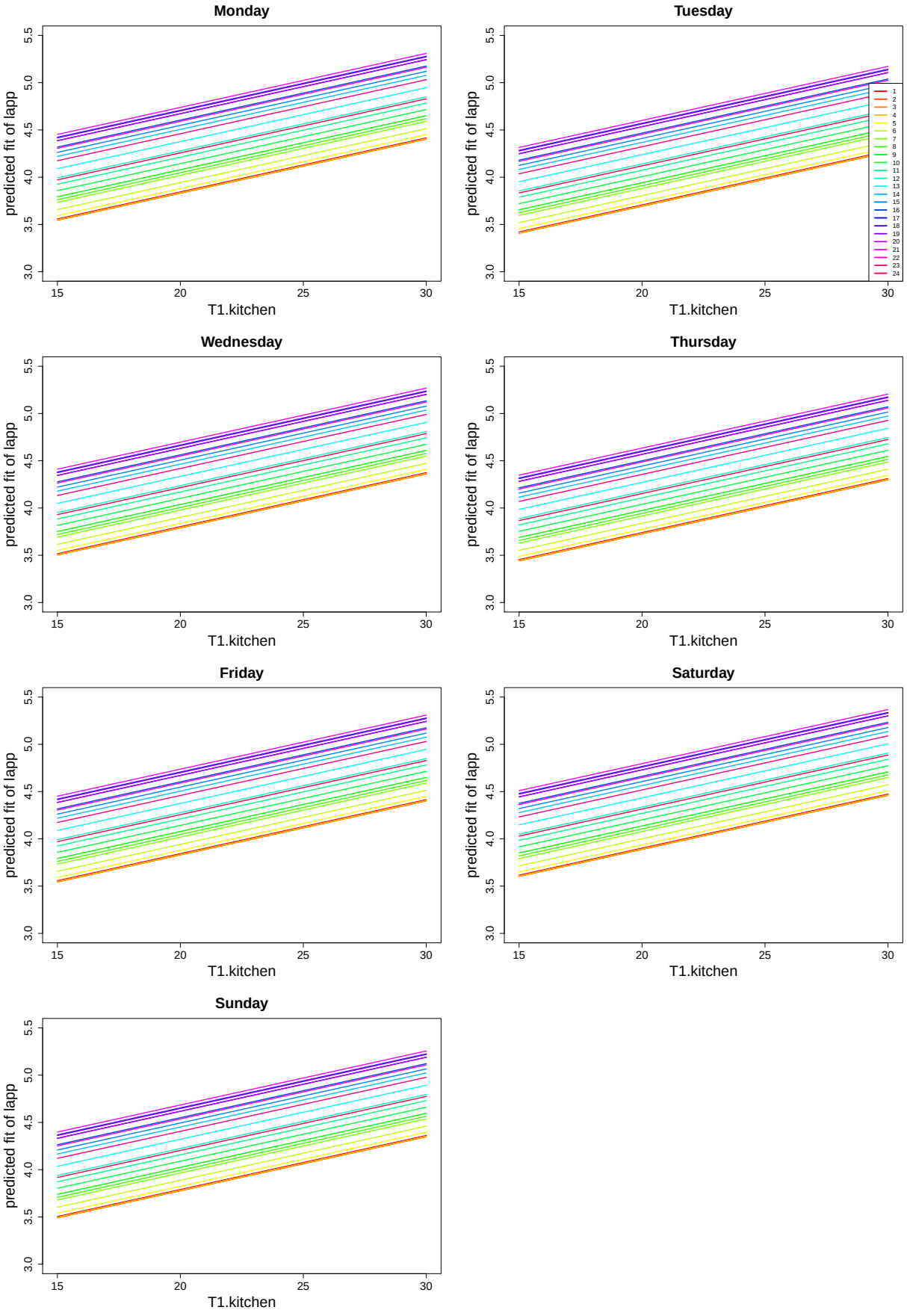



Figure 7.13: Prediction of the full main model (Model 7) restricted on weekdays. The plot shows the relationship between fitted $(\widehat{\texttt{lapp}}_{i_h})_{i_h=(i\in[19735]|\texttt{hour}=h,\text{with}h=1,\ldots,24)}$ and temperature `T1.kitchen` to see the variation of hourly-wise pattern for all the weekdays, while the other covariates fixed at their medians. Condition $h$ is colored by hours 1 to 24.

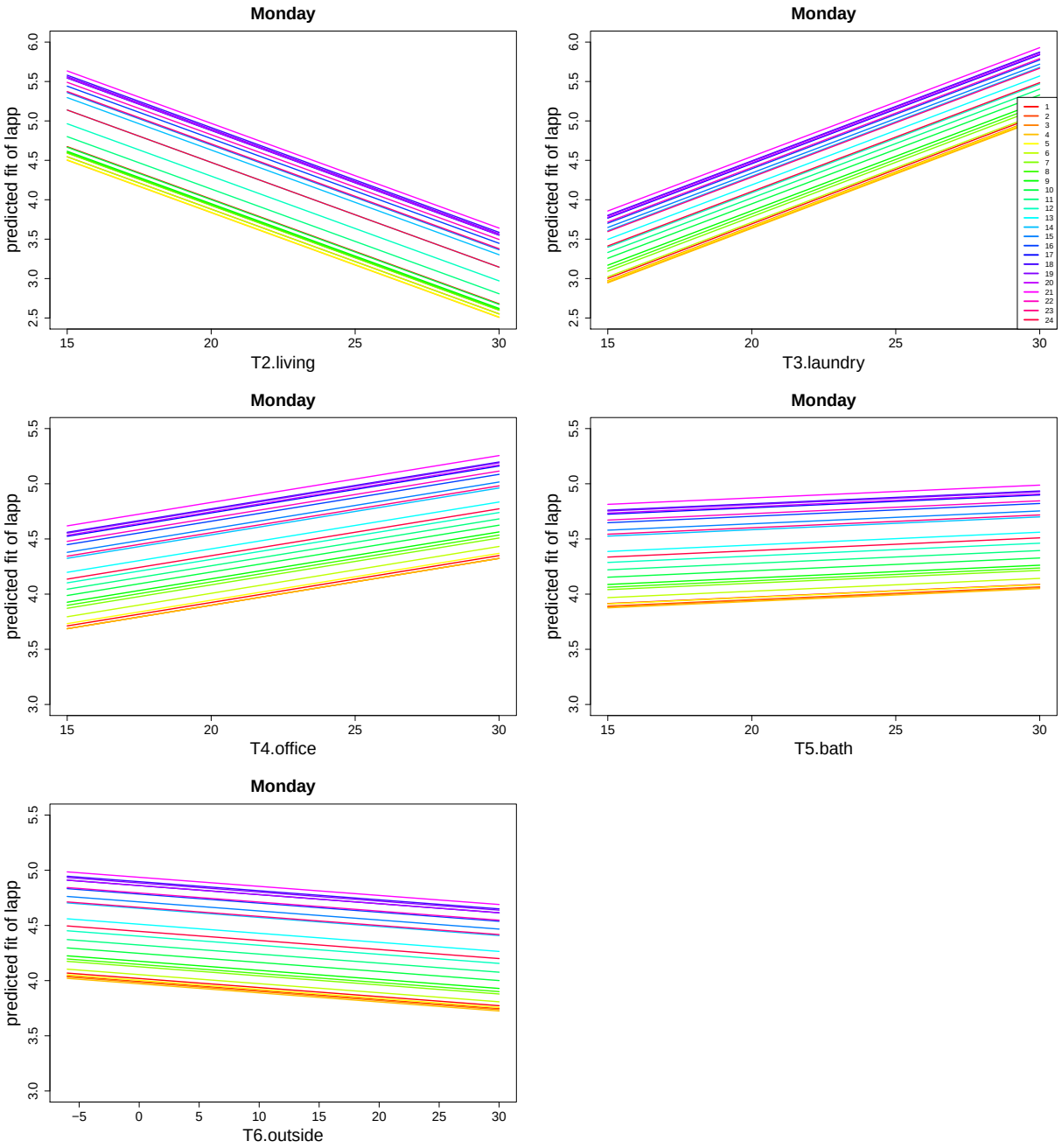



Figure 7.14: Prediction of the full main model (Model 7) restricted on weekdays. The plot shows the relationship between fitted $(\widehat{\texttt{lapp}}_{i_h})_{i_h=(i\in[19735]|\texttt{hour}=h,\text{with}h=1,\ldots,24)}$ and temperatures $(\texttt{Tj})_{j=\texttt{2.living},\texttt{3.laundry},\texttt{4.office},\texttt{5.bath},\texttt{6.outside}}$ to see the variation of hourly-wise pattern for Mondays, while the other covariates fixed at their medians. All weekdays of these room temperatures can be found in the Appendix, Figure A.8 - A.12. Condition $h$ is colored by hours 1 to 24.

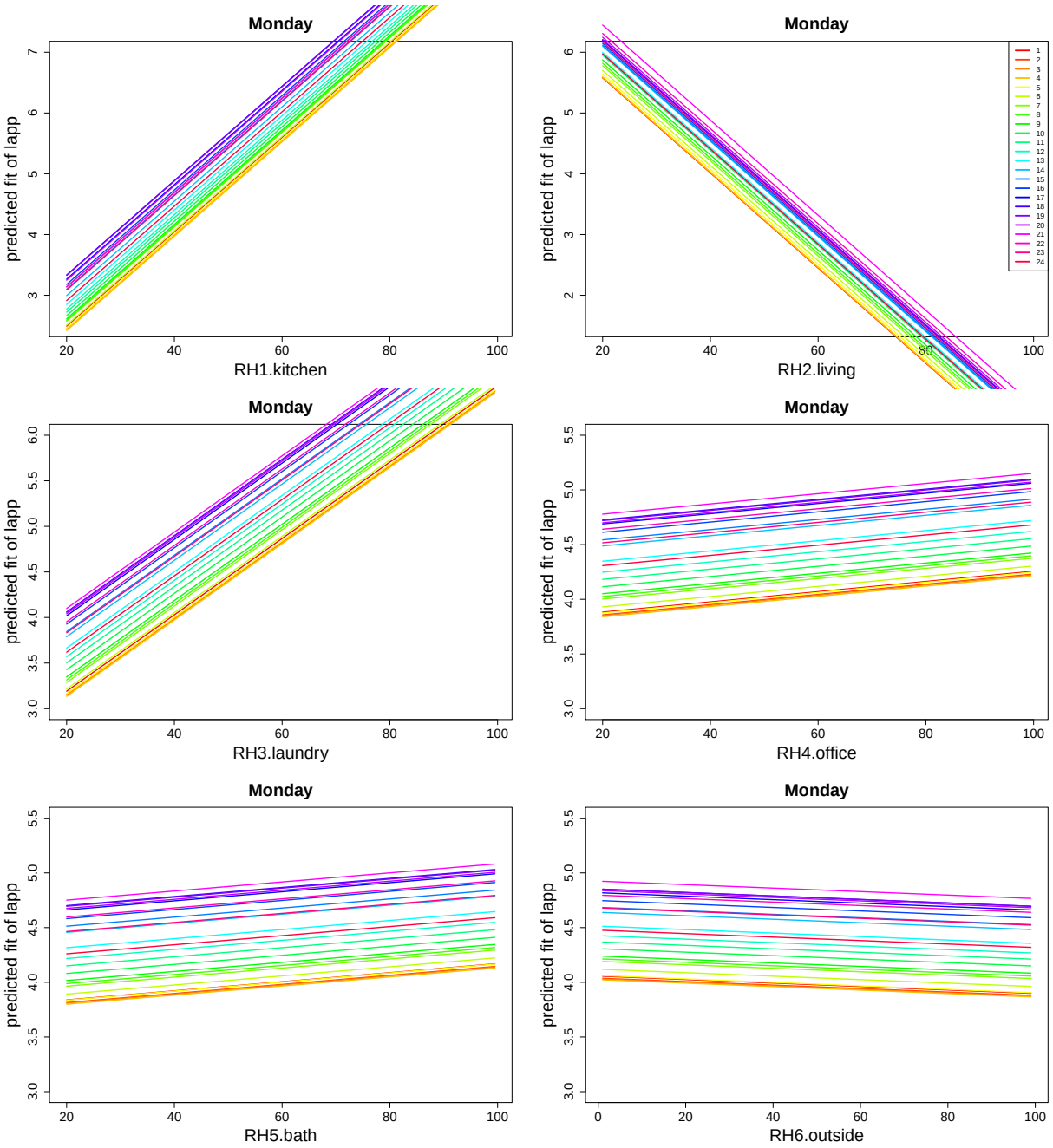



Figure 7.15: Prediction of full main model (Model 7) restricted on weekdays. The plot shows the relationship between fitted $(\widehat{\texttt{lapp}}_{i_h})_{i_h=(i\in[19735]|\texttt{hour}=h,\text{with}h=1,\ldots,24)}$ and humidities $(\texttt{RHj})_{j=\texttt{1.kitchen,2.living,3.laundry,4.office,5.bath,6.outside}}$ to see the variation of hourly-wise pattern for Mondays, while the other covariates fixed at their medians. Condition $h$ is colored by hours 1 to 24.

### 7.5.2 Prediction for interaction effects

Finally we look at the interaction effect predictions. For that, we use the same procedure as for the main effect prediction, but now based on the full interaction model (LMinter). For our prediction we prepare the following simplified formula with notation as (7.2):

$$\begin{aligned}\hat{Y}_i = \hat{\beta}_0 \quad & +\hat{\beta}_1 x_{i1} + \hat{\boldsymbol{\beta}}\tilde{\boldsymbol{x}}_i + \hat{\beta}_{21}\texttt{poly}(h_i,\texttt{1}) + \hat{\beta}_{22}\texttt{poly}(h_i,\texttt{2}) + \hat{\beta}_{23}\texttt{poly}(h_i,\texttt{3}) + \hat{\beta}_{24}\delta_i\texttt{weekday}_i \\ & +\hat{\beta}_{25}x_{i1}\texttt{poly}(h_i,\texttt{1}) + \hat{\beta}_{26}x_{i1}\texttt{poly}(h_i,\texttt{2}) + \hat{\beta}_{27}x_{i1}\texttt{poly}(h_i,\texttt{3}) \\ & +\hat{\beta}_{28}\delta_i\texttt{weekday}_i\texttt{poly}(h_i,\texttt{1}) + \hat{\beta}_{29}\delta_i\texttt{weekday}_i\texttt{poly}(h_i,\texttt{2}) + \hat{\beta}_{30}\delta_i\texttt{weekday}_i\texttt{poly}(h_i,\texttt{3}) \\ & +\hat{\boldsymbol{\beta}}\tilde{\boldsymbol{x}}_i\texttt{poly}(h_i,\texttt{1}) + \hat{\boldsymbol{\beta}}\tilde{\boldsymbol{x}}_i\texttt{poly}(h_i,\texttt{2}) + \hat{\boldsymbol{\beta}}\tilde{\boldsymbol{x}}_i\texttt{poly}(h_i,\texttt{3}).\end{aligned} \tag{7.3}$$

Intersections can be observed now in Figure 7.16 and 7.17, which is the summary of Figure A.13, A.14, A.15, A.16 and A.17 in the Appendix, since we are allowing the covariates to interact with the time effect `poly(hour, 3)`. Thus contrary to the main effect predictions the hourly-wise slopes are not parallel to each other. Since in our full interaction model (LMinter) where no interaction allowed with weekdays, there are similarities between the weekdays. Identical to the main effect predictions, we detect some parallel shifts of the hourly-wise slopes comparing the seven weekday predictions.

**Interpretation:** First, focusing on the covariate `T1.kitchen` in Figure 7.16, almost all 24 hourly slopes intersect. If we now take a closer look at the prediction, we discover interesting similar behaviors for some hours. That is, the afternoon hours, $h = 14, \ldots, 19$, which are represented in the color variation blue, behave the same with a positive relation between the kitchen temperature and the appliances energy consumption, whereas the late evening or the morning hours almost intersect vertically with the afternoon hours, i.e. a negative relation between temperature and energy use.

This means that if the behavior in the afternoon will change, so the evening and morning behavior.

For the similarity reasons between the weekdays, remaining rooms summarized in one plot in Figure 7.17.

As visible in this plot the afternoon hours also behave the same in the other main rooms, but in different shapes. The least interaction can be seen in the bathroom where only the later evening hour lines cut the morning hour lines. So only the morning hours and evening hours influence each other.

Furthermore, we provide the LMinter prediction for the relevant and corresponding humidities in Figure 7.18. The humidity predictions support our findings in the temperature predictions. The figure also shows the same afternoon behavior like we detected in the temperature predictions.

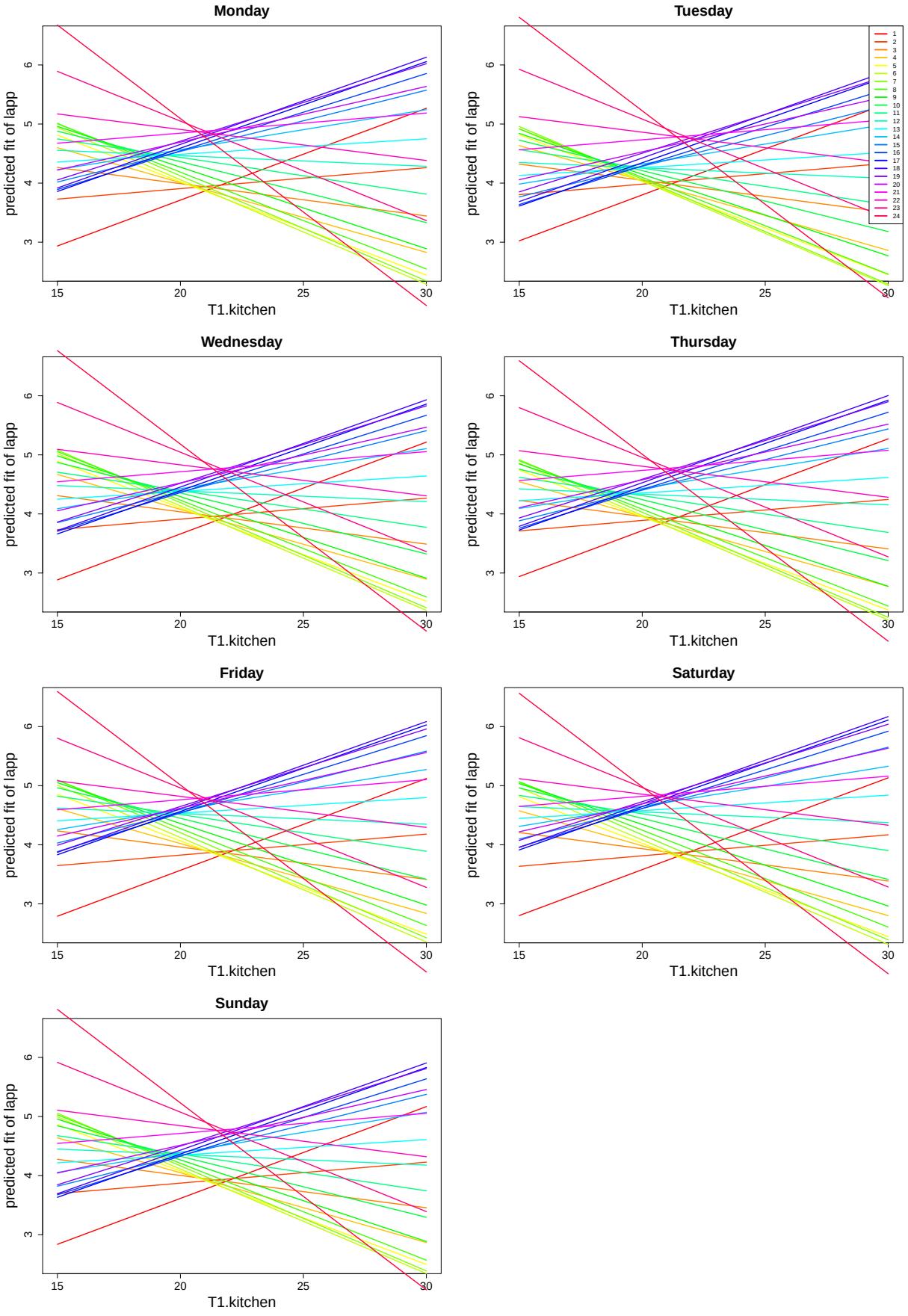



Figure 7.16: Prediction of LMinter restricted on weekdays. The plot shows the relationship between fitted $(\widehat{\texttt{lapp}}_{i_h})_{i_h=(i\in[19735]|\texttt{hour}=h,\text{with}h=1,\ldots,24)}$ and temperature `T1.kitchen` to see the variation of hourly-wise pattern for all the weekdays, while the other covariates fixed at their medians. Condition $h$ is colored by hours 1 to 24.

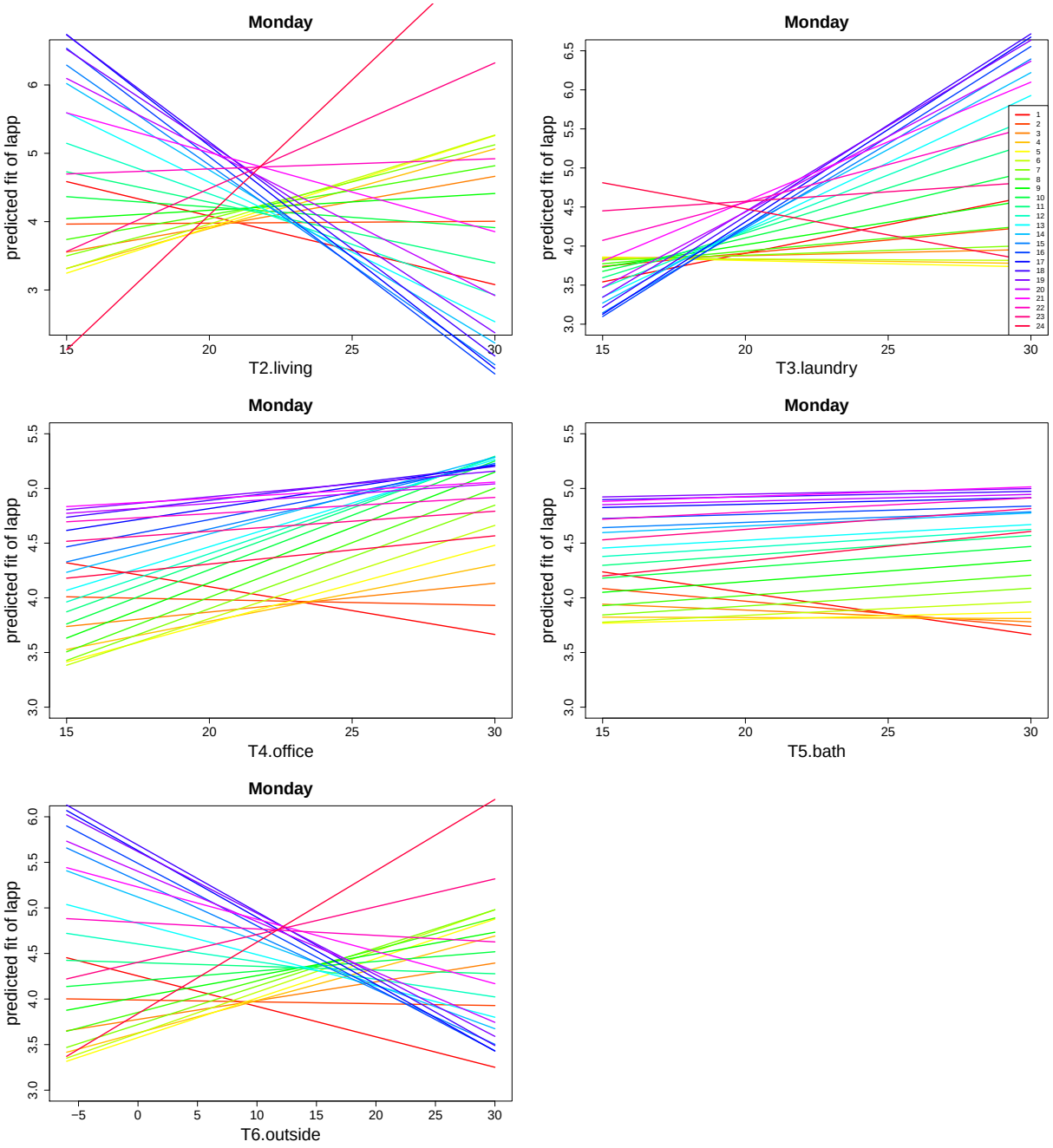



Figure 7.17: Prediction of LMinter restricted on weekdays. The plot shows the relationship between fitted $(\widehat{\texttt{lapp}}_{i_h})_{i_h=(i\in[19735]|\texttt{hour}=h,\text{with}h=1,\ldots,24)}$ and temperatures $(\texttt{Tj})_{j=\texttt{2.living},\texttt{3.laundry},\texttt{4.office},\texttt{5.bath},\texttt{6.outside}}$ to see the variation of hourly-wise pattern for Mondays, while the other covariates fixed at their medians. All weekdays of these room temperatures can be found in the Appendix, Figure A.13 - A.17. Condition $h$ is colored by hours 1 to 24.

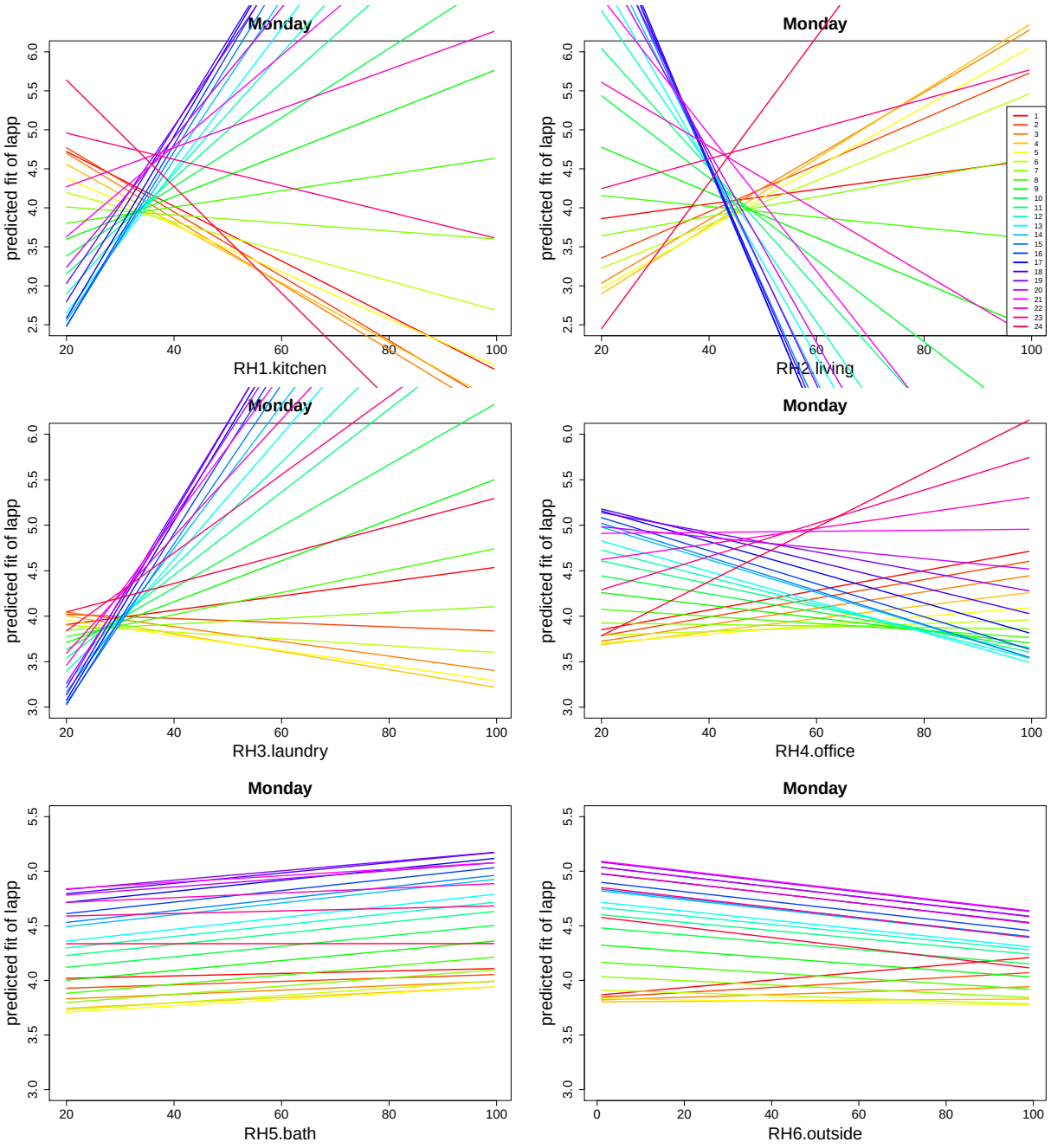



Figure 7.18: Prediction of LMinter restricted on weekdays. The plot shows the relationship between fitted $(\widehat{\texttt{lapp}}_{i_h})_{i_h=(i\in[19735]|\texttt{hour}=h,\text{with}h=1,\ldots,24)}$ and humidities $(\texttt{RHj})_{j=\texttt{1.kitchen,2.living,3.laundry,4.office,5.bath,6.outside}}$ to see the variation of hourly-wise pattern for Mondays, while the other covariates fixed at their medians. Condition $h$ is colored by hours 1 to 24.

# Chapter 8

# Generalized additive models (GAM's) for energy consumption within a house

At this point, we already inspected our energy data set with some plots and fitted a linear models on our energy use data set with the `lm()`-function, where we said that `lapp` is a linear function of temperature, humidity and an additional time effect. Now we are also doing predictions using GAM for modeling the relationship.

## 8.1 Main effect model

Generally, with a non-linear relationship, the linear model does a poor job fitting the data. Review the data set and our residual plots which clearly has some non-linear pattern, we now applying the method of the generalized additive model. Does this non-linear approach fitting the data better? Do we get a higher $R^2_{adj}$-value with GAM? Or does the model support our previous results?

We are starting with fitting a model to the energy use data where `lapp` has a smooth, non-linear relation to temperature, humidity and time effects using the `gam()`-function from the package `mgcv`. That is, the independent variables are modeled by smoothing spline function $s$. The method we use for the model fitting is REML, due to the arguments we discussed in Section 4.3.6. And at the end, we visualize the model fit and doing some predictions.

### 8.1.1 Setting the model

Since we already inspected the relation of the covariate and the response variable in the data exploration and the multiple linear regression setting separately, we immediately take all the covariates given into account and set our full GAM with $i = 1, \ldots, 19735$.

Our full main GAM model (GAMmain) has the structure

$$
\begin{aligned}
g(\texttt{lapp}_i) = \mathbf{A}_i\boldsymbol{\theta}_i \quad & +f_1(\texttt{T1.kitchen}_i) + f_2(\texttt{T2.living}_i) + f_3(\texttt{T3.laundry}_i) \\
& +f_4(\texttt{T4.office}_i) + f_5(\texttt{T5.bath}_i) + f_6(\texttt{T6.outside}_i) \\
& +f_7(\texttt{T7.ironing}_i) + f_8(\texttt{T8.teenager}_i) + f_9(\texttt{T9.parents}_i) \\
& +f_{10}(\texttt{T.outstation}_i) + f_{11}(\texttt{RH1.kitchen}_i) + f_{12}(\texttt{RH2.living}_i) \\
& +f_{13}(\texttt{RH3.laundry}_i) + f_{14}(\texttt{RH4.office}_i) + f_{15}(\texttt{RH5.bath}_i) \\
& +f_{16}(\texttt{RH6.outside}_i) + f_{17}(\texttt{RH7.ironing}_i) + f_{18}(\texttt{RH8.teenager}_i) \\
& +f_{19}(\texttt{RH9.parents}_i) + f_{20}(\texttt{RH.outstation}_i) + f_{21}(\texttt{hour}_i) + \varepsilon_i,
\end{aligned} \tag{8.1}
$$

where all the temperature, humidity and hour covariates set into a smooth function $f$ so that these are non-parametric coefficients, and the parametric model matrix $\boldsymbol{A} \in \mathbb{R}^{19735\times 7}$ with its $i$-th row

$$
\mathbf{A}_i = \begin{pmatrix} \texttt{weekday.Monday}_i \\ \texttt{weekday.Tuesday}_i \\ \texttt{weekday.Wednesday}_i \\ \texttt{weekday.Thursday}_i \\ \texttt{weekday.Friday}_i \\ \texttt{weekday.Saturday}_i \\ \texttt{weekday.Sunday}_i \end{pmatrix}^T \in \mathbb{R}^{1\times 7}.
$$

The covariates in $\mathbf{A}$ included in the model as a factor, which is hence parametric in the model. Also note that the weekday included as factor and therefore not set in a smooth function, otherwise setting `weekday` in a smooth function there are not enough unique values for it and the `weekday` covariate will be ignored. Moreover $g(\cdot)$ is a known smooth monotonic link function, $\varepsilon_i$ is the Gaussian error term and $\boldsymbol{\theta} \in \mathbb{R}^7$ is the coefficient parameter.

The corresponding summary to our GAMmain (8.1) is presented in Table 8.1.

| Family: gaussian; Link function: identity | | | | |
|---|---|---|---|---|
| Parametric coefficients: | | | | |
| | Estimate | Std. Error | t value | Pr(>\|t\|) |
| (Intercept) | 4.35299 | 0.01080 | 403.22 | < 2e-16 *** |
| weekday.Tuesday | -0.13303 | 0.01508 | -8.82 | < 2e-16 *** |
| weekday.Wednesday | -0.04530 | 0.01559 | -2.91 | 0.0037 ** |
| weekday.Thursday | -0.10854 | 0.01559 | -6.96 | 3.5e-12 *** |
| weekday.Friday | -0.00748 | 0.01612 | -0.46 | 0.6427 |
| weekday.Saturday | 0.03275 | 0.01581 | 2.07 | 0.0383 * |
| weekday.Sunday | -0.07864 | 0.01531 | -5.14 | 2.8e-07 *** |
| Approximate significance of smooth terms: | | | | |
| | edf | Ref.df | F | p-value |
| s(T1.kitchen) | 8.56 | 8.91 | 12.94 | < 2e-16 *** |
| s(T2.living) | 6.55 | 7.67 | 38.86 | < 2e-16 *** |
| s(T3.laundry) | 5.19 | 6.39 | 66.06 | < 2e-16 *** |
| s(T4.office) | 8.57 | 8.92 | 26.34 | < 2e-16 *** |
| s(T5.bath) | 8.57 | 8.93 | 9.36 | 3.5e-14 *** |
| s(T6.outside) | 8.65 | 8.93 | 10.49 | 7.9e-13 *** |
| s(T7.ironing) | 5.84 | 7.08 | 6.53 | 9.0e-08 *** |
| s(T8.teenager) | 7.82 | 8.63 | 46.50 | < 2e-16 *** |
| s(T9.parents) | 8.35 | 8.84 | 18.11 | < 2e-16 *** |
| s(T.outstation) | 6.87 | 8.02 | 4.82 | 6.1e-06 *** |
| s(RH1.kitchen) | 8.30 | 8.83 | 31.31 | < 2e-16 *** |
| s(RH2.living) | 5.42 | 6.70 | 57.24 | < 2e-16 *** |
| s(RH3.laundry) | 7.30 | 8.32 | 43.88 | < 2e-16 *** |
| s(RH4.office) | 6.76 | 7.90 | 4.82 | 6.7e-06 *** |
| s(RH5.bath) | 7.06 | 8.13 | 10.21 | 1.9e-14 *** |
| s(RH6.outside) | 7.20 | 8.23 | 10.58 | 3.7e-15 *** |
| s(RH7.ironing) | 7.59 | 8.51 | 14.22 | < 2e-16 *** |
| s(RH8.teenager) | 8.29 | 8.84 | 16.02 | < 2e-16 *** |
| s(RH9.parents) | 8.02 | 8.74 | 13.74 | < 2e-16 *** |
| s(RH.outstation) | 7.29 | 8.17 | 9.78 | 7.9e-14 *** |
| s(hour) | 8.92 | 9.00 | 341.48 | < 2e-16 *** |

Table 8.1: Summary of full main GAM (GAMmain) (8.1): `lapp` $\sim$ `s(temperatures) + s(humidities) + as.factor(weekday) + s(hour)`. The full significant GAMmain reaches a coefficient of determination of $R^2_{adj} = 0.42$ and deviance explained $D = 42.5\%$.

### 8.1.2 Analyzing the GAM main model

In comparison to the linear fit, we predict the appliances as the sum of smooth functions of temperature, humidity and hours. Our full model GAMmain (c.f. (8.1) and Table 8.1) has significant smooth functions of all covariates, so we do not have to think about removing covariates like in our linear models before. Moreover, we reach a higher adjusted coefficient of determination than with the linear full main model (Model 7):

$$R^2_{adj}(\text{GAMmain}) = 0.42 > 0.33 = R^2_{adj}(\text{Model 7})$$

The higher $R^2_{adj}$ can be explained by the structure of the general additive model which is much better considering the subtleties and at fitting these data because it can capture the non-linear relationships between the variables.

Besides the $R^2_{adj}$, the summary in Table 8.1 shows the EDF, estimated degree of freedom, which practically says on how much the covariate is smoothed. That is, the higher the EDF, the more complex the splines. The $p$-value is also given, which still shows the statistical significance of given covariate on the response variable, which is tested by the $F$-test - lower $F$-value is better. In our summary, all the covariates are highly significance, except for the non-significant factor `weekday.Friday`. This problem, we already discussed in the linear regression application.

As illustrated by Figure 8.1 and 8.2 we see the marginal fits of the different areas influencing the appliances energy consumption. From the plots we see that we have done a good job in fitting the linear model in the last chapter. Except for the tails, the fitted line and the confidence interval a quite linear with slight gradient.

Results are given in Figure 8.3, which represents the marginal influence of the smooth hour function on the response, are coincide with the box-plot we provided in Figure 6.21, which supports a polynomial of degree three as argued before depending on Figure 6.21.

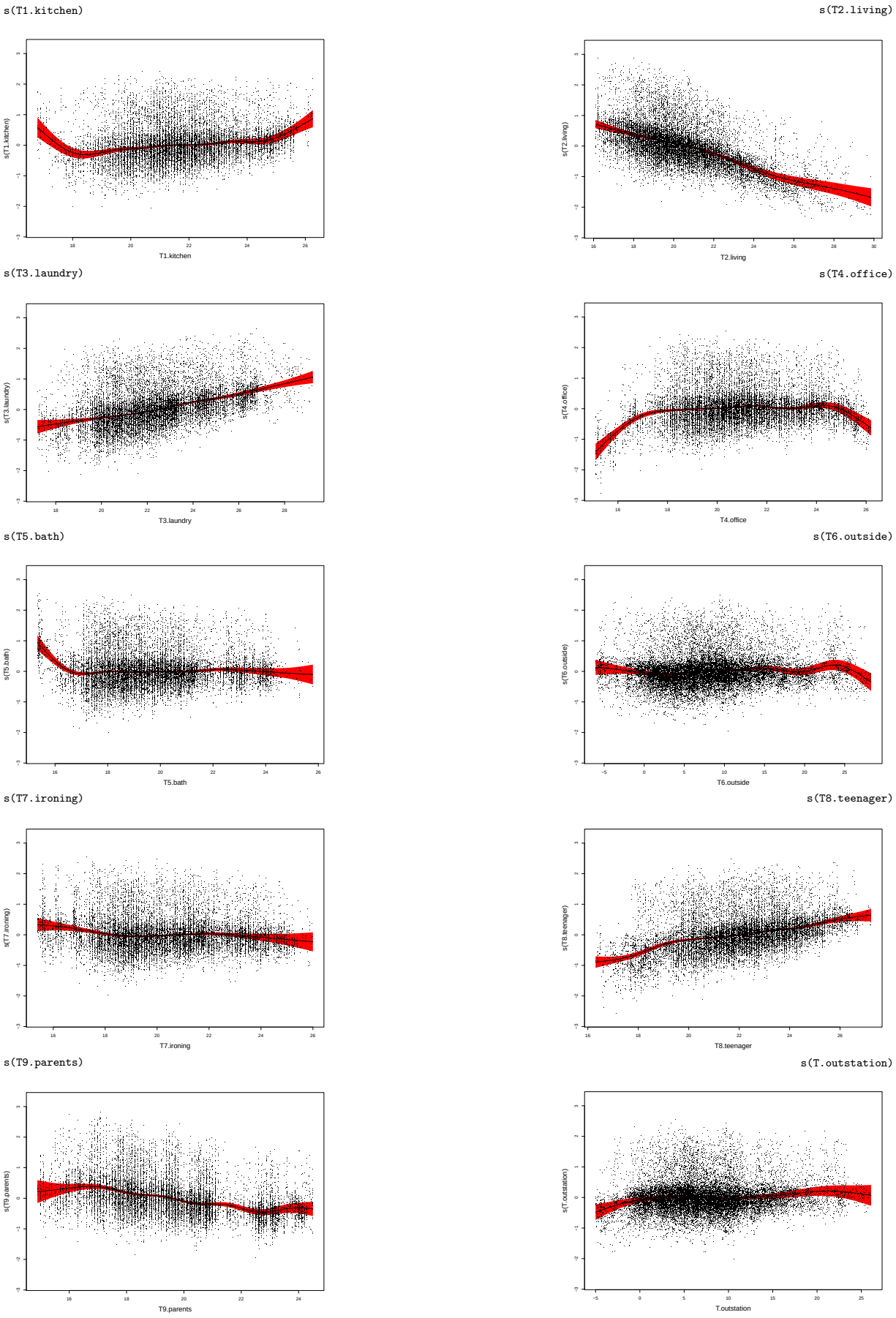



Figure 8.1: Plot of full main GAM (GAMmain) (8.1) from Table 8.1. Estimated marginal main effects on $\widehat{\texttt{lapp}}$ classified by room temperatures with fitted line, interval and residuals.

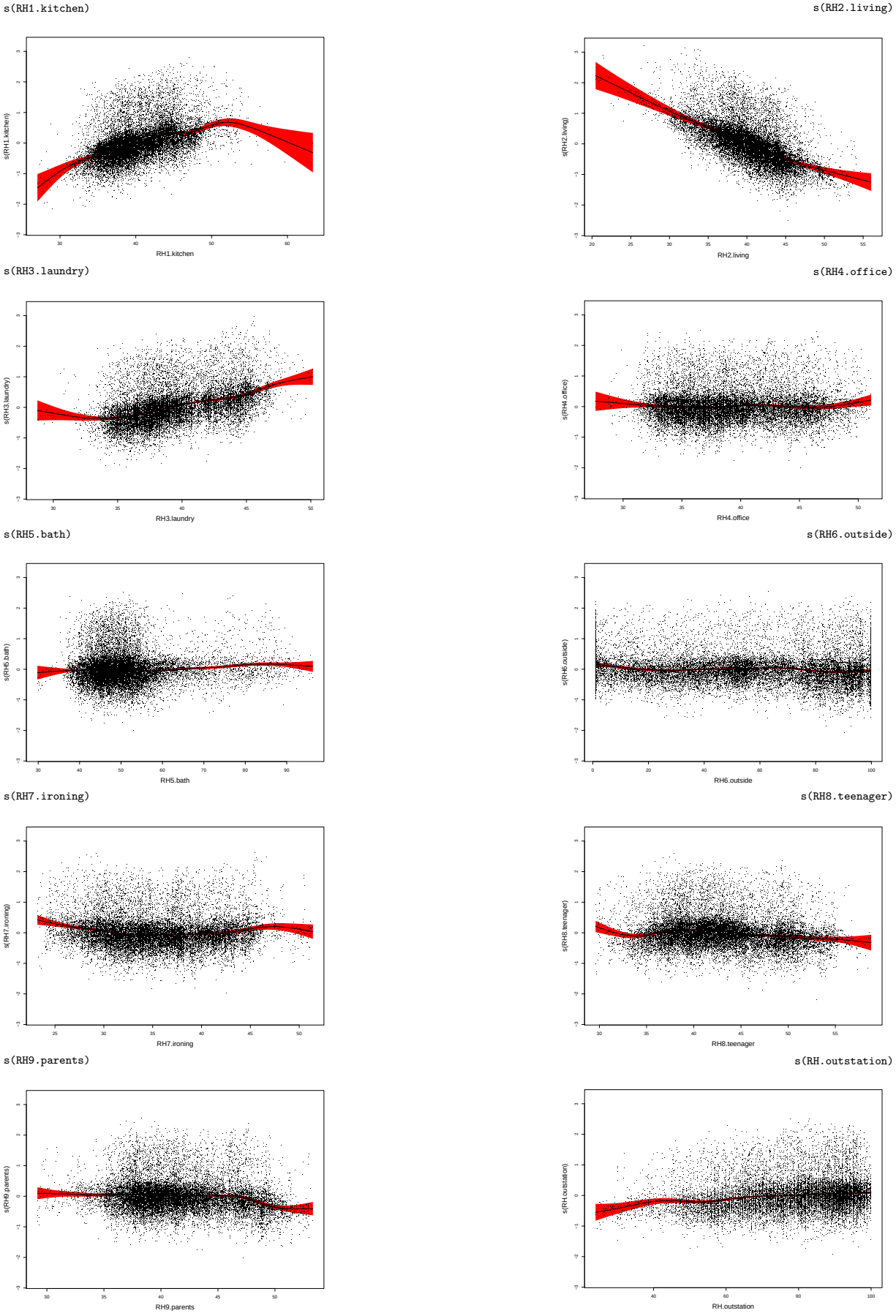



Figure 8.2: Plot of full main GAM (GAMmain) (8.1) from Table 8.1. Estimated marginal main effects on $\widehat{\texttt{lapp}}$ classified by room humidities with fitted line, interval and residuals.

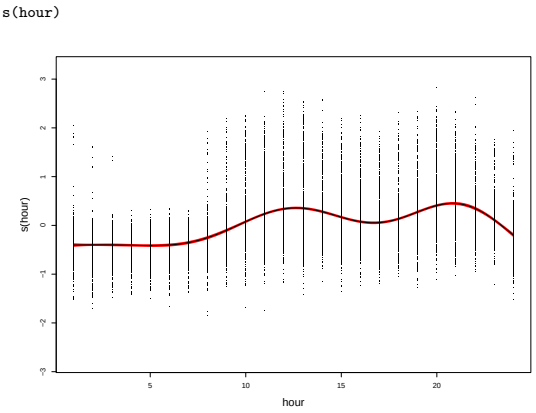



Figure 8.3: Plot of full main GAM (GAMmain) (8.1) from Table 8.1. Estimated marginal main effects on $\widehat{\texttt{lapp}}$ classified by time effect hours with fitted line, interval and residuals.

## 8.2 Interaction model

At this point we just fitted `lapp` with just the main effect, with temperatures, humidities and times. But now we also want to fit a model with smooths and tensor interactions to separate out the independent and interacting effects of covariates.

### 8.2.1 Setting the interaction model

Before setting the interaction model we have to choose between the methods to include tensor product interaction term. (c.f. Wood and Wood (2019), package `mgcv` and Wood (2017), Section 7.1) `ti()` gives a tensor product interaction that is appropriate with included main effects.

Our full interaction GAM model (GAMinter) has the structure

$$\begin{aligned} g(\texttt{lapp}_i) = \mathbf{A}_i\boldsymbol{\theta}_i \quad & +f_1(\texttt{T1.kitchen}_i) + f_2(\texttt{T2.living}_i) + f_3(\texttt{T3.laundry}_i) \\ & +f_4(\texttt{T4.office}_i) + f_5(\texttt{T5.bath}_i) + f_6(\texttt{T6.outside}_i) \\ & +f_7(\texttt{T7.ironing}_i) + f_8(\texttt{T8.teenager}_i) + f_9(\texttt{T9.parents}_i) \\ & +f_{10}(\texttt{T.outstation}_i) + f_{11}(\texttt{RH1.kitchen}_i) + f_{12}(\texttt{RH2.living}_i) \\ & +f_{13}(\texttt{RH3.laundry}_i) + f_{14}(\texttt{RH4.office}_i) + f_{15}(\texttt{RH5.bath}_i) \\ & +f_{16}(\texttt{RH6.outside}_i) + f_{17}(\texttt{RH7.ironing}_i) + f_{18}(\texttt{RH8.teenager}_i) \\ & +f_{19}(\texttt{RH9.parents}_i) + f_{20}(\texttt{RH.outstation}_i) + f_{21}(\texttt{hour}_i) \\ & +ti_1(\texttt{T1.kitchen}_i, \texttt{hour}_i) + ti_2(\texttt{T2.living}_i, \texttt{hour}_i) \\ & +ti_3(\texttt{T3.laundry}_i, \texttt{hour}_i) + ti_4(\texttt{T4.office}_i, \texttt{hour}_i) \\ & +ti_5(\texttt{T5.bath}_i, \texttt{hour}_i) + ti_6(\texttt{T6.outside}_i, \texttt{hour}_i) \\ & +ti_7(\texttt{T7.ironing}_i, \texttt{hour}_i) + ti_8(\texttt{T8.teenager}_i, \texttt{hour}_i) \\ & +ti_9(\texttt{T9.parents}_i, \texttt{hour}_i) + ti_{10}(\texttt{T.outstation}_i, \texttt{hour}_i) \\ & +ti_{11}(\texttt{RH1.kitchen}_i, \texttt{hour}_i) + ti_{12}(\texttt{RH2.living}_i, \texttt{hour}_i) \\ & +ti_{13}(\texttt{RH3.laundry}_i, \texttt{hour}_i) + ti_{14}(\texttt{RH4.office}_i, \texttt{hour}_i) \\ & +ti_{15}(\texttt{RH5.bath}_i, \texttt{hour}_i) + ti_{16}(\texttt{RH6.outside}_i, \texttt{hour}_i) \\ & +ti_{17}(\texttt{RH7.ironing}_i, \texttt{hour}_i) + ti_{18}(\texttt{RH8.teenager}_i, \texttt{hour}_i) \\ & +ti_{19}(\texttt{RH9.parents}_i, \texttt{hour}_i) + ti_{20}(\texttt{RH.outstation}_i, \texttt{hour}_i) \\ & +ti_{21}(\texttt{weekday}_i, \texttt{hour}_i) + \varepsilon_i, \end{aligned} \tag{8.2}$$

where the setting is as explained in the full model GAMmain (8.1). The summary with the significant covariates are shown in Table 8.2. We do not have to reduce it. One more advantage of the GAM models.

The corresponding summary to our GAMinter (8.2) is presented in Table 8.2.

| Family: gaussian; Link function: identity | | | | |
|---|---|---|---|---|
| Parametric coefficients: | | | | |
| | Estimate | Std. Error | t value | Pr(>\|t\|) |
| (Intercept) | 4.44644 | 0.01842 | 241.36 | < 2e-16 *** |
| weekday.Tuesday | -0.10764 | 0.01622 | -6.64 | 3.3e-11 *** |
| weekday.Wednesday | -0.06195 | 0.01683 | -3.68 | 0.00023 *** |
| weekday.Thursday | -0.11872 | 0.01662 | -7.14 | 9.4e-13 *** |
| weekday.Friday | -0.00748 | 0.01761 | -0.42 | 0.67101 |
| weekday.Saturday | -0.03687 | 0.01748 | -2.11 | 0.03493 * |
| weekday.Sunday | -0.13407 | 0.01683 | -7.96 | 1.8e-15 *** |
| Approximate significance of smooth terms: | | | | |
| | edf | Ref.df | F | p-value |
| s(T1.kitchen) | 8.51 | 8.87 | 8.64 | 1.1e-12 *** |
| s(T2.living) | 4.98 | 6.20 | 6.81 | 2.2e-07 *** |
| s(T3.laundry) | 6.55 | 7.60 | 12.30 | < 2e-16 *** |
| s(T4.office) | 8.66 | 8.93 | 20.91 | < 2e-16 *** |
| s(T5.bath) | 8.63 | 8.93 | 9.20 | 5.6e-14 *** |
| s(T6.outside) | 8.24 | 8.81 | 4.99 | 7.7e-07 *** |
| s(T7.ironing) | 8.41 | 8.83 | 9.81 | 5.9e-15 *** |
| s(T8.teenager) | 8.13 | 8.75 | 41.73 | < 2e-16 *** |
| s(T9.parents) | 5.71 | 6.89 | 6.64 | 9.7e-08 *** |
| s(T.outstation) | 8.26 | 8.80 | 7.97 | 1.8e-10 *** |
| s(RH1.kitchen) | 7.93 | 8.64 | 6.87 | 8.5e-10 *** |
| s(RH2.living) | 5.75 | 6.94 | 11.25 | 7.8e-14 *** |
| s(RH3.laundry) | 7.69 | 8.54 | 18.77 | < 2e-16 *** |
| s(RH4.office) | 7.85 | 8.61 | 6.53 | 9.3e-08 *** |
| s(RH5.bath) | 5.42 | 6.58 | 3.57 | 0.00104 ** |
| s(RH6.outside) | 4.87 | 6.00 | 4.70 | 8.8e-05 *** |
| s(RH7.ironing) | 6.67 | 7.84 | 8.86 | 4.1e-12 *** |
| s(RH8.teenager) | 7.89 | 8.64 | 4.26 | 0.00019 *** |
| s(RH9.parents) | 8.45 | 8.88 | 16.05 | < 2e-16 *** |
| s(RH.outstation) | 3.39 | 4.30 | 2.33 | 0.03643 * |
| s(hour) | 8.84 | 8.99 | 88.49 | < 2e-16 *** |
| ti(T1.kitchen,hour) | 10.97 | 12.89 | 5.09 | 6.7e-09 *** |
| ti(T2.living,hour) | 12.29 | 13.61 | 10.14 | < 2e-16 *** |
| ti(T3.laundry,hour) | 12.81 | 14.40 | 17.38 | < 2e-16 *** |
| ti(T4.office,hour) | 12.61 | 14.37 | 6.42 | 2.5e-13 *** |
| ti(T5.bath,hour) | 12.73 | 14.56 | 5.63 | 2.6e-11 *** |
| ti(T6.outside,hour) | 4.39 | 4.89 | 9.41 | 6.2e-09 *** |
| ti(T7.ironing,hour) | 13.08 | 14.59 | 8.56 | < 2e-16 *** |
| ti(T8.teenager,hour) | 14.00 | 15.26 | 14.05 | < 2e-16 *** |
| ti(T9.parents,hour) | 6.60 | 8.20 | 8.95 | 1.7e-12 *** |
| ti(T.outstation,hour) | 12.91 | 14.41 | 8.74 | 5.9e-15 *** |
| ti(RH1.kitchen,hour) | 12.57 | 14.21 | 12.56 | < 2e-16 *** |
| ti(RH2.living,hour) | 11.70 | 13.39 | 13.29 | < 2e-16 *** |
| ti(RH3.laundry,hour) | 13.80 | 15.20 | 16.00 | < 2e-16 *** |
| ti(RH4.office,hour) | 11.70 | 13.86 | 4.13 | 3.0e-07 *** |
| ti(RH5.bath,hour) | 3.92 | 4.01 | 14.85 | 3.5e-12 *** |
| ti(RH6.outside,hour) | 12.03 | 14.10 | 5.87 | 8.6e-12 *** |
| ti(RH7.ironing,hour) | 13.26 | 14.65 | 9.30 | < 2e-16 *** |
| ti(RH8.teenager,hour) | 13.49 | 15.03 | 10.84 | < 2e-16 *** |
| ti(RH9.parents,hour) | 12.41 | 14.25 | 8.93 | < 2e-16 *** |
| ti(RH.outstation,hour) | 10.68 | 12.39 | 5.01 | 2.9e-08 *** |
| ti(weekday,hour) | 14.56 | 15.67 | 14.57 | < 2e-16 *** |

Table 8.2: Summary of interaction GAM (GAMinter) with hour: `lapp` $\sim$ `s(temperatures) + s(humidities) + as.factor(weekday) + s(hour) + ti(temperature, hour) + ti(humidities, hour) + ti(weekday, hour)`. The full significant GAMinter reaches a coefficient of determination of $R^2_{adj} = 0.5$ and deviance explained $D = 51\%$.

### 8.2.2 Analyzing the GAM interaction model

Like we see in Table 8.2 we have a better fit with using GAM-method, since

$$R^2_{adj}(\text{GAMinter}) = 0.5 > 0.39 = R^2_{adj}(\text{LMinter})$$

Furthermore, we see that for the main effects, the EDF-values are approximately the same, but as for the interaction terms, we see that the EDF-values are mostly higher which means that there is more complexity, i.e. more smooth basis, in the fitting.

As detailed in the plots below, we again see the main effects on the energy consumption, i.e. in Figure 8.4 the estimated marginal plots classified by temperatures, in Figure 8.7 the estimated marginal plots classified by humidities and in Figure 8.10 the estimated marginal plots of the time effect hour, using model GAMinter (8.2) in Table 8.2. Whereas in Figure 8.5, 8.8 and 8.11 we now see the interaction terms in form of an 3D surface. The surfaces do not have extremely changes or wiggles. i.e. the highest value is 2 and the lowest value is $-1.5$. Taking Figure 8.11 as an example, we see that on Saturday, i.e. weekday 6, we have the highest value of energy consumption between 10 and 15 pm. Additionally we can see that contrary to the weekends, in the beginning of weeks the energy consumption is higher in the morning hours and lower in the afternoon hours.

Furthermore we added the heat maps in Figure 8.6, 8.9 and 8.12 to support the interaction 3D surface plots. The contour plot shows the covariates on the axes and the contour lines and the corresponding number on the lines provides the effects of the covariates on the response variable `lapp`. Overall we can say that we have smooth contours which means that the model adapts less to response variable and the smoothing factor is higher to penalize the waves. In heat maps the extreme points can be recognized not only by the contours, but also by the colors. The whiter the color, the higher the value and so the influence, and the darker the color, up to red, the lower the values. The heat map of the interaction term of `ti(weekday, hour)` on `lapp`, in Figure 8.12, confirms our results we made for the corresponding 3D surface in Figure 8.11.

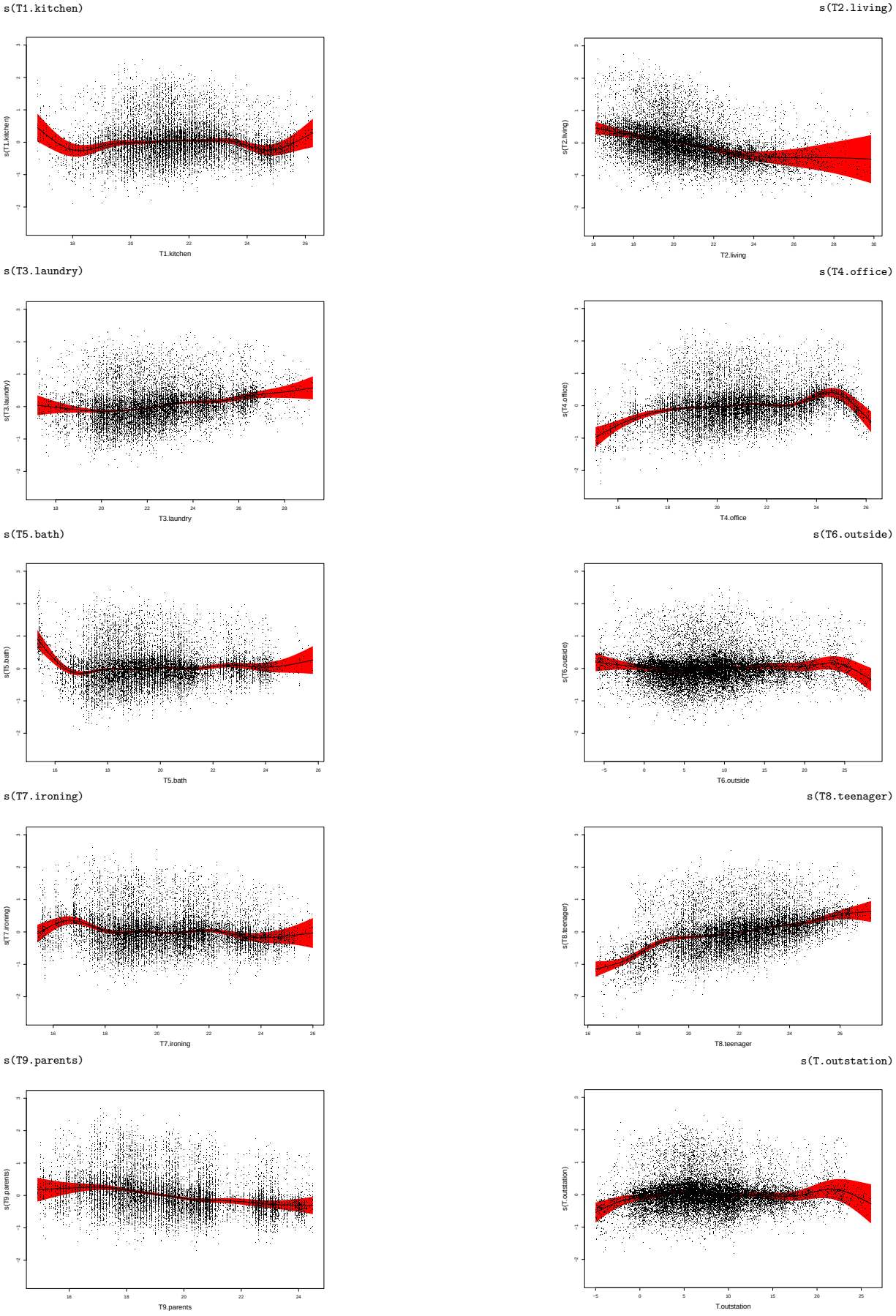



Figure 8.4: Plot of GAMinter (8.2). Estimated marginal main effects on $\widehat{\texttt{lapp}}$ classified by room temperatures with fitted line, interval and residuals.

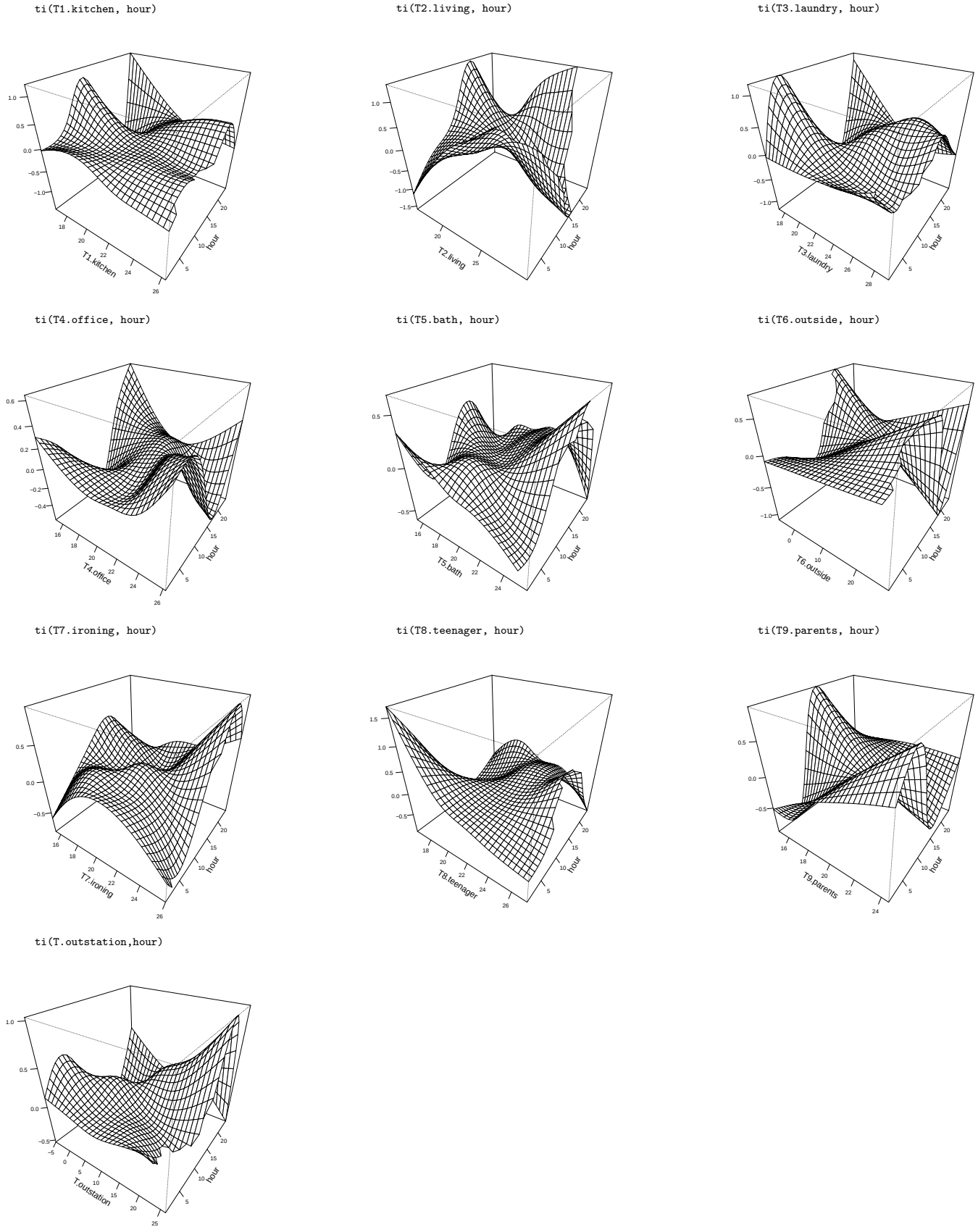



Figure 8.5: Plot of GAMinter (8.2). Estimated marginal interaction effects on $\widehat{\texttt{lapp}}$ classified by interaction terms between room temperatures and hours displayed as 3D surfaces.

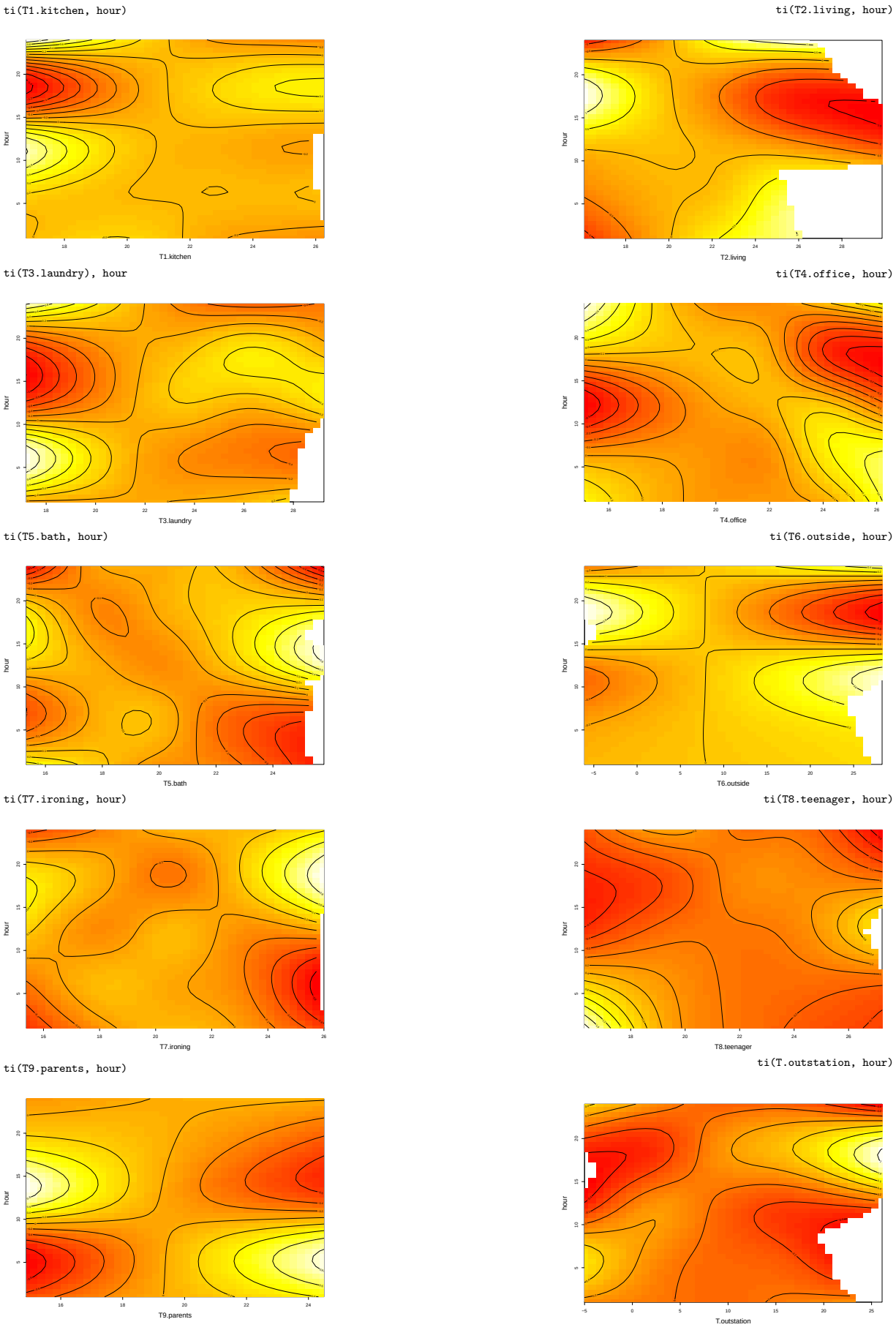



Figure 8.6: Plot of GAMinter (8.2). Estimated marginal interaction effects on $\widehat{\texttt{lapp}}$ classified by interaction terms between temperatures and hours displayed as colored heat maps.

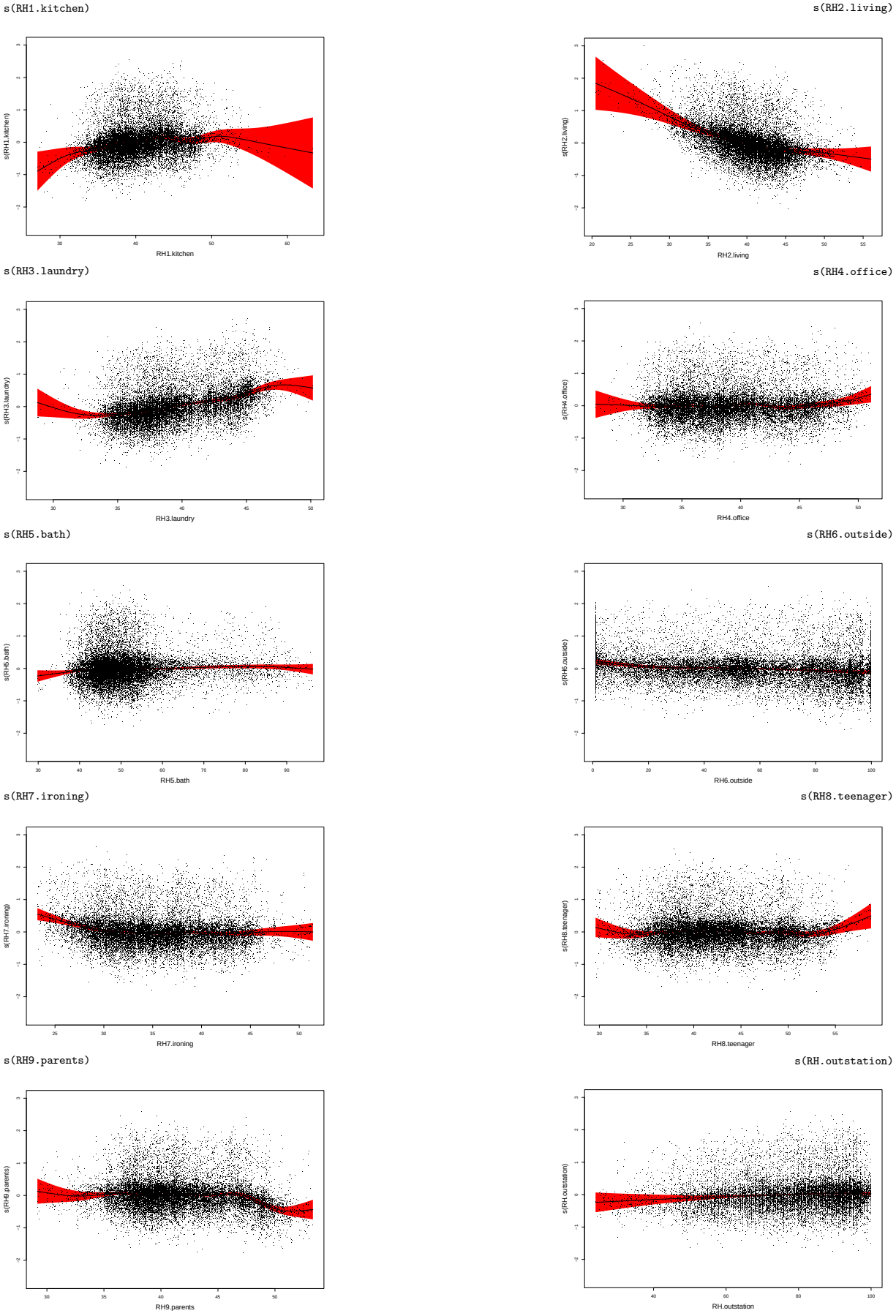



Figure 8.7: Plot of GAMinter (8.2). Estimated marginal main effects on $\widehat{\texttt{lapp}}$ classified by room humidities with fitted line, interval and residuals.

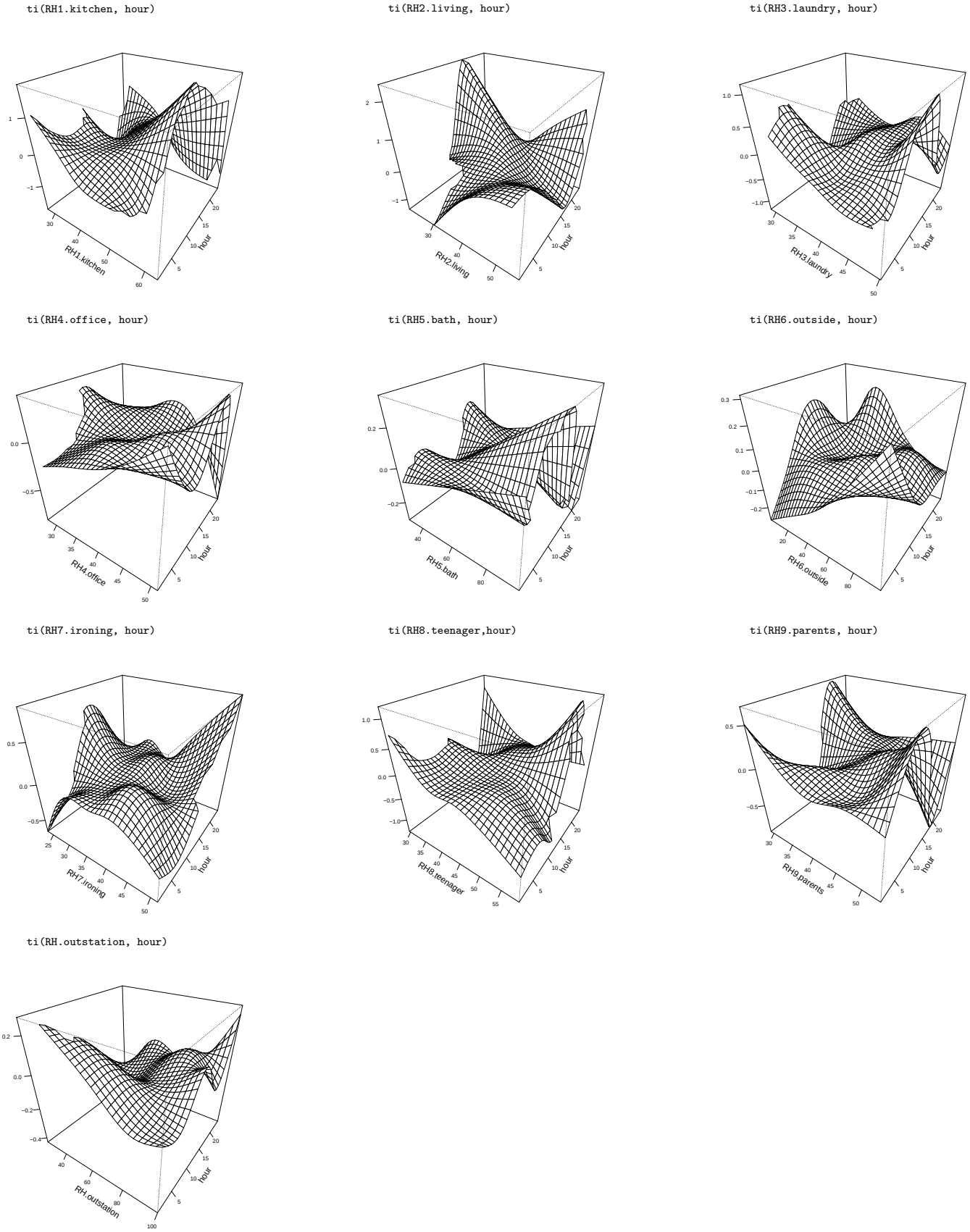



Figure 8.8: Plot of GAMinter (8.2). Estimated marginal interaction effects on $\widehat{\texttt{lapp}}$ classified by interaction terms between room humidities and hours displayed as 3D surfaces.

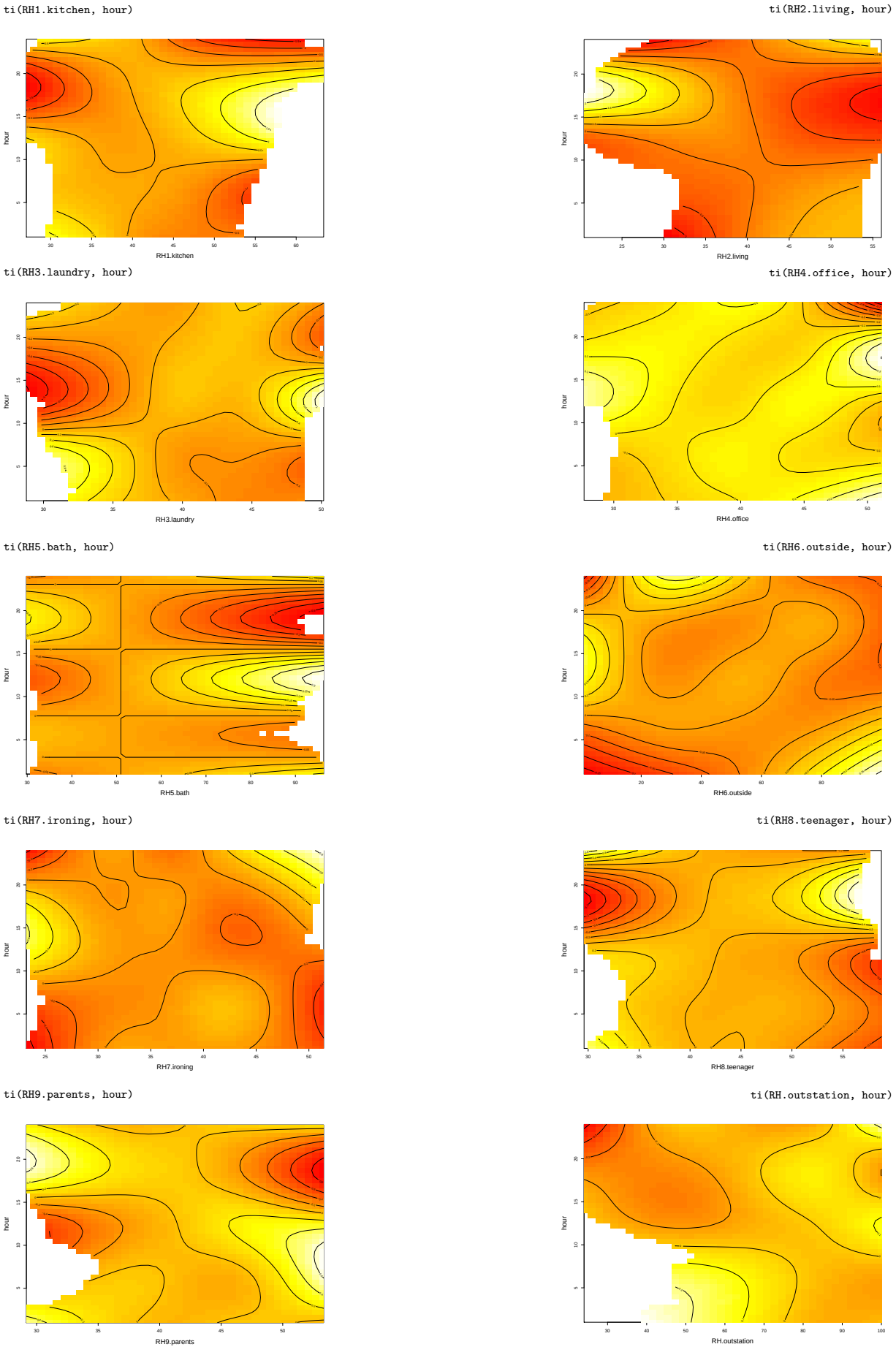



Figure 8.9: Plot of GAMinter (8.2). Estimated marginal interaction effects on $\widehat{\texttt{lapp}}$ classified by interaction terms between humidities and hours displayed as colored heat maps.

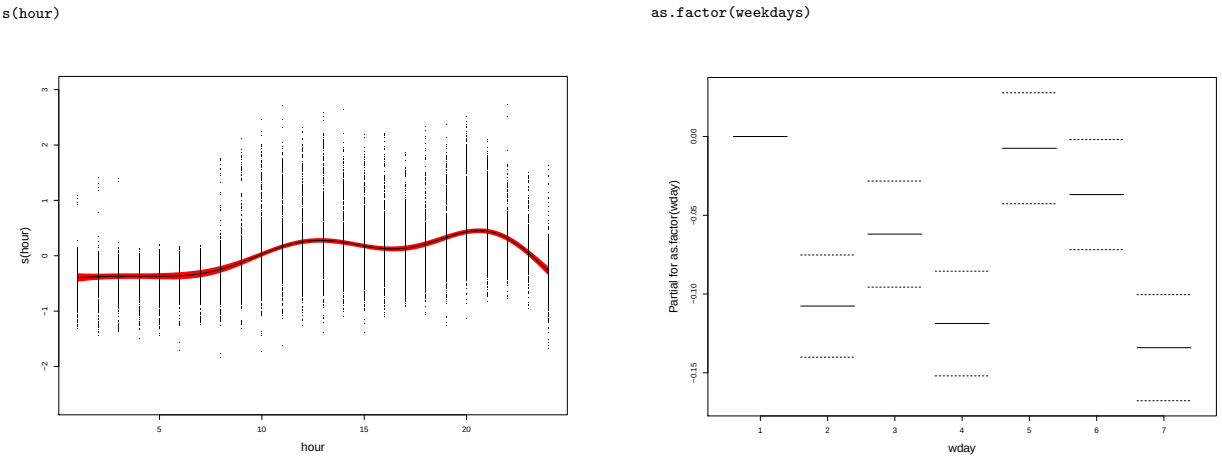



Figure 8.10: Plot of GAMinter (8.2). Estimated marginal main effects on $\widehat{\texttt{lapp}}$ classified by time effects, left panel hours with fitted line, interval and residuals, and right panel weekdays with their factorial influence.

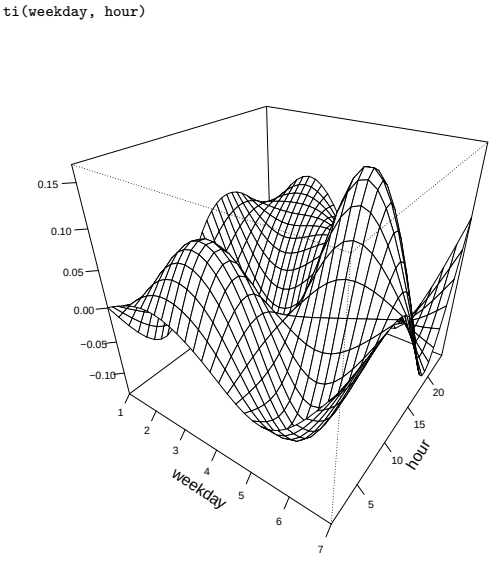



Figure 8.11: Plot of GAMinter (8.2). Estimated marginal interaction effects on $\widehat{\texttt{lapp}}$ classified by interaction terms between weekdays and hours displayed as 3D surface.

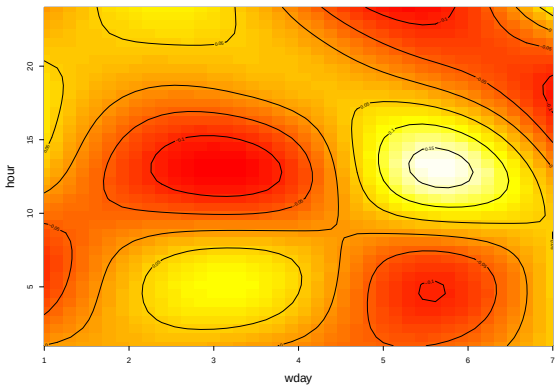



Figure 8.12: Plot of GAMinter (8.2). Estimated marginal interaction effects on $\widehat{\texttt{lapp}}$ classified by interaction terms between weekdays and hours displayed as colored heat maps.

## 8.3 Comparing main effect and interaction model

Finally we conclude the model setting with some statistics to compare these models.

| **GAMmain** | **GAMinter** |
|---|---|
| $R^2_{adj} = 0.42$ | $R^2_{adj} = 0.50$ |

Table 8.3: Comparing $R^2$ of full main and interaction GAM, GAMmain (8.1) and GAMinter (8.2), respectively.

| | df | AIC |
|---|---|---|
| **Main effect model** | | |
| GAMmain | 175.33 | 28808.11 |
| **Interaction effect model** | | |
| GAMinter | 438.59 | 26189.10 |

Table 8.4: Comparing all the created full main and interaction GAM, GAMmain (8.1) and GAMinter (8.2), with model selection criterion AIC (4.32).

We continue with the widely used model selection criterion in statistics, that is coefficients of determinations and AIC. For the equation of AIC we apply (4.32), where we have the number of parameters to be estimated subtracted by the maximized value of the likelihood function of the model. So we want the model with lower AIC values.

Finally, we can conclude from the above results in Table 8.3 and 8.4 that GAMinter is a better fit model due to the higher $R^2_{adj}$ and lower $AIC$ value.

$$\begin{array}{lllll} R^2_{adj}(\text{GAMmain}) & = 0.42 & < & 0.5 & = R^2_{adj}(\text{GAMinter}) \\ AIC(\text{GAMmain}) & = 28808.11 & > & 26189.10 & = AIC(\text{GAMinter}) \end{array}$$

Since the Table 8.1 and 8.2 showing significant $p$-values with suitable $F$-values, we can work with the full GAM models. Thus for main effect prediction we use the model GAMmain and for the interaction effect prediction we operate with the model GAMinter.

## 8.4 Predictions on the energy consumption

In line with the linear models, we also predict our created main and interaction GAM models, GAMmain and GAMinter.

### 8.4.1 Prediction for main effect

First, we inspect the model GAMmain (8.1) and handle the prediction input like we did with the linear model case. That is letting one covariate of interest be variable, while the other covariates are fixed to their median and time effect fixed to their values depending on the weekday and hour of interest.

**Interpretation** Starting with an overview of all the six temperature marginal prediction, we now identify more curvy, non-linear prediction graphs. Like hinted in the GAM regression plots in Figure 8.1, we have small wiggles or fluctuations which are indeed highly smoothed and big slope changes at the tails or limits in temperature prediction of `T1.kitchen` and `T4.office`. This can happen due to outliers and a lack of observations at the temperature limits. Cutting out these limits, we see that these prediction are consistent with the linear ones with some wiggles, compare to Figure 8.13 and 7.14. As for the covariates `T2.living` and `T3.laundry` we have quite similarities with the linear prediction we found in Figure 7.14, but here in Figure 8.14 we have slightly smooth curves.

Exploring `T5.bath` and `T6.outside`, there are differences between the linear and the GAM predictions, in Figure 7.14 and 8.14 respectively. While with the full main model (Model 7) we have clearly just one gradient, we can identify a quadratic or even cubic curve using the model GAMmain, i.e. a non-linear relationship. Its energy consumption peak is at 23 °C for the bathroom and highest peak near 24°C for the outside temperature.

Moreover, we have the same line pattern as for the linear model (Model 7) prediction. Since the curves of the hours are parallel to each other and are again slightly shifted comparing the weekdays, see Figure 8.13 and in the Appendix Figure A.18 to A.22 all the weekdays are listed of the summarized Figure 8.14.

Additionally, we provide the GAMmain prediction for the relevant and corresponding humidities in Figure 8.15. The predictions on humidity support our findings in the temperature predictions. The presence of the occupants increases the temperature and humidity and thus leads to device application. Note that due to thermodynamic reasons that rising temperature also causes a higher humidity absorption (c.f. Figure A.1). The figure also shows the same afternoon and evening behavior like we detected in the temperature predictions. As expected, the bathroom humidity stays almost constant with a slightly higher humidity level when devices are used.

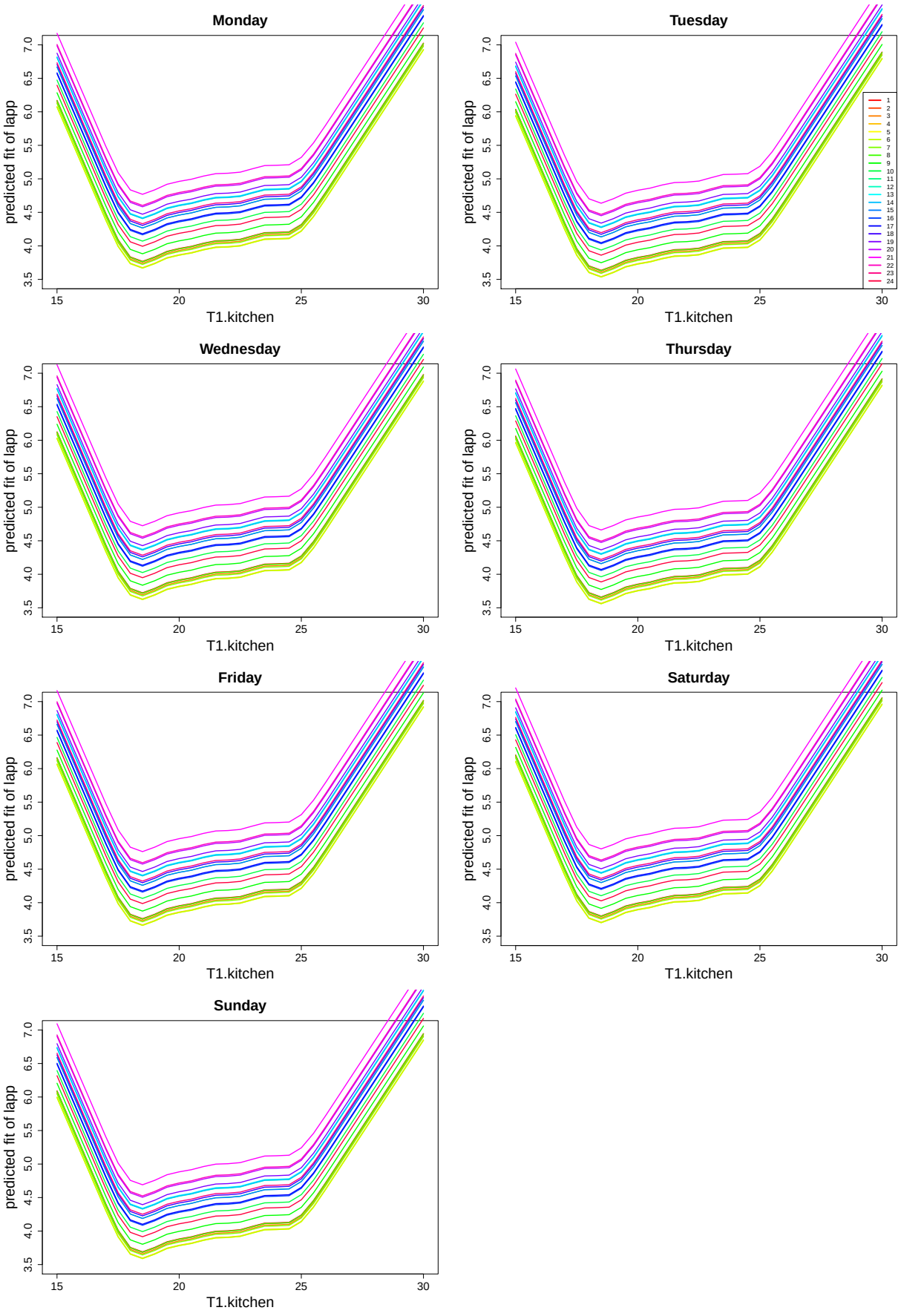



Figure 8.13: Prediction of GAMmain restricted on weekdays. The plot shows the relationship between fitted $(\widehat{\texttt{lapp}}_{i_h})_{i_h=(i\in[19735]|\texttt{hour}=h,\text{with}h=1,\ldots,24)}$ and temperature `T1.kitchen` to see the variation of hourly-wise pattern for all the weekdays, while the other covariates fixed at their medians. Condition $h$ is colored by hours 1 to 24.

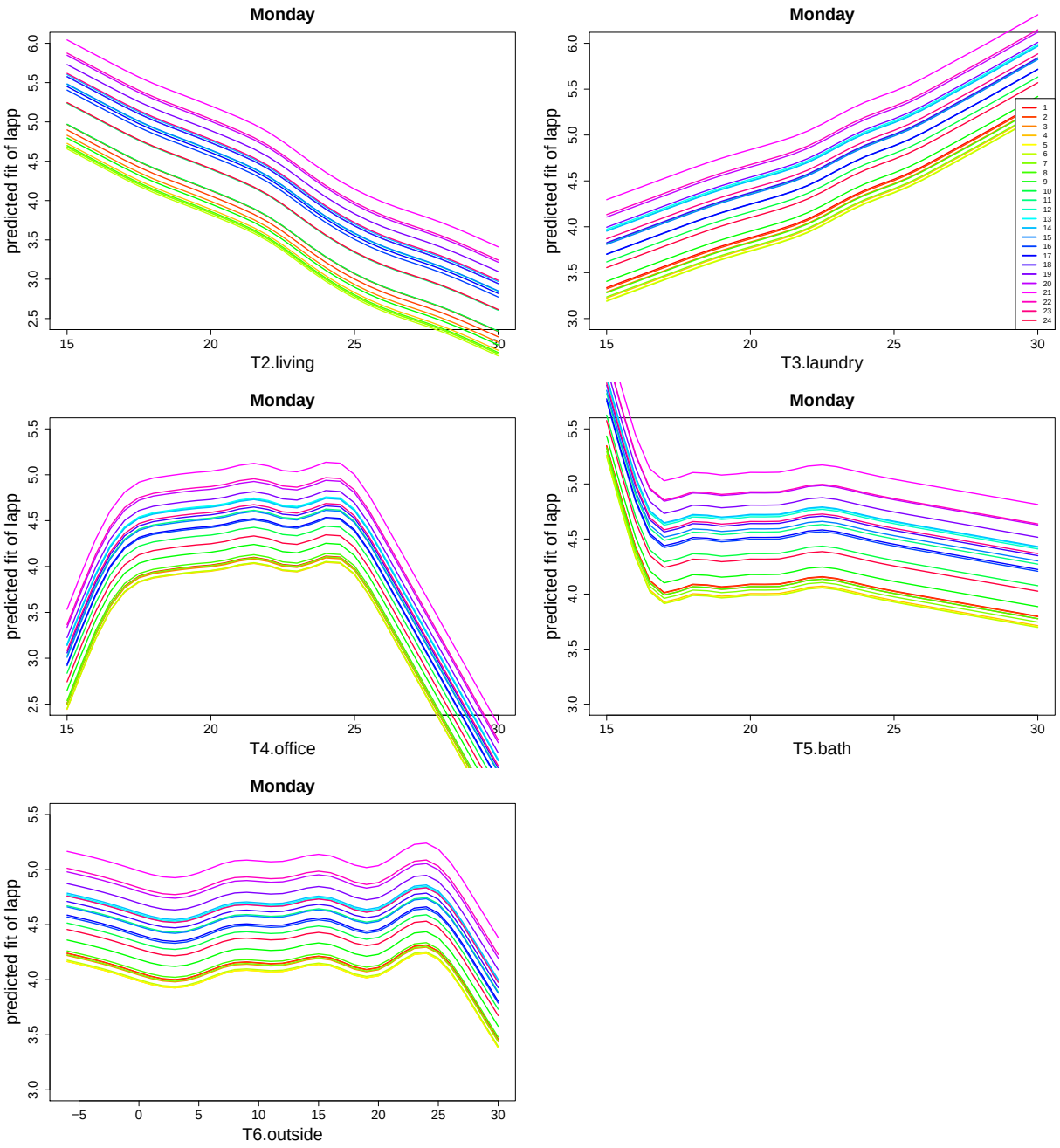



Figure 8.14: Prediction of GAMmain restricted on weekdays. The plot shows the relationship between fitted $(\widehat{\texttt{lapp}}_{i_h})_{i_h=(i\in[19735]|\texttt{hour}=h,\text{with}h=1,\ldots,24)}$ and temperatures $(\texttt{Tj})_{j=\texttt{2.living},\texttt{3.laundry},\texttt{4.office},\texttt{5.bath},\texttt{6.outside}}$ to see the variation of hourly-wise pattern for Mondays, while the other covariates fixed at their medians. All weekdays of these room temperatures can be found in the Appendix, Figure A.18 - A.22. Condition $h$ is colored by hours 1 to 24.

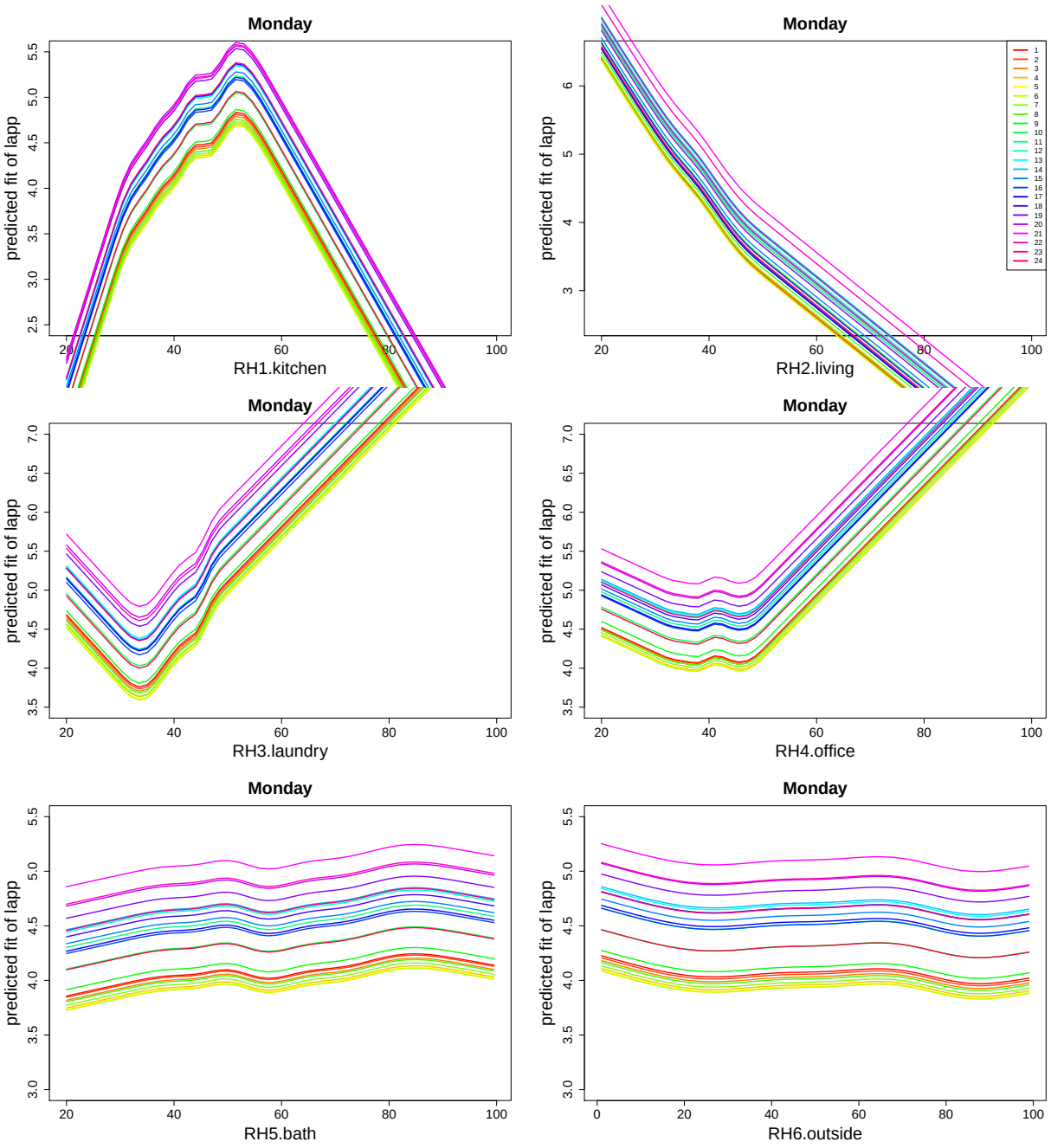



Figure 8.15: Prediction of GAMmain restricted on weekdays. The plot shows the relationship between fitted $(\widehat{\texttt{lapp}}_{i_h})_{i_h=(i\in[19735]|\texttt{hour}=h,\text{with}h=1,\ldots,24)}$ and humidities $(\texttt{RHj})_{j=\texttt{1.kitchen,2.living,3.laundry,4.office,5.bath,6.outside}}$ to see the variation of hourly-wise pattern for Mondays, while the other covariates fixed at their medians. Condition $h$ is colored by hours 1 to 24.

### 8.4.2 Prediction for interaction effect

Finally, we reveal the prediction of the model GAMinter we established in (8.2). The prediction input is managed analogously as for the GAMmain predictions and the linear models in the previous sections.

**Interpretation** The constructed hourly-wise prediction for GAMinter (8.2) has a higher difference to the final linear model LMinter.

But when we get more into details, we observe that the tendencies of the afternoon hours are just the same as for the linear predictions of LMinter. As anticipated in Figure 8.16 and 8.17, the hourly-wise prediction curves are more non-linear. In the afternoon hours we have the same pattern in the curved fittings, such as for the early morning hours. An overall impression is that the afternoon to evening hours interacting more powerful with the other hours. Since the blue to purple lines, which indicates the afternoon till evening hours, intersecting with the other hour curves virtually more obvious. Remember that the interaction says that a change in the afternoon hours will influence the behavior of the remaining hours.

Since the curves of the hours are parallel to each other and are again slightly shifted comparing the weekdays, see Figure 8.16 and in the Appendix Figure A.23 to A.27 all the weekdays are listed of the summarized Figure 8.17.

Additionally, we provide the prediction of GAMinter for the relevant and corresponding humidities in Figure 8.18. These predictions support our findings in the temperature predictions. The presence of the occupants increases the temperature and decreasing humidity and thus leads to device application. The figure also shows the same afternoon behavior and a higher usage in the evening hours, like we detected in the temperature predictions. But contrary to the temperature predictions the Figure 8.18 obviously has a higher degree of confusion or chaos for the kitchen, living, laundry room and office. This can be explained by the short or temporarily staying in this area. It is quite different in the bathroom and outside humidity with smoother curves. This confirms our assumption about almost constant high humidity level in the bathroom.

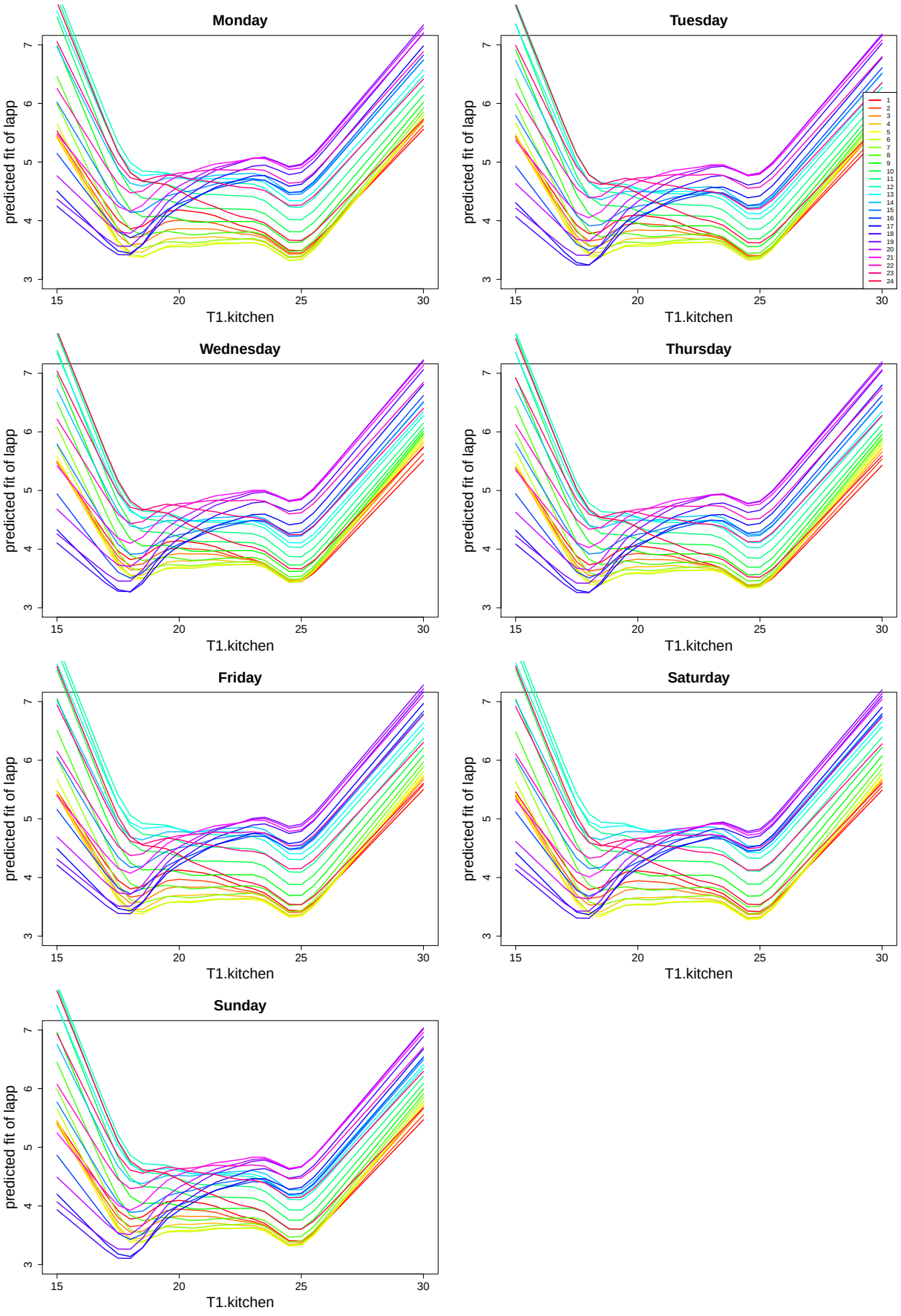



Figure 8.16: Prediction of GAMinter restricted on weekdays. The plot shows the relationship between fitted $(\widehat{\texttt{lapp}}_{i_h})_{i_h=(i\in[19735]|\texttt{hour}=h,\text{with}h=1,\ldots,24)}$ and temperature `T1.kitchen` to see the variation of hourly-wise pattern for all the weekdays, while the other covariates fixed at their medians. Condition $h$ is colored by hours 1 to 24.

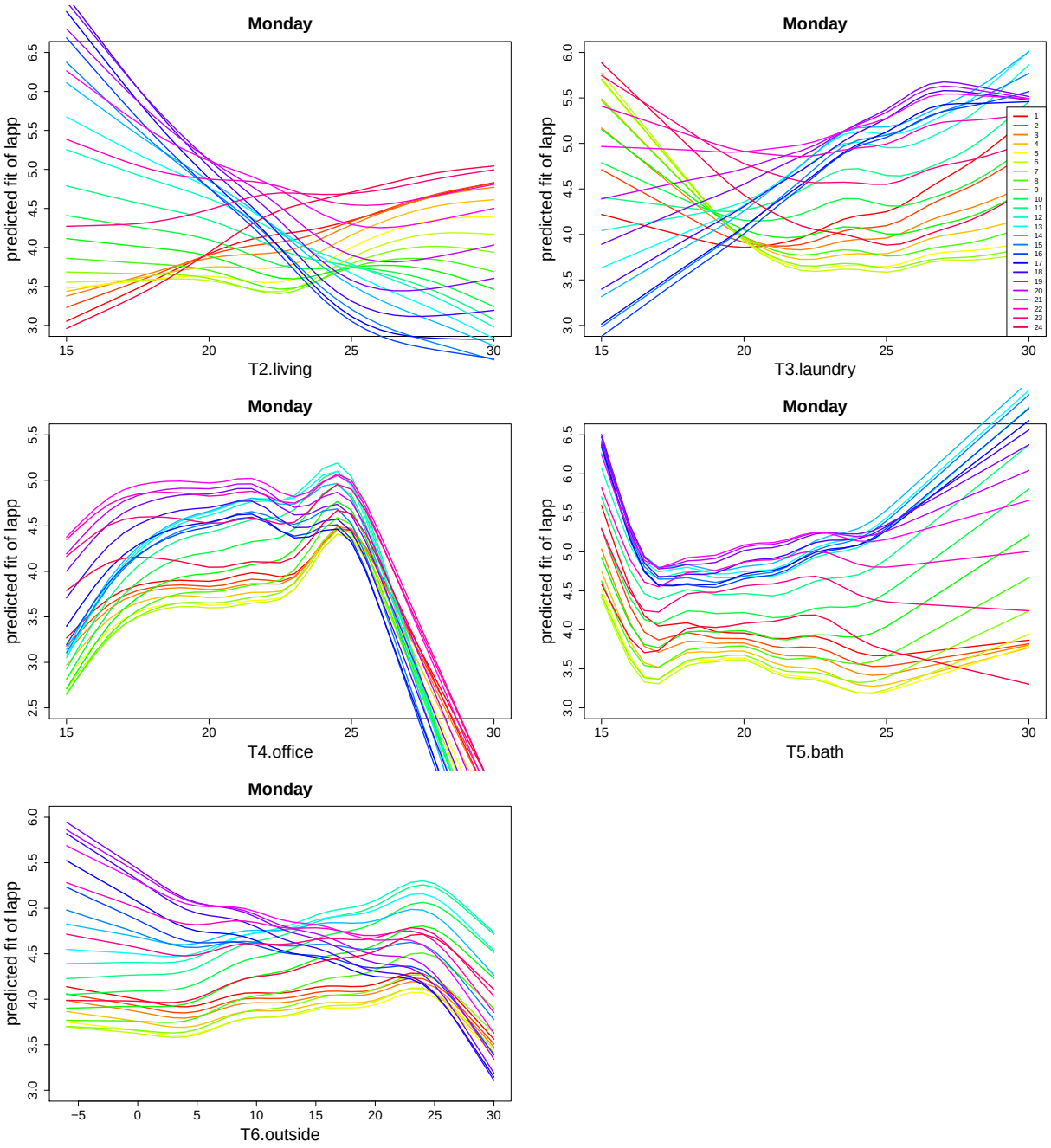



Figure 8.17: Prediction of GAMinter restricted on weekdays. The plot shows the relationship between fitted $(\widehat{\texttt{lapp}}_{i_h})_{i_h=(i\in[19735]|\texttt{hour}=h,\text{with}h=1,\ldots,24)}$ and temperatures $(\texttt{Tj})_{j=\texttt{2.living},\texttt{3.laundry},\texttt{4.office},\texttt{5.bath},\texttt{6.outside}}$ to see the variation of hourly-wise pattern for Mondays, while the other covariates fixed at their medians. All weekdays of these room temperatures can be found in the Appendix, Figure A.18 - A.22. Condition $h$ is colored by hours 1 to 24.

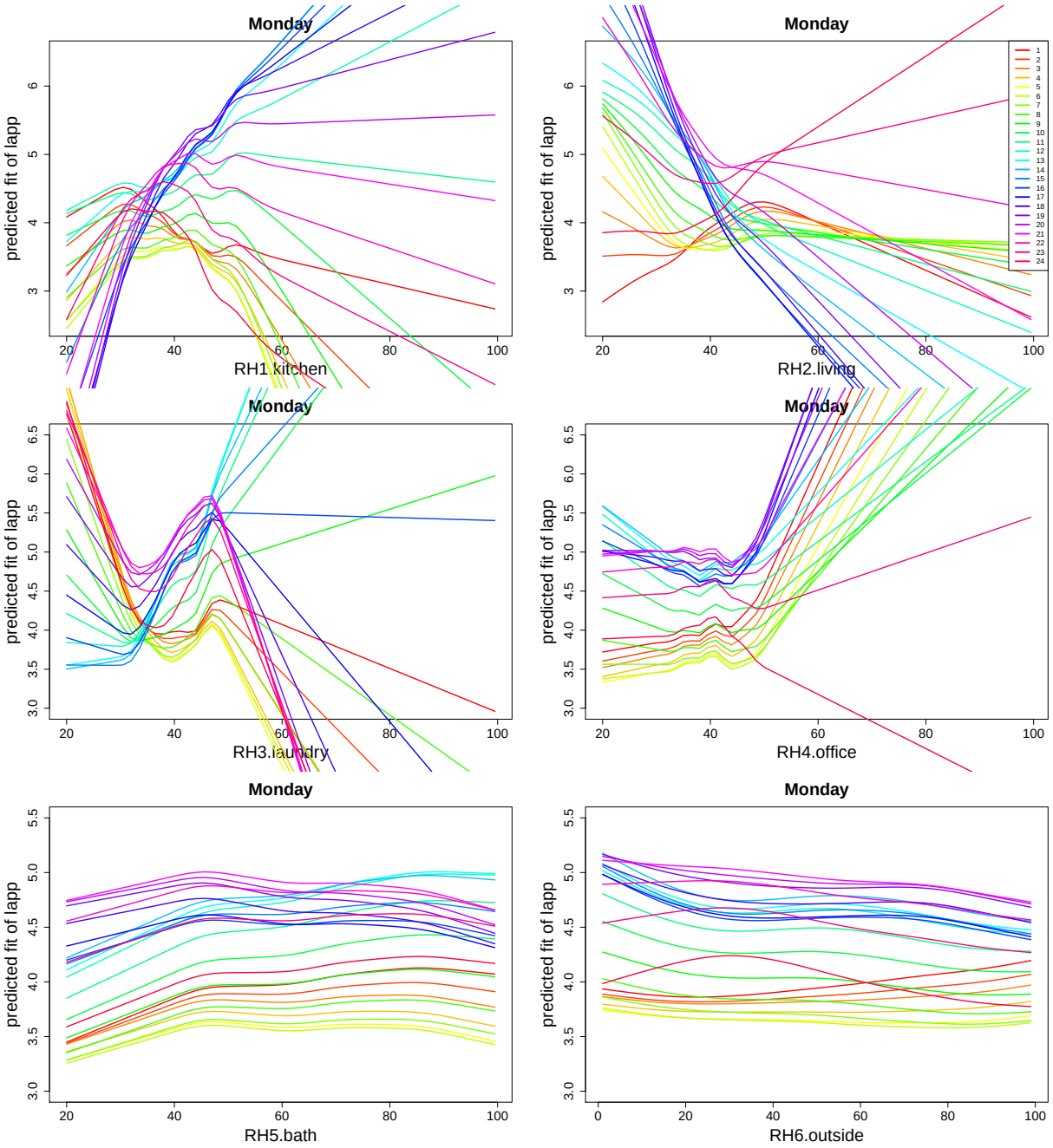



Figure 8.18: Prediction of GAMinter restricted on weekdays. The plot shows the relationship between fitted $(\widehat{\texttt{lapp}}_{i_h})_{i_h=(i\in[19735]|\texttt{hour}=h,\text{with}h=1,\ldots,24)}$ and humidities $(\texttt{RHj})_{j=\texttt{1.kitchen,2.living,3.laundry,4.office,5.bath,6.outside}}$ to see the variation of hourly-wise pattern for Mondays, while the other covariates fixed at their medians. Condition $h$ is colored by hours 1 to 24.

# Chapter 9

# Comparison of LM and GAM for energy consumption within a house

## 9.1 Evaluation and comparison based on model selection criterion AIC and $R^2_{adj}$

We fitted a pretty satisfying linear model and selected them with help of the adjusted coefficient of determination and the selection criteria AIC. We also computed the generalized additive model with the R-package `mgcv` for our data set on the energy consumption of a house. Now the question arises, if we really can simply compare the two different regression models to choose between LM and GAM.

In many statistical journals, the $R^2_{adj}$ is used as indicator to compare the two models. (c.f. in Candanedo et al. (2017), Khouloud et al. (2017) and Abeare (2009)). According to the manual of the package `mgcv` of Wood and Wood (2019), the adjusted coefficient of determination is defined as the proportion of variance explained. For this, the original variance and residual variance are determined with their unbiased estimator. The deviance $D$ of a model is defined as discussed in Section 4.3.4. The theoretical definition is $D = 2[l(\hat{\boldsymbol{\beta}}_{max}) - l(\hat{\boldsymbol{\beta}})]\phi$, where $l(\hat{\boldsymbol{\beta}}_{max})$ is the maximized likelihood of the saturated model, i.e. a model with one parameter for each data point so that the likelihood is maximized. $l(\hat{\boldsymbol{\beta}})$ is the maximized likelihood of the fitted model and $\phi$ the scale parameter. Due to these points, it is reasonable to say, that the deviance explained has a high similarity to the coefficient of determination, see the definitions of the coefficient of determination in Section 3.5.2. We have to be careful with comparing $R^2_{adj}$ of linear models and generalized additive models, but is a satisfying indicator of performance.

Furthermore, according to the GAM construction (c.f. the general GAM equation (4.6)), it is possible to combine the parametric and the non-parametric components in one model. For our models, GAMmain (8.1) and GAMinter (8.2), we established both components within the GAM estimation procedure. And since the GAM method selects the best model within the calculation also by AIC, we cannot argue against a comparison between both regression approach based on the selection criteria AIC. Nevertheless it is fundamental to only apply AIC, when checking nested models against each other, see Section 3.5.3. As we do use the same structure in both models, we will apply AIC as a

model selection criteria. This means that both full models contain all temperature and humidity, weekday and hour covariates and due to all arguments of the GAM construction, this indicates that our linear models are nested in generalized additive models. This is also supported with a statement in Guisan et al. (2002).

| | df | AIC | $R^2_{adj}$ |
|---|---|---|---|
| **Main Effect Model** | | | |
| Model 7 | 31.00 | 31572.14 | 0.328 |
| GAMmain | 175.33 | 28808.11 | 0.420 |
| **Interaction Model** | | | |
| LMinter | 109.00 | 29640.86 | 0.393 |
| GAMinter | 438.59 | 26189.10 | 0.500 |

Table 9.1: Comparing all the created Main and Interaction LM and GAM Models with model selection criterion AIC, degree of freedom (df) and the coefficient of determination $R^2_{adj}$.

Both, the full main and interaction models improved their adjusted coefficient of determination by approximately 0.1 using GAM, see Table 9.1. Looking at the values, the model GAMinter is providing the best $R^2_{adj}$ and AIC, i.e. the highest $R^2_{adj}$ and the smallest AIC, which corroborate the best fitted model for our energy consumption data set with appliances as the response variable.

## 9.2 Further model comparison

With parametric models one simply use the AIC to choose the best fit model. However, the non-parametric analogues can be obtained by trading-off the number of parameters. For more information on effective number of parameters, see Friedman et al. (2001) in Section 7.6.

A non-parametric AIC was issued on regression models by Hurvich et al. (1998). The results on the AIC usage were convincing, but a comparison of a parametric and non-parametric model was not directly verified. Thus, we want to search for a valid selection criteria for checking these models against each other. One solution is the parametricness index for regression models which was introduced in the Paper of Liu et al. (2011). The performance reached an asymptotic efficiency when the true model is infinite dimensional in the case of a non-parametric model. Though it all, the behavior of the parametricness index was simulated with real data and showed its usefulness in practice.

Another solution for model comparison between generalized linear and additive models is applied in Czado et al. (2009) the non-randomized probability integral transforms and proper scores. These methods allows to evaluate model fits and predictive power. In the paper these model selections were used on non-nested insurance models. It shows that GAM indeed provide a better model fit, but due to the computational cost, the generalized linear model is a appropriate and good alternative with non-linear components.

# Chapter 10

# Conclusion

The statistical data analysis has revealed interesting results in exploratory analysis and model setting. As seen in the time series plots, the appliances energy consumption profile is highly variable and in the box-plots it was obvious that the data above the median is higher dispersed. The pairwise plots show that the temperature and humidity have indeed effects on the appliances energy consumption. Temperature and humidity increases when a resident enters a room. If the occupant then uses devices in this area that also consequently increase the temperature, a dependence on the appliances is visible. The time information was ranked as an important covariate as the explanatory power of the model increased impressively. So the time, especially the hours, have an effect on the appliances energy consumption. Additionally, setting an interaction with the time effect hour improved the regression. It was clear from the predictions that the fitted lines of afternoon hours intersect with the lines of morning and evening hours, as there are interactions. These are the occupants showing daily routine. Changing the behavior in the afternoon hours will affect the evening and morning hours.

We have also seen interesting results when considering the LM and GAM for fitting the appliances energy consumption. To receive a satisfying model, we transformed our response variable appliances with the logarithm function and used linear relationships for the temperature and humidity covariates and non-linear relationship for the hour covariate. The weekday was included as an factor. This composition yield a $R^2_{adj}$ of more than 0.3 and with an included interaction effect even nearly 0.4. For comparison reasons, the same regression structure were used for GAM and reached an $R^2_{adj}$ improvement of 0.1. Even though the GAM did a better job in fitting the model with its non-parametric structure, we have seen in the GAM marginal plots that we fitted the linear model well with help of polynomial hourly effects. Also the predictions of the LM and GAM supported each other in their interpretation. With this nested structure we were able to use the adjusted coefficient of determination and the AIC for model selection. The GAM also provided a better AIC, but only with an enhancement of approximately 10%. Based on these results, one can say that the LM, with careful chosen non-linear covariate and interaction effect, is also satisfying regarding the simpler straightforward setting and the lower computational cost.

# Appendix A

# Additional supporting plots and tables

Further supporting figures and study results are listed here.

## A.1 Psychrometric chart

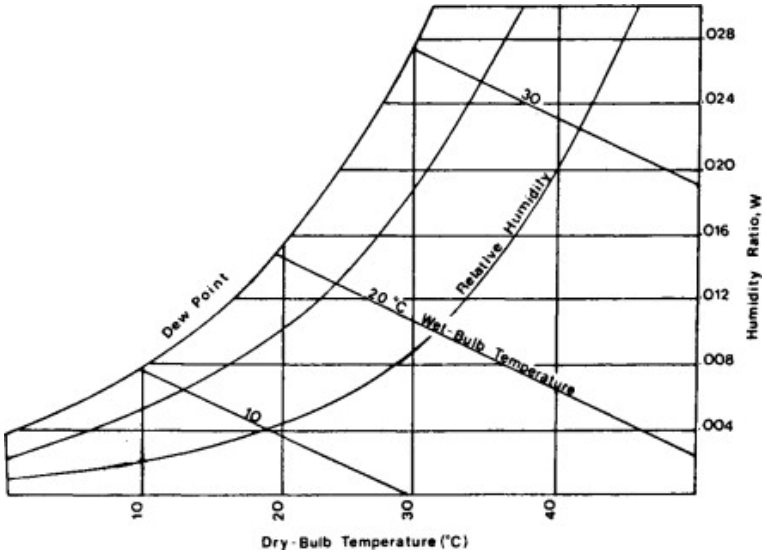



Figure A.1: The simplified psychrometric chart adopted from Swartman (1981). The figure shows the inter-relationship of humidity and temperature. For more insight on the theory of thermodynamics, see Sattelmayer (2008)

## A.2 Interaction plots

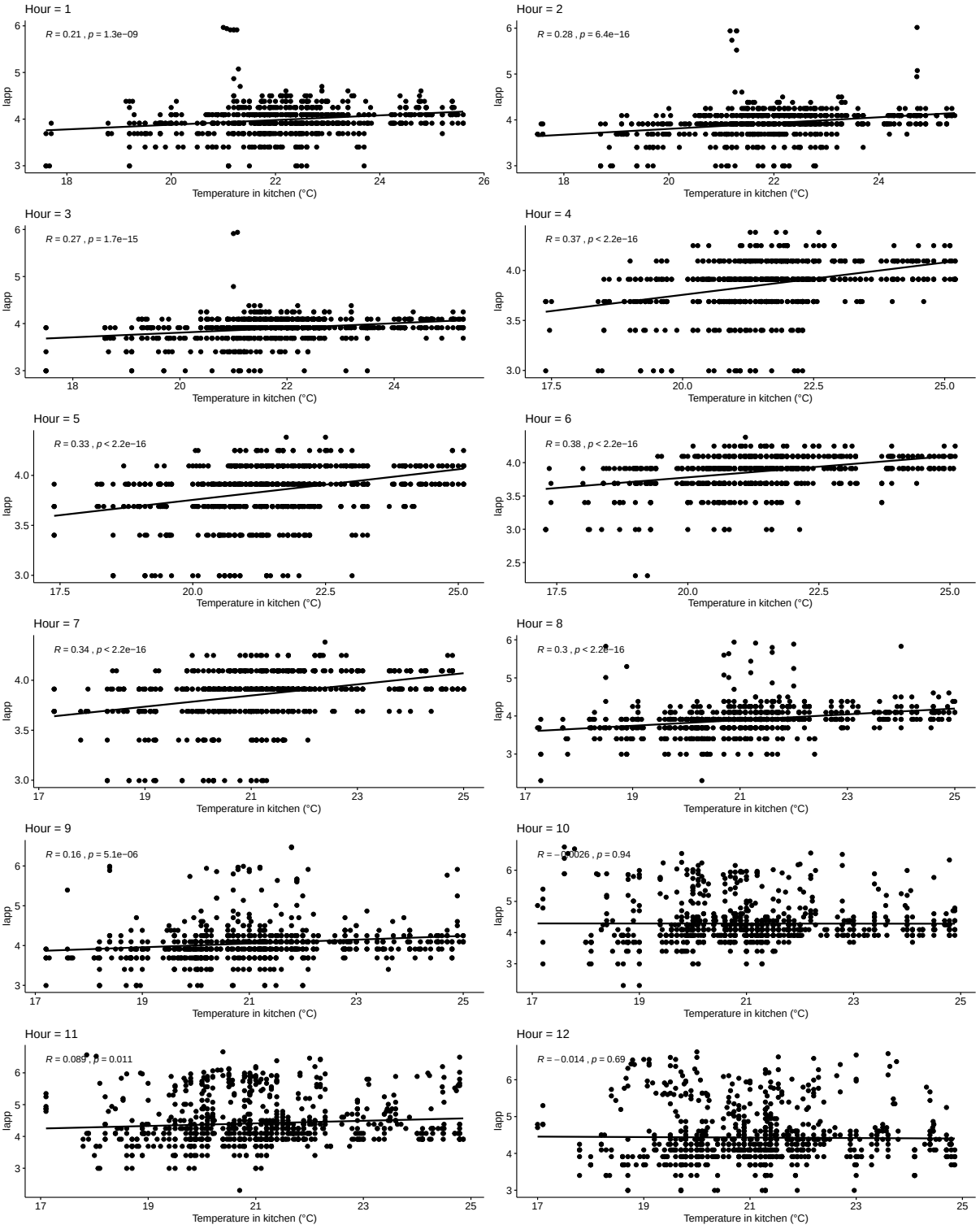



Figure A.2: Scatter-plots (1) of `T1.kitchen` separated by hour versus `lapp`, where $R$ denotes the correlation coefficient and $p$ the p-value.

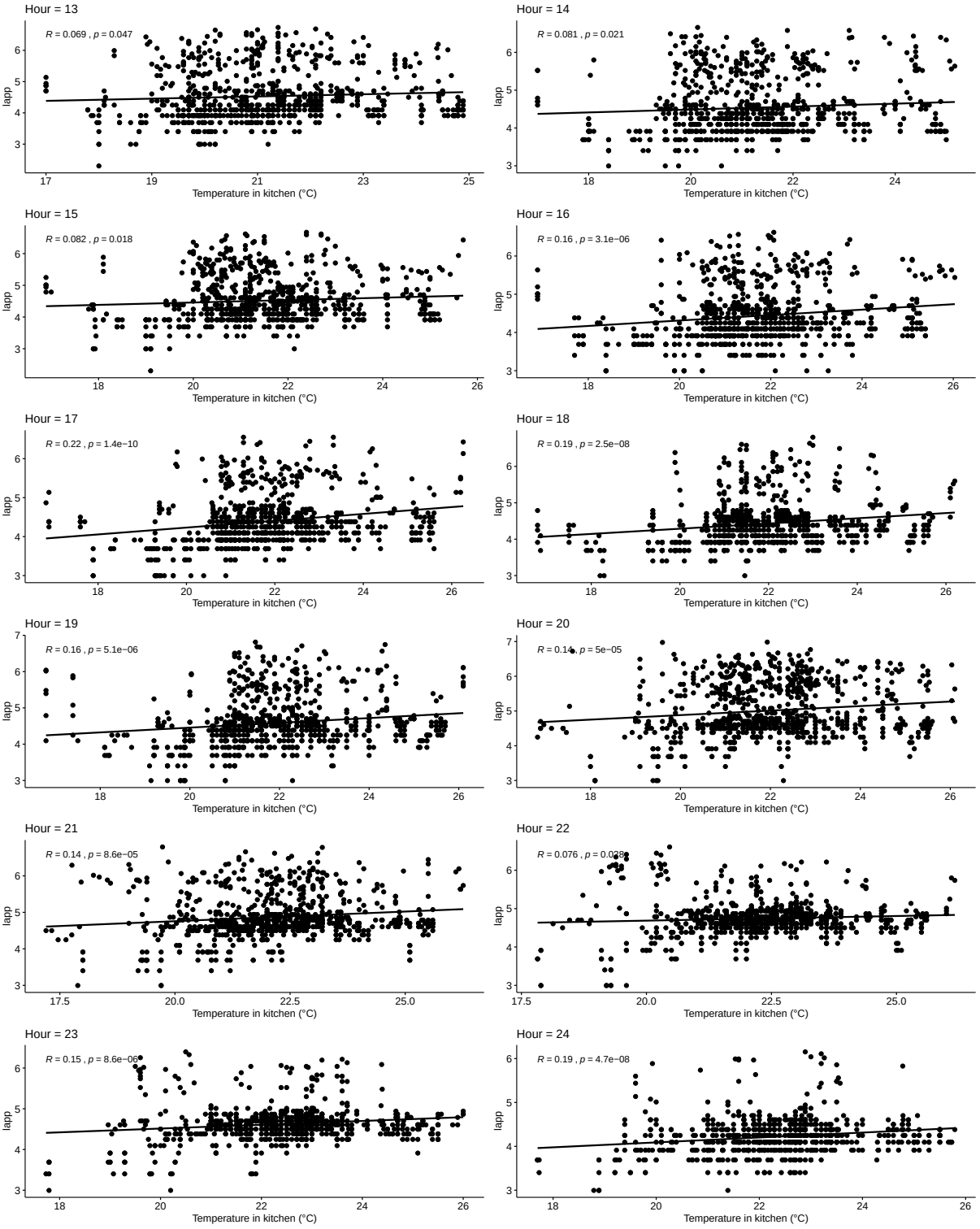



Figure A.3: Scatter-plots (2) of `T1.kitchen` separated by hour versus `lapp`, where $R$ denotes the correlation coefficient and $p$ the p-value.

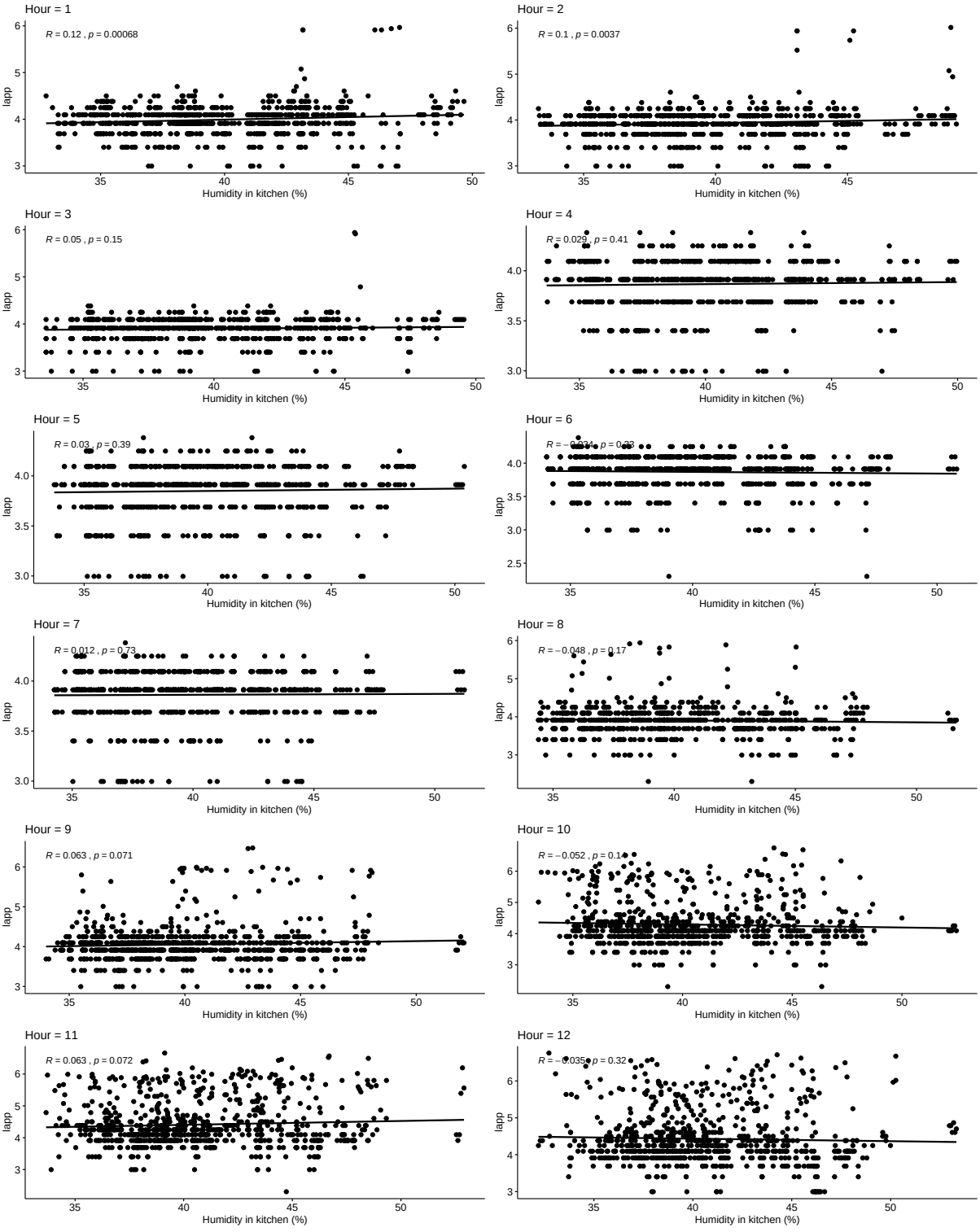



Figure A.4: Scatter-plots (1) of `RH1.kitchen` separated by hour versus `lapp`, where $R$ denotes the correlation coefficient and $p$ the p-value.

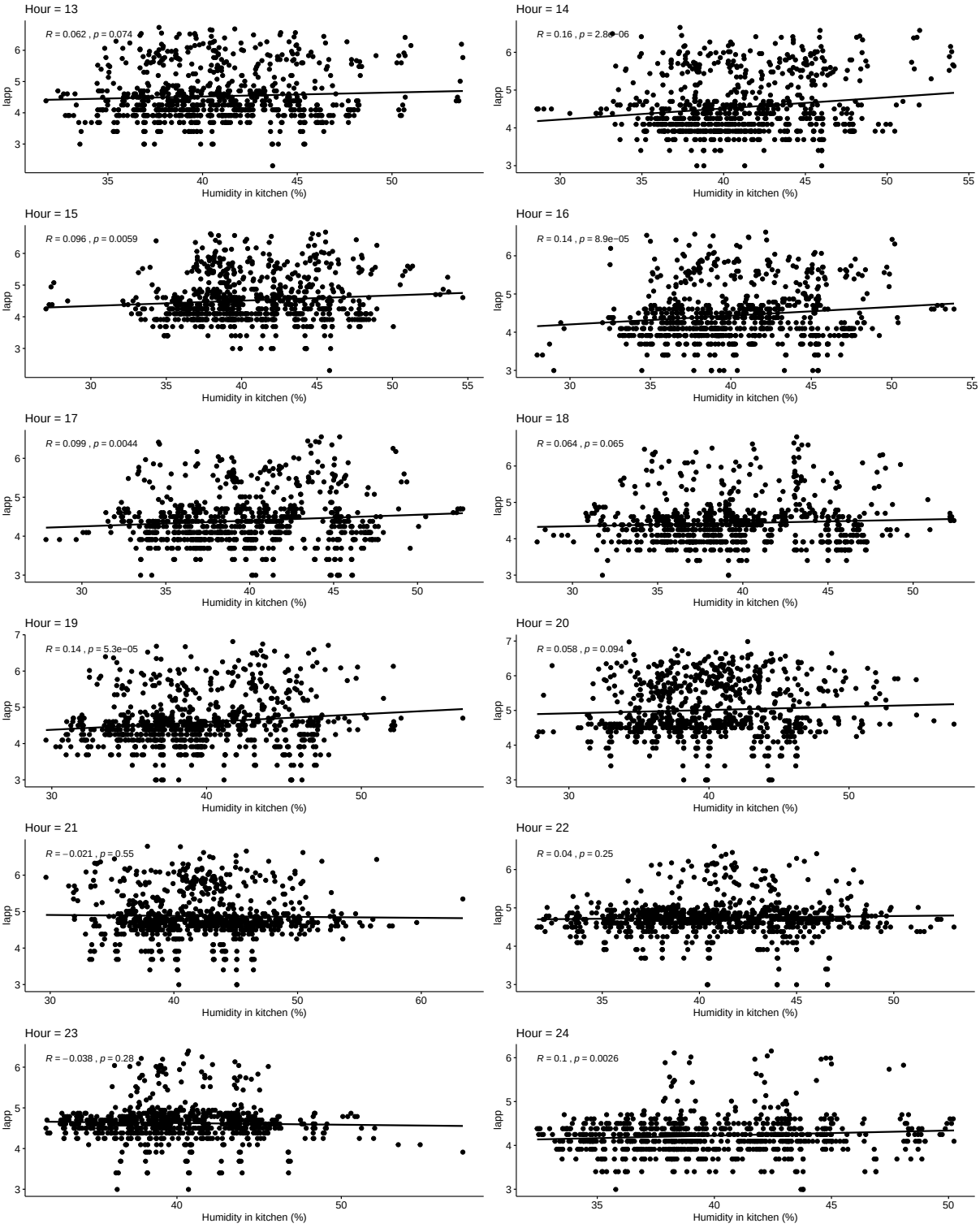



Figure A.5: Scatter-plots (2) of `RH1.kitchen` separated by hour versus `lapp`, where $R$ denotes the correlation coefficient and $p$ the p-value.

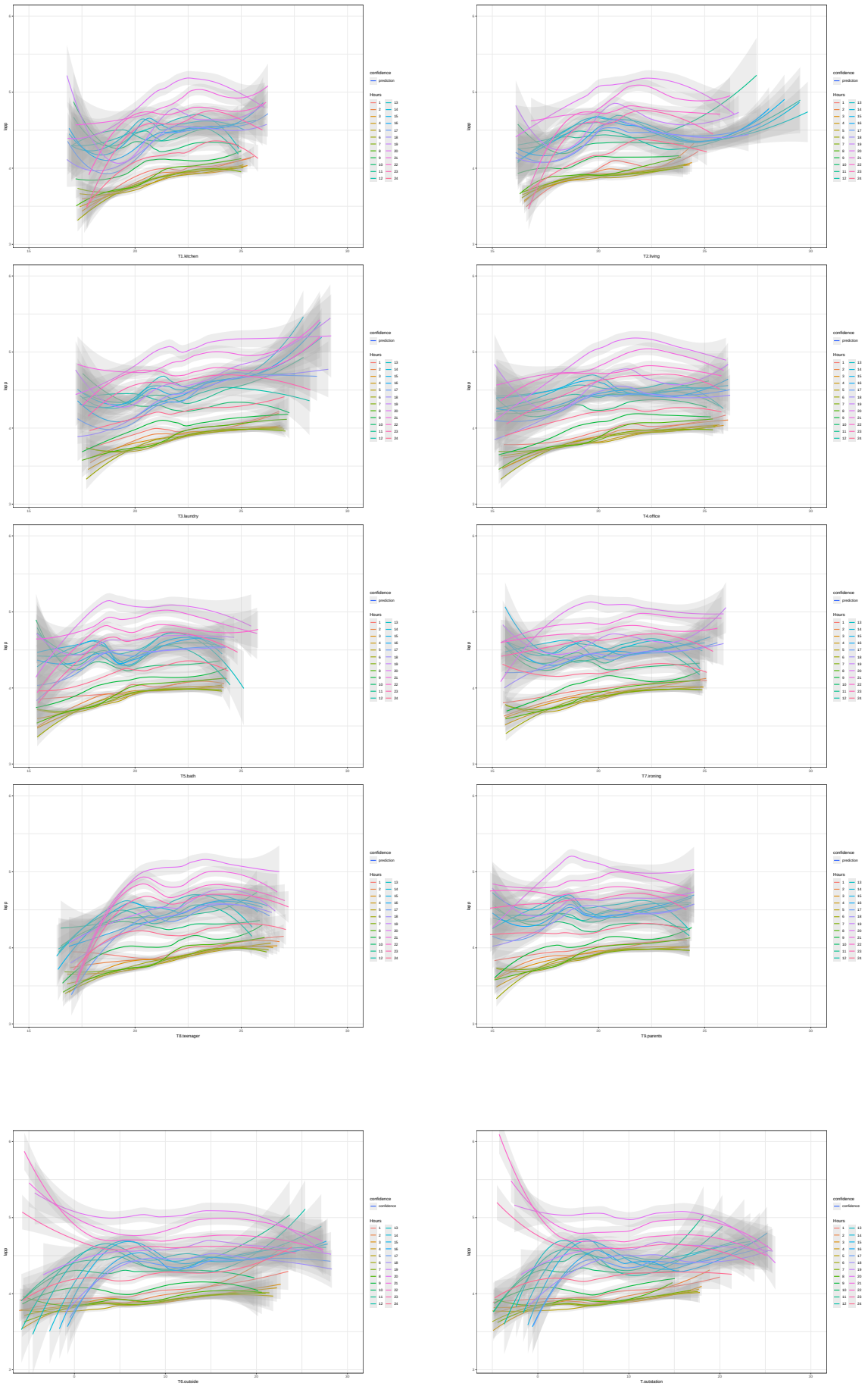


Figure A.6: Interaction plot of $\widehat{\texttt{lapp}}_{i_h} = \hat{\beta}_0 + \sum_{j=1}^{10} \hat{\beta}_j \texttt{Tj}_{i_h}$, $j = \texttt{1.kitchen}, \ldots, \texttt{9.parents}, \texttt{.outstation}$, separated by observation $i_h$ corresponds to the `hour`. With an additional 95% confidence interval marked as a gray shadow.

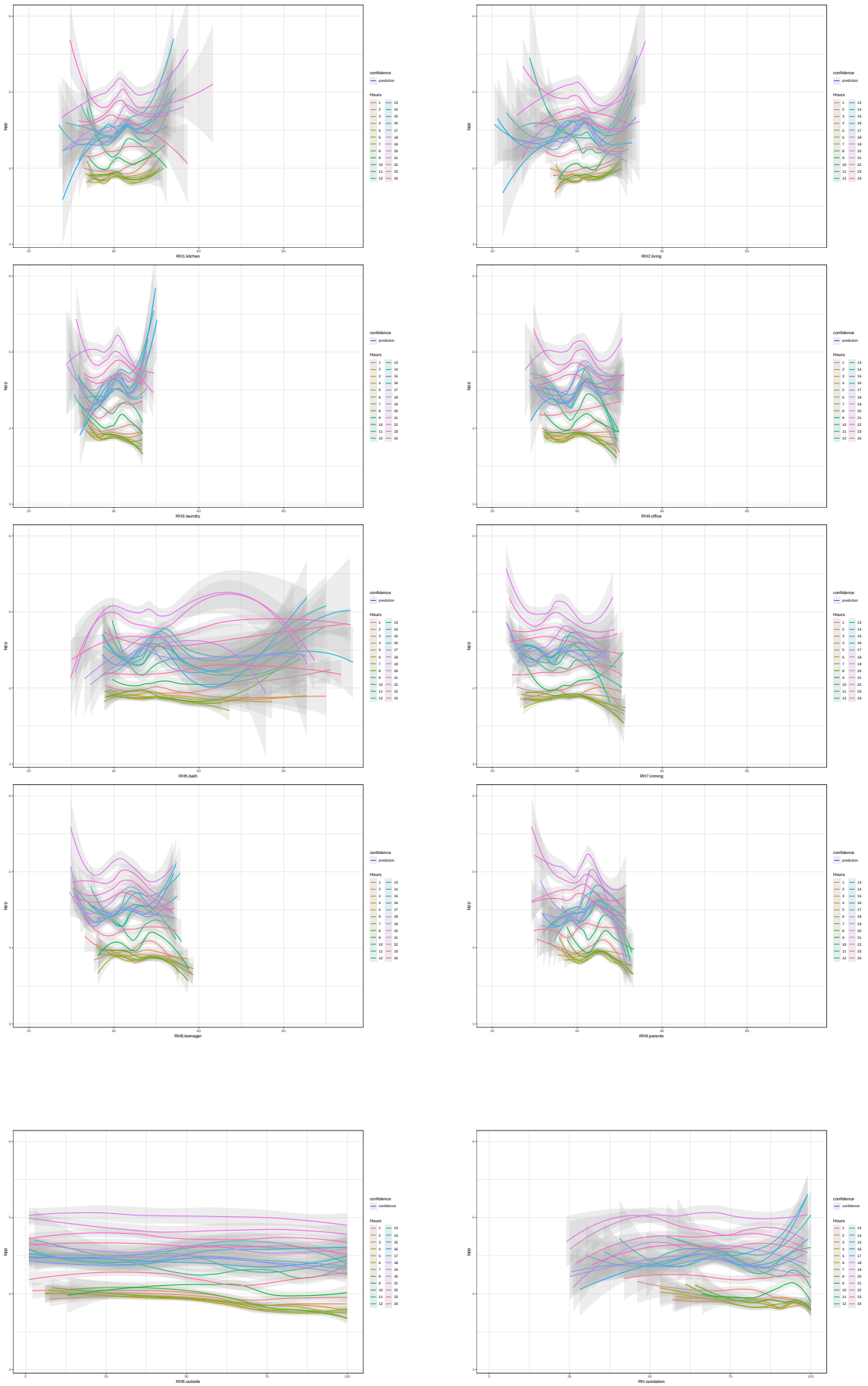



Figure A.7: Interaction plot of $\widehat{\texttt{lapp}}_{i_h} = \hat{\beta}_0 + \sum_{j=1}^{10} \hat{\beta}_j \texttt{RHj}_{i_h}$, $j = \texttt{1.kitchen}, \ldots, \texttt{9.parents}, \texttt{.outstation}$, separated by observation $i_h$ corresponds to the `hour`. With an additional 95% confidence interval marked as a gray shadow.

## A.3 Predictions

### A.3.1 Statistics

### A.3.2 Predictions using the LM's and GAM's

| hour | statistics | T1.kit | T2.liv | T3.laun | T4.off | T5.bath | T6.out | T7.iron | T8.teen | T9.par | T.outstation |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | min | 17.60 | 16.60 | 17.89 | 15.53 | 15.42 | -6.01 | 15.50 | 17.00 | 15.10 | -4.90 |
| | median | 22.20 | 20.29 | 22.00 | 21.08 | 19.79 | 5.95 | 20.16 | 22.73 | 19.29 | 6.22 |
| | mean | 22.22 | 20.51 | 22.28 | 21.16 | 19.94 | 6.29 | 20.30 | 22.59 | 19.39 | 6.70 |
| | max | 25.60 | 24.70 | 26.79 | 26.00 | 24.32 | 23.50 | 25.10 | 27.00 | 24.10 | 20.00 |
| 2 | min | 17.50 | 16.60 | 17.89 | 15.50 | 15.39 | -6.07 | 15.54 | 16.94 | 15.10 | -4.75 |
| | median | 22.00 | 19.89 | 22.03 | 21.07 | 19.63 | 5.80 | 20.04 | 22.50 | 19.29 | 6.16 |
| | mean | 22.04 | 20.20 | 22.25 | 21.10 | 19.81 | 5.99 | 20.27 | 22.45 | 19.41 | 6.30 |
| | max | 25.45 | 24.50 | 26.78 | 26.10 | 24.20 | 23.08 | 25.07 | 26.80 | 24.16 | 18.90 |
| 3 | min | 17.50 | 16.50 | 17.79 | 15.50 | 15.39 | -5.61 | 15.60 | 16.89 | 15.19 | -4.86 |
| | median | 21.79 | 19.58 | 22.10 | 20.89 | 19.50 | 5.67 | 20.00 | 22.39 | 19.29 | 5.55 |
| | mean | 21.86 | 19.91 | 22.25 | 20.96 | 19.69 | 5.72 | 20.24 | 22.29 | 19.42 | 5.96 |
| | max | 25.29 | 24.40 | 26.84 | 25.89 | 24.17 | 22.67 | 24.95 | 26.70 | 24.20 | 18.00 |
| 4 | min | 17.39 | 16.50 | 17.79 | 15.46 | 15.39 | -5.75 | 15.60 | 16.80 | 15.19 | -4.92 |
| | median | 21.60 | 19.32 | 22.10 | 20.67 | 19.39 | 5.40 | 20.00 | 22.27 | 19.39 | 5.33 |
| | mean | 21.70 | 19.69 | 22.27 | 20.79 | 19.61 | 5.49 | 20.23 | 22.16 | 19.44 | 5.73 |
| | max | 25.20 | 24.30 | 26.90 | 25.70 | 24.10 | 22.26 | 24.89 | 26.50 | 24.27 | 17.70 |
| 5 | min | 17.39 | 16.42 | 17.70 | 15.39 | 15.39 | -5.86 | 15.60 | 16.79 | 15.19 | -4.99 |
| | median | 21.43 | 19.10 | 22.20 | 20.39 | 19.29 | 5.08 | 19.98 | 22.10 | 19.39 | 5.10 |
| | mean | 21.55 | 19.48 | 22.28 | 20.65 | 19.54 | 5.22 | 20.20 | 22.05 | 19.46 | 5.47 |
| | max | 25.10 | 24.19 | 26.97 | 25.60 | 24.10 | 21.84 | 24.89 | 26.39 | 24.29 | 17.80 |
| 6 | min | 17.29 | 16.39 | 17.70 | 15.39 | 15.39 | -5.51 | 15.63 | 16.70 | 15.19 | -5.00 |
| | median | 21.29 | 18.96 | 22.20 | 20.23 | 19.20 | 4.56 | 19.89 | 22.00 | 19.39 | 4.97 |
| | mean | 21.40 | 19.30 | 22.28 | 20.53 | 19.47 | 5.00 | 20.17 | 21.95 | 19.47 | 5.20 |
| | max | 25.10 | 24.09 | 27.03 | 25.50 | 24.10 | 21.43 | 24.89 | 26.29 | 24.29 | 17.60 |
| 7 | min | 17.29 | 16.39 | 17.70 | 15.30 | 15.38 | -5.54 | 15.63 | 16.68 | 15.19 | -4.43 |
| | median | 21.20 | 18.79 | 22.20 | 20.10 | 19.10 | 4.52 | 19.82 | 21.96 | 19.39 | 4.84 |
| | mean | 21.27 | 19.14 | 22.26 | 20.43 | 19.42 | 4.87 | 20.13 | 21.85 | 19.48 | 5.00 |
| | max | 25.00 | 23.99 | 27.09 | 25.39 | 24.07 | 21.01 | 24.79 | 26.20 | 24.29 | 16.80 |
| 8 | min | 17.23 | 16.29 | 17.50 | 15.30 | 15.37 | -5.61 | 15.62 | 16.60 | 15.10 | -4.75 |
| | median | 21.10 | 18.70 | 22.10 | 20.00 | 19.15 | 4.40 | 19.79 | 21.79 | 19.39 | 4.80 |
| | mean | 21.16 | 19.00 | 22.22 | 20.34 | 19.37 | 4.77 | 20.10 | 21.75 | 19.48 | 4.88 |
| | max | 25.00 | 23.88 | 27.15 | 25.39 | 24.26 | 20.60 | 24.79 | 26.10 | 24.32 | 14.80 |
| 9 | min | 17.20 | 16.20 | 17.50 | 15.30 | 15.34 | -5.71 | 15.67 | 16.58 | 15.10 | -4.80 |
| | median | 21.00 | 18.60 | 22.00 | 19.95 | 19.20 | 4.73 | 19.79 | 21.70 | 19.39 | 4.82 |
| | mean | 21.08 | 18.91 | 22.11 | 20.27 | 19.35 | 4.87 | 20.07 | 21.65 | 19.48 | 4.85 |
| | max | 25.00 | 23.78 | 27.21 | 25.29 | 24.32 | 20.19 | 24.79 | 26.00 | 24.39 | 15.05 |
| 10 | min | 17.10 | 16.20 | 17.53 | 15.26 | 15.34 | -5.62 | 15.60 | 16.50 | 15.10 | -4.80 |
| | median | 21.00 | 18.50 | 21.83 | 20.03 | 19.13 | 5.59 | 19.63 | 21.58 | 19.39 | 5.02 |
| | mean | 21.04 | 18.96 | 22.01 | 20.35 | 19.35 | 5.53 | 19.99 | 21.52 | 19.46 | 5.15 |
| | max | 24.89 | 24.01 | 27.27 | 25.10 | 24.00 | 19.77 | 24.76 | 25.87 | 24.16 | 15.52 |
| 11 | min | 17.10 | 16.20 | 17.60 | 15.19 | 15.39 | -5.83 | 15.60 | 16.50 | 15.10 | -4.90 |
| | median | 21.00 | 19.00 | 21.79 | 20.29 | 19.20 | 6.51 | 19.50 | 21.39 | 19.39 | 5.70 |
| | mean | 21.04 | 19.54 | 21.98 | 20.54 | 19.36 | 6.98 | 19.93 | 21.40 | 19.52 | 6.00 |
| | max | 24.86 | 27.45 | 27.93 | 25.20 | 24.00 | 23.67 | 24.76 | 25.74 | 24.25 | 18.22 |
| 12 | min | 17.00 | 16.20 | 17.50 | 15.19 | 15.39 | -5.82 | 15.60 | 16.40 | 15.10 | -3.70 |
| | median | 21.00 | 19.73 | 21.79 | 20.50 | 19.20 | 8.11 | 19.50 | 21.36 | 19.39 | 6.50 |
| | mean | 21.09 | 20.38 | 21.95 | 20.73 | 19.39 | 8.45 | 19.90 | 21.32 | 19.55 | 7.07 |
| | max | 24.86 | 29.50 | 28.22 | 25.70 | 24.48 | 25.37 | 24.76 | 25.50 | 24.27 | 20.28 |
| 13 | min | 17.00 | 16.20 | 17.39 | 15.19 | 15.39 | -4.58 | 15.60 | 16.39 | 15.03 | -2.40 |
| | median | 21.03 | 20.29 | 21.79 | 20.60 | 19.20 | 9.20 | 19.60 | 21.39 | 19.39 | 7.60 |
| | mean | 21.20 | 20.92 | 22.01 | 20.90 | 19.45 | 9.65 | 19.94 | 21.33 | 19.52 | 8.10 |
| | max | 24.89 | 29.86 | 27.53 | 25.88 | 25.12 | 27.13 | 24.79 | 25.52 | 24.29 | 22.68 |
| 14 | min | 17.00 | 16.20 | 17.39 | 15.19 | 15.39 | -2.73 | 15.60 | 16.31 | 15.00 | -1.60 |
| | median | 21.20 | 20.43 | 22.00 | 20.89 | 19.22 | 9.80 | 19.89 | 21.54 | 19.43 | 8.46 |
| | mean | 21.34 | 21.14 | 22.17 | 21.02 | 19.43 | 10.59 | 20.06 | 21.45 | 19.53 | 8.95 |
| | max | 25.17 | 29.46 | 27.93 | 26.00 | 24.10 | 27.66 | 24.89 | 25.79 | 24.36 | 23.35 |
| 15 | min | 16.89 | 16.20 | 17.29 | 15.19 | 15.39 | -1.65 | 15.60 | 16.37 | 15.00 | -1.10 |
| | median | 21.41 | 20.50 | 22.10 | 20.89 | 19.26 | 10.29 | 20.10 | 21.70 | 19.46 | 9.10 |
| | mean | 21.55 | 21.24 | 22.31 | 21.04 | 19.42 | 11.24 | 20.22 | 21.65 | 19.54 | 9.55 |
| | max | 25.70 | 28.76 | 28.79 | 26.10 | 24.10 | 27.83 | 25.00 | 26.00 | 24.39 | 24.40 |
| 16 | min | 16.89 | 16.20 | 17.29 | 15.19 | 15.39 | -0.77 | 15.60 | 16.89 | 15.00 | -0.60 |
| | median | 21.62 | 20.60 | 22.20 | 20.85 | 19.27 | 10.67 | 20.29 | 21.96 | 19.50 | 9.50 |
| | mean | 21.73 | 21.27 | 22.38 | 21.07 | 19.44 | 11.67 | 20.41 | 21.85 | 19.57 | 9.99 |
| | max | 26.03 | 28.03 | 28.73 | 26.10 | 24.20 | 27.70 | 25.26 | 26.24 | 24.50 | 25.18 |
| 17 | min | 16.82 | 16.10 | 17.29 | 15.10 | 15.39 | -0.49 | 15.60 | 17.00 | 15.00 | -0.48 |
| | median | 21.78 | 20.79 | 22.29 | 20.89 | 19.29 | 10.60 | 20.59 | 22.39 | 19.50 | 9.80 |
| | mean | 21.88 | 21.27 | 22.40 | 21.09 | 19.48 | 11.80 | 20.61 | 22.04 | 19.59 | 10.25 |
| | max | 26.26 | 27.63 | 28.57 | 26.18 | 24.23 | 28.14 | 25.60 | 26.33 | 24.50 | 25.30 |
| 18 | min | 16.79 | 16.10 | 17.29 | 15.10 | 15.36 | -1.02 | 15.56 | 16.96 | 15.00 | -1.08 |
| | median | 21.79 | 20.79 | 22.30 | 20.89 | 19.29 | 10.50 | 20.68 | 22.60 | 19.46 | 9.73 |
| | mean | 21.96 | 21.23 | 22.42 | 21.07 | 19.52 | 11.71 | 20.72 | 22.21 | 19.59 | 10.25 |
| | max | 26.20 | 27.07 | 29.10 | 26.20 | 24.29 | 28.29 | 25.89 | 26.44 | 24.50 | 25.27 |
| 19 | min | 16.79 | 16.10 | 17.20 | 15.10 | 15.36 | -3.42 | 15.49 | 16.89 | 15.00 | -1.87 |
| | median | 21.79 | 20.70 | 22.29 | 20.89 | 19.29 | 9.96 | 20.63 | 22.60 | 19.39 | 9.35 |
| | mean | 21.97 | 21.12 | 22.41 | 21.03 | 19.53 | 11.20 | 20.68 | 22.30 | 19.56 | 10.02 |
| | max | 26.10 | 26.60 | 29.20 | 26.10 | 24.67 | 28.20 | 26.00 | 26.60 | 24.50 | 25.82 |
| 20 | min | 16.82 | 16.10 | 17.20 | 15.10 | 15.33 | -4.37 | 15.39 | 17.28 | 15.00 | -2.58 |
| | median | 21.96 | 20.75 | 22.29 | 20.89 | 19.29 | 9.46 | 20.46 | 22.70 | 19.43 | 8.78 |
| | mean | 22.03 | 21.03 | 22.40 | 21.04 | 19.59 | 10.48 | 20.62 | 22.44 | 19.53 | 9.64 |
| | max | 26.10 | 26.39 | 29.24 | 26.00 | 25.47 | 27.35 | 25.89 | 26.79 | 24.50 | 26.10 |
| 21 | min | 17.20 | 16.82 | 17.29 | 15.19 | 15.35 | -4.98 | 15.39 | 17.29 | 15.00 | -2.95 |
| | median | 22.20 | 20.79 | 22.29 | 20.79 | 19.39 | 8.73 | 20.29 | 22.70 | 19.39 | 8.23 |
| | mean | 22.24 | 21.06 | 22.46 | 20.98 | 19.67 | 9.46 | 20.51 | 22.53 | 19.47 | 9.10 |
| | max | 26.26 | 26.20 | 28.70 | 25.89 | 25.75 | 27.26 | 25.79 | 26.70 | 24.33 | 25.70 |
| 22 | min | 17.82 | 16.79 | 17.50 | 15.39 | 15.39 | -5.51 | 15.40 | 17.26 | 14.89 | -4.25 |
| | median | 22.39 | 20.73 | 22.31 | 20.89 | 19.91 | 8.03 | 20.25 | 22.76 | 19.37 | 7.88 |
| | mean | 22.36 | 21.07 | 22.51 | 21.07 | 20.02 | 8.61 | 20.43 | 22.61 | 19.44 | 8.57 |
| | max | 26.17 | 25.73 | 28.70 | 25.79 | 25.80 | 27.37 | 25.60 | 27.05 | 24.20 | 25.10 |
| 23 | min | 17.73 | 16.70 | 17.79 | 15.66 | 15.44 | -5.74 | 15.42 | 17.19 | 14.89 | -4.50 |
| | median | 22.50 | 20.79 | 22.29 | 21.10 | 20.16 | 7.38 | 20.20 | 22.79 | 19.29 | 7.32 |
| | mean | 22.41 | 20.99 | 22.45 | 21.20 | 20.24 | 7.48 | 20.36 | 22.65 | 19.39 | 7.90 |
| | max | 26.00 | 25.39 | 28.26 | 25.70 | 24.83 | 24.33 | 25.39 | 27.23 | 24.14 | 23.80 |
| 24 | min | 17.70 | 16.70 | 17.86 | 15.60 | 15.48 | -5.69 | 15.48 | 17.10 | 14.89 | -4.75 |
| | median | 22.39 | 20.60 | 22.10 | 21.08 | 20.04 | 6.33 | 20.20 | 22.79 | 19.25 | 6.81 |
| | mean | 22.36 | 20.80 | 22.36 | 21.17 | 20.13 | 6.76 | 20.31 | 22.64 | 19.36 | 7.23 |
| | max | 25.79 | 25.10 | 27.03 | 25.79 | 24.39 | 23.91 | 25.10 | 27.10 | 24.04 | 21.30 |

Table A.1: Statistics of all temperature variables.

| hour | statistics | RH1.kit | RH2.liv | RH3.laun | RH4.off | RH5.bath | RH6.out | RH7.iron | RH8.teen | RH9.par | RH.outstation |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | min | 32.80 | 33.63 | 34.05 | 31.67 | 37.90 | 2.05 | 25.76 | 33.13 | 30.50 | 46.00 |
| | median | 39.02 | 40.59 | 38.59 | 38.70 | 51.67 | 54.51 | 35.43 | 42.36 | 39.66 | 85.58 |
| | mean | 39.91 | 40.64 | 39.35 | 39.25 | 54.29 | 58.11 | 35.63 | 43.28 | 40.66 | 83.24 |
| | max | 49.66 | 50.59 | 46.56 | 49.90 | 91.00 | 99.90 | 49.96 | 55.25 | 51.59 | 100.00 |
| 2 | min | 33.29 | 34.30 | 34.47 | 31.89 | 38.17 | 5.93 | 26.50 | 35.33 | 33.18 | 57.67 |
| | median | 39.25 | 40.90 | 38.76 | 38.73 | 51.40 | 58.32 | 36.01 | 43.68 | 40.47 | 86.42 |
| | mean | 39.95 | 40.98 | 39.43 | 39.27 | 52.76 | 60.03 | 36.08 | 44.38 | 41.40 | 85.05 |
| | max | 49.13 | 50.59 | 46.50 | 49.70 | 90.01 | 99.90 | 50.44 | 56.34 | 52.06 | 100.00 |
| 3 | min | 33.59 | 34.66 | 34.12 | 32.09 | 38.70 | 7.44 | 26.82 | 35.79 | 35.40 | 57.00 |
| | median | 39.23 | 41.09 | 38.83 | 38.59 | 50.92 | 58.88 | 36.46 | 44.40 | 41.16 | 88.00 |
| | mean | 39.95 | 41.21 | 39.50 | 39.19 | 51.78 | 61.75 | 36.37 | 45.22 | 42.06 | 86.41 |
| | max | 49.55 | 50.59 | 46.56 | 49.50 | 85.35 | 99.90 | 50.28 | 57.31 | 52.00 | 100.00 |
| 4 | min | 33.70 | 34.90 | 34.00 | 32.16 | 38.70 | 6.00 | 27.13 | 35.79 | 35.73 | 58.00 |
| | median | 39.40 | 41.29 | 38.90 | 38.50 | 50.52 | 61.28 | 36.47 | 44.95 | 41.86 | 88.83 |
| | mean | 39.96 | 41.43 | 39.58 | 39.15 | 51.12 | 63.12 | 36.57 | 45.74 | 42.60 | 87.53 |
| | max | 49.97 | 50.40 | 46.50 | 49.36 | 77.27 | 99.90 | 50.58 | 58.67 | 52.30 | 100.00 |
| 5 | min | 33.79 | 35.00 | 33.50 | 32.20 | 38.09 | 6.21 | 27.29 | 36.20 | 36.36 | 53.17 |
| | median | 39.40 | 41.42 | 39.06 | 38.50 | 50.19 | 63.59 | 36.29 | 44.85 | 42.33 | 89.25 |
| | mean | 39.94 | 41.63 | 39.64 | 39.12 | 50.64 | 64.23 | 36.68 | 45.88 | 43.04 | 88.30 |
| | max | 50.38 | 50.26 | 46.63 | 49.29 | 71.81 | 99.90 | 51.15 | 58.78 | 52.73 | 100.00 |
| 6 | min | 34.03 | 35.29 | 33.33 | 32.40 | 37.67 | 7.00 | 27.50 | 36.20 | 36.97 | 53.00 |
| | median | 39.42 | 41.50 | 39.13 | 38.42 | 50.00 | 64.56 | 36.53 | 44.56 | 42.65 | 90.50 |
| | mean | 39.94 | 41.78 | 39.70 | 39.10 | 50.28 | 65.30 | 36.74 | 45.72 | 43.38 | 89.06 |
| | max | 50.80 | 50.09 | 46.70 | 49.29 | 67.24 | 99.90 | 51.40 | 58.51 | 52.90 | 99.83 |
| 7 | min | 34.26 | 35.50 | 32.80 | 32.44 | 37.75 | 8.02 | 27.67 | 36.40 | 37.59 | 61.00 |
| | median | 39.50 | 41.59 | 39.11 | 38.29 | 49.90 | 65.14 | 36.78 | 44.48 | 42.90 | 91.33 |
| | mean | 39.94 | 41.93 | 39.74 | 39.09 | 49.99 | 66.28 | 36.79 | 45.50 | 43.63 | 89.82 |
| | max | 51.22 | 49.90 | 46.76 | 49.23 | 65.26 | 99.90 | 51.20 | 57.75 | 52.90 | 100.00 |
| 8 | min | 34.40 | 35.70 | 32.29 | 32.50 | 37.97 | 13.82 | 27.70 | 36.40 | 37.63 | 64.00 |
| | median | 39.59 | 41.83 | 39.00 | 38.20 | 49.611 | 65.28 | 36.90 | 44.50 | 43.09 | 91.92 |
| | mean | 40.08 | 42.21 | 39.69 | 39.09 | 49.81 | 66.78 | 36.84 | 45.29 | 43.82 | 90.30 |
| | max | 51.64 | 50.59 | 46.86 | 49.20 | 84.90 | 99.90 | 51.05 | 57.44 | 53.22 | 100.00 |
| 9 | min | 34.00 | 35.83 | 31.96 | 32.23 | 39.50 | 12.93 | 26.76 | 36.29 | 37.53 | 66.17 |
| | median | 39.62 | 42.20 | 38.81 | 38.29 | 49.00 | 65.53 | 36.91 | 44.10 | 43.09 | 92.00 |
| | mean | 40.28 | 42.48 | 39.51 | 39.16 | 49.88 | 67.31 | 36.80 | 45.05 | 43.78 | 90.51 |
| | max | 52.06 | 51.04 | 46.86 | 49.59 | 83.97 | 99.90 | 50.85 | 57.65 | 53.33 | 100.00 |
| 10 | min | 33.43 | 32.95 | 31.52 | 32.09 | 39.59 | 2.69 | 26.29 | 36.00 | 36.33 | 58.50 |
| | median | 39.76 | 42.36 | 38.47 | 38.47 | 49.00 | 66.75 | 35.68 | 43.33 | 42.65 | 91.67 |
| | mean | 40.50 | 42.50 | 39.24 | 39.22 | 49.94 | 67.16 | 36.29 | 44.15 | 43.15 | 89.81 |
| | max | 52.48 | 51.67 | 46.79 | 49.86 | 83.42 | 99.90 | 50.53 | 55.93 | 52.90 | 100.00 |
| 11 | min | 33.67 | 28.80 | 31.72 | 31.57 | 38.78 | 1.00 | 26.20 | 35.20 | 34.50 | 55.17 |
| | median | 39.90 | 41.78 | 38.29 | 38.58 | 48.61 | 63.02 | 35.923 | 42.62 | 41.48 | 89.00 |
| | mean | 40.69 | 41.81 | 39.04 | 39.20 | 49.59 | 63.25 | 36.01 | 43.40 | 42.18 | 87.11 |
| | max | 52.90 | 52.30 | 46.90 | 48.97 | 82.33 | 99.90 | 49.12 | 54.57 | 52.47 | 100.00 |
| 12 | min | 32.20 | 25.60 | 33.02 | 31.50 | 37.25 | 1.00 | 26.20 | 34.53 | 33.19 | 48.83 |
| | median | 40.11 | 40.82 | 38.28 | 38.73 | 48.46 | 56.77 | 35.50 | 42.15 | 40.80 | 84.00 |
| | mean | 40.72 | 40.59 | 38.92 | 39.21 | 50.38 | 57.33 | 35.75 | 42.97 | 41.68 | 82.68 |
| | max | 53.32 | 52.93 | 46.89 | 49.23 | 89.60 | 99.90 | 47.68 | 54.46 | 51.63 | 100.00 |
| 13 | min | 31.73 | 23.37 | 32.79 | 30.97 | 37.58 | 1.00 | 25.89 | 33.76 | 33.55 | 40.50 |
| | median | 40.43 | 39.96 | 38.40 | 38.59 | 48.74 | 49.24 | 35.09 | 41.40 | 40.58 | 79.00 |
| | mean | 40.87 | 39.74 | 39.09 | 39.19 | 51.38 | 52.01 | 35.47 | 42.48 | 41.47 | 77.60 |
| | max | 53.74 | 53.56 | 47.99 | 49.48 | 90.00 | 99.90 | 46.67 | 54.94 | 51.17 | 100.00 |
| 14 | min | 28.59 | 20.46 | 31.67 | 28.89 | 37.50 | 1.00 | 25.50 | 32.13 | 33.09 | 35.67 |
| | median | 40.33 | 39.48 | 38.75 | 38.53 | 48.40 | 41.16 | 34.43 | 40.67 | 40.40 | 73.67 |
| | mean | 40.90 | 39.16 | 39.26 | 39.16 | 50.75 | 47.44 | 35.09 | 41.89 | 41.27 | 73.26 |
| | max | 54.09 | 54.09 | 49.36 | 49.86 | 96.32 | 99.90 | 46.93 | 55.63 | 51.40 | 99.00 |
| 15 | min | 27.02 | 21.23 | 30.66 | 28.79 | 37.38 | 1.00 | 25.07 | 31.67 | 31.97 | 30.00 |
| | median | 39.68 | 39.01 | 38.70 | 38.23 | 47.93 | 35.47 | 33.96 | 40.20 | 40.00 | 69.33 |
| | mean | 40.58 | 38.85 | 39.28 | 38.89 | 49.89 | 43.80 | 34.71 | 41.35 | 41.07 | 70.12 |
| | max | 54.67 | 53.83 | 49.80 | 49.83 | 95.61 | 99.90 | 47.40 | 54.70 | 51.50 | 99.00 |
| 16 | min | 27.96 | 22.46 | 32.00 | 29.03 | 37.29 | 1.00 | 24.60 | 31.03 | 31.77 | 28.17 |
| | median | 39.54 | 38.52 | 38.36 | 37.94 | 47.52 | 32.30 | 33.79 | 40.04 | 39.90 | 66.25 |
| | mean | 40.12 | 38.58 | 39.09 | 38.72 | 49.14 | 41.74 | 34.39 | 41.00 | 40.98 | 67.80 |
| | max | 53.87 | 52.93 | 50.16 | 50.40 | 85.46 | 99.90 | 47.59 | 54.09 | 51.47 | 99.00 |
| 17 | min | 27.86 | 24.57 | 32.43 | 29.44 | 37.06 | 1.00 | 24.17 | 30.46 | 31.32 | 25.50 |
| | median | 39.42 | 38.20 | 38.03 | 38.13 | 47.05 | 28.30 | 33.37 | 39.66 | 39.85 | 65.58 |
| | mean | 39.72 | 38.34 | 38.83 | 38.57 | 48.34 | 40.54 | 34.03 | 40.59 | 40.89 | 66.35 |
| | max | 52.7 | 52.07 | 47.8 | 50.29 | 85.40 | 99.90 | 47.50 | 53.17 | 51.40 | 97.83 |
| 18 | min | 27.93 | 24.90 | 29.49 | 28.72 | 34.50 | 1.00 | 23.67 | 29.82 | 31.72 | 25.00 |
| | median | 39.00 | 37.78 | 37.93 | 37.90 | 46.31 | 26.61 | 32.92 | 39.36 | 39.90 | 65.67 |
| | mean | 39.45 | 38.21 | 38.61 | 38.42 | 47.52 | 40.33 | 33.76 | 40.41 | 40.71 | 66.32 |
| | max | 52.40 | 51.73 | 47.69 | 50.66 | 82.83 | 99.90 | 47.72 | 52.94 | 51.09 | 98.00 |
| 19 | min | 29.62 | 26.95 | 29.00 | 29.10 | 33.10 | 1.00 | 23.20 | 29.60 | 31.29 | 26.17 |
| | median | 39.00 | 37.98 | 37.79 | 37.93 | 46.10 | 31.06 | 32.65 | 39.64 | 39.79 | 68.08 |
| | mean | 39.58 | 38.34 | 38.59 | 38.50 | 46.96 | 41.12 | 33.73 | 40.53 | 40.52 | 67.95 |
| | max | 56.56 | 54.66 | 47.79 | 50.06 | 75.79 | 99.90 | 48.09 | 54.00 | 51.29 | 98.67 |
| 20 | min | 27.73 | 25.76 | 28.77 | 27.66 | 30.89 | 1.00 | 23.23 | 29.79 | 29.93 | 25.17 |
| | median | 39.75 | 38.70 | 37.74 | 38.02 | 46.03 | 35.25 | 33.34 | 39.72 | 39.59 | 70.17 |
| | mean | 40.19 | 38.84 | 38.61 | 38.65 | 47.13 | 42.71 | 33.91 | 40.85 | 40.25 | 69.88 |
| | max | 57.50 | 56.03 | 46.73 | 50.63 | 85.47 | 99.90 | 48.42 | 53.85 | 51.54 | 99.00 |
| 21 | min | 29.66 | 27.20 | 31.16 | 29.67 | 29.82 | 1.00 | 23.93 | 30.00 | 29.23 | 24.00 |
| | median | 42.19 | 40.23 | 38.26 | 38.33 | 47.23 | 37.95 | 33.38 | 39.71 | 39.21 | 74.00 |
| | mean | 42.27 | 40.17 | 39.24 | 39.05 | 48.80 | 45.30 | 34.03 | 40.84 | 39.96 | 71.97 |
| | max | 63.36 | 53.09 | 49.56 | 51.00 | 87.27 | 99.90 | 48.76 | 53.90 | 51.50 | 98.00 |
| 22 | min | 31.65 | 29.96 | 32.93 | 29.66 | 30.03 | 1.00 | 24.07 | 30.36 | 29.17 | 27.00 |
| | median | 40.46 | 40.19 | 38.50 | 38.63 | 51.55 | 43.35 | 33.57 | 39.95 | 39.00 | 76.83 |
| | mean | 40.75 | 40.21 | 39.39 | 39.30 | 56.50 | 48.20 | 34.21 | 41.01 | 39.73 | 74.37 |
| | max | 53.10 | 52.22 | 49.22 | 51.09 | 95.80 | 99.90 | 49.55 | 53.75 | 51.15 | 97.50 |
| 23 | min | 32.03 | 31.20 | 33.40 | 30.57 | 37.59 | 1.00 | 24.07 | 30.53 | 29.29 | 32.00 |
| | median | 39.51 | 40.01 | 38.40 | 38.50 | 53.49 | 48.53 | 33.81 | 39.90 | 38.56 | 80.50 |
| | mean | 40.11 | 40.20 | 39.25 | 39.10 | 59.43 | 51.73 | 34.51 | 41.15 | 39.45 | 77.79 |
| | max | 57.40 | 51.43 | 46.83 | 50.96 | 94.99 | 99.90 | 50.38 | 54.30 | 51.00 | 98.83 |
| 24 | min | 32.43 | 32.49 | 33.70 | 31.07 | 37.76 | 1 | 24.60 | 32.83 | 29.76 | 42.00 |
| | median | 39.06 | 40.16 | 38.50 | 38.50 | 52.25 | 52.05 | 34.47 | 40.93 | 38.83 | 83.83 |
| | mean | 39.84 | 40.29 | 39.25 | 39.04 | 56.56 | 55.30 | 34.94 | 41.86 | 39.61 | 80.98 |
| | max | 50.23 | 50.70 | 46.79 | 50.20 | 93.55 | 99.90 | 50.40 | 54.88 | 51.50 | 99.67 |

Table A.2: Statistics of all humidity variables.

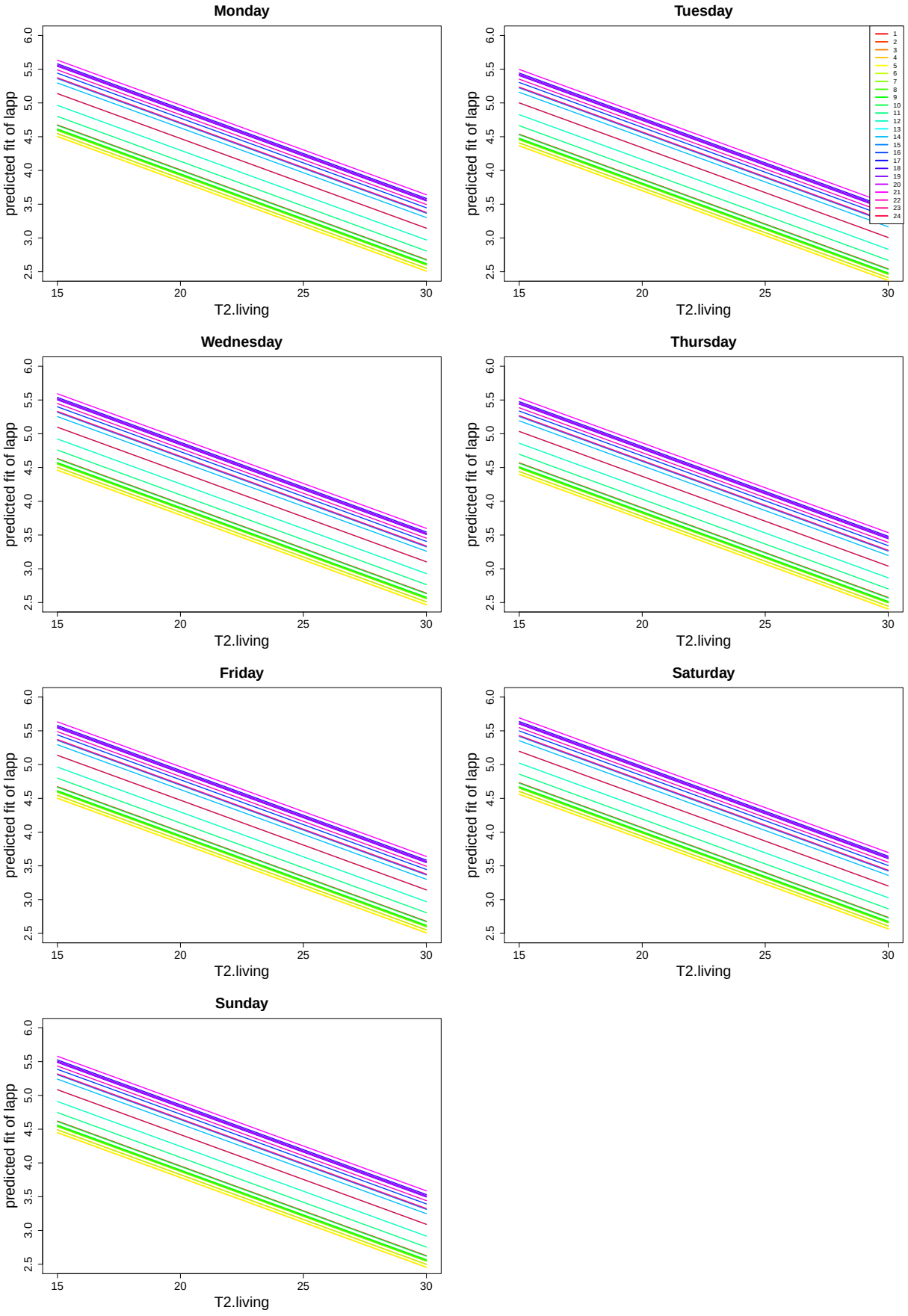



Figure A.8: Prediction of the full main model (Model 7) restricted on weekdays. The plot shows the relationship between fitted $(\widehat{\texttt{lapp}}_{i_h})_{i_h=(i\in[19735]|\texttt{hour}=h,\text{with}h=1,\ldots,24)}$ and `T2.living` to see the variation of hourly-wise pattern for all the weekdays, while the other covariates fixed at their medians. Condition $h$ is colored by hours 1 to 24.

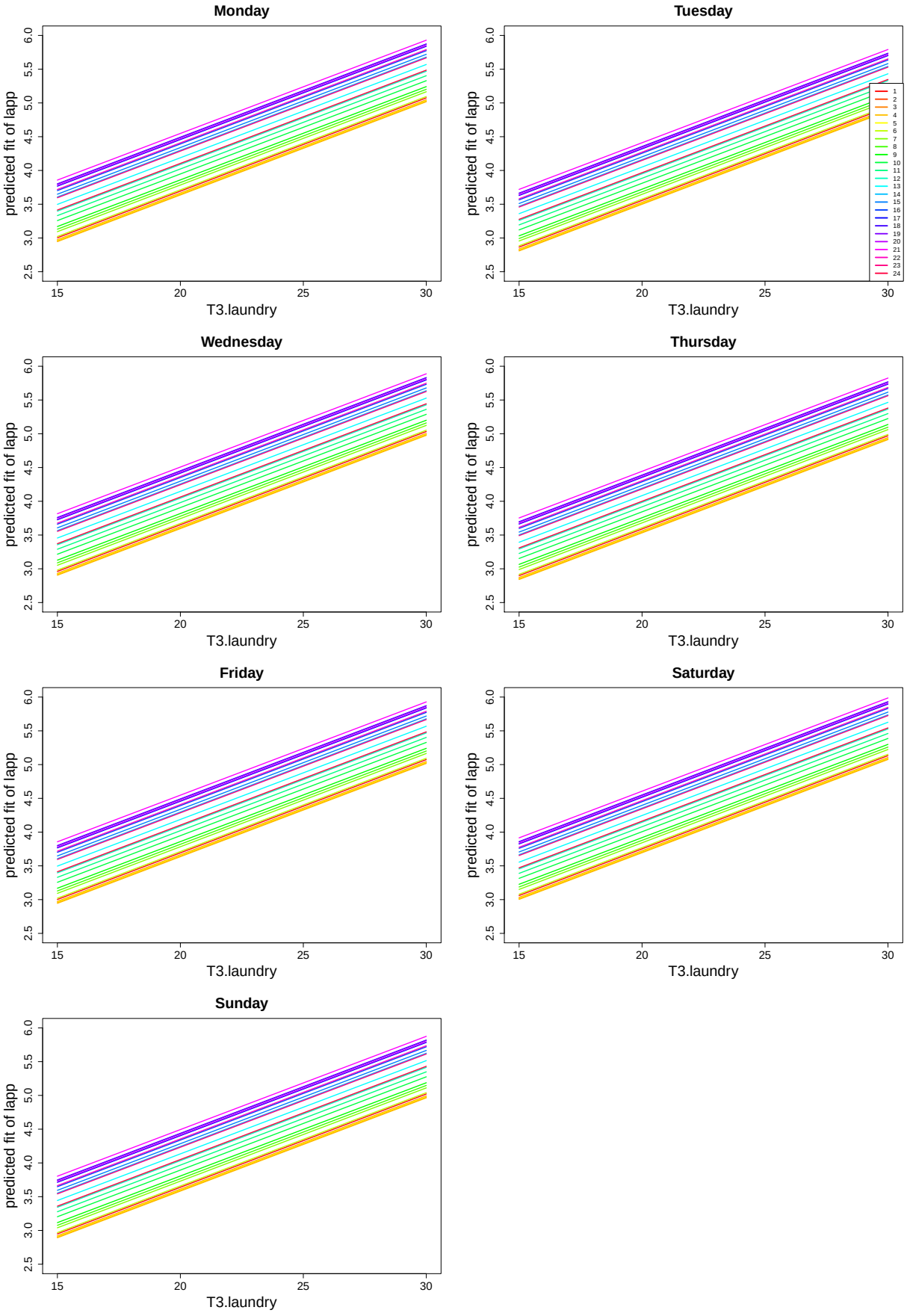



Figure A.9: Prediction of the full main model (Model 7) restricted on weekdays. The plot shows the relationship between fitted $(\widehat{\texttt{lapp}}_{i_h})_{i_h=(i\in[19735]|\texttt{hour}=h,\text{with}h=1,\ldots,24)}$ and `T3.laundry` to see the variation of hourly-wise pattern for all the weekdays, while the other covariates fixed at their medians. Condition $h$ is colored by hours 1 to 24.

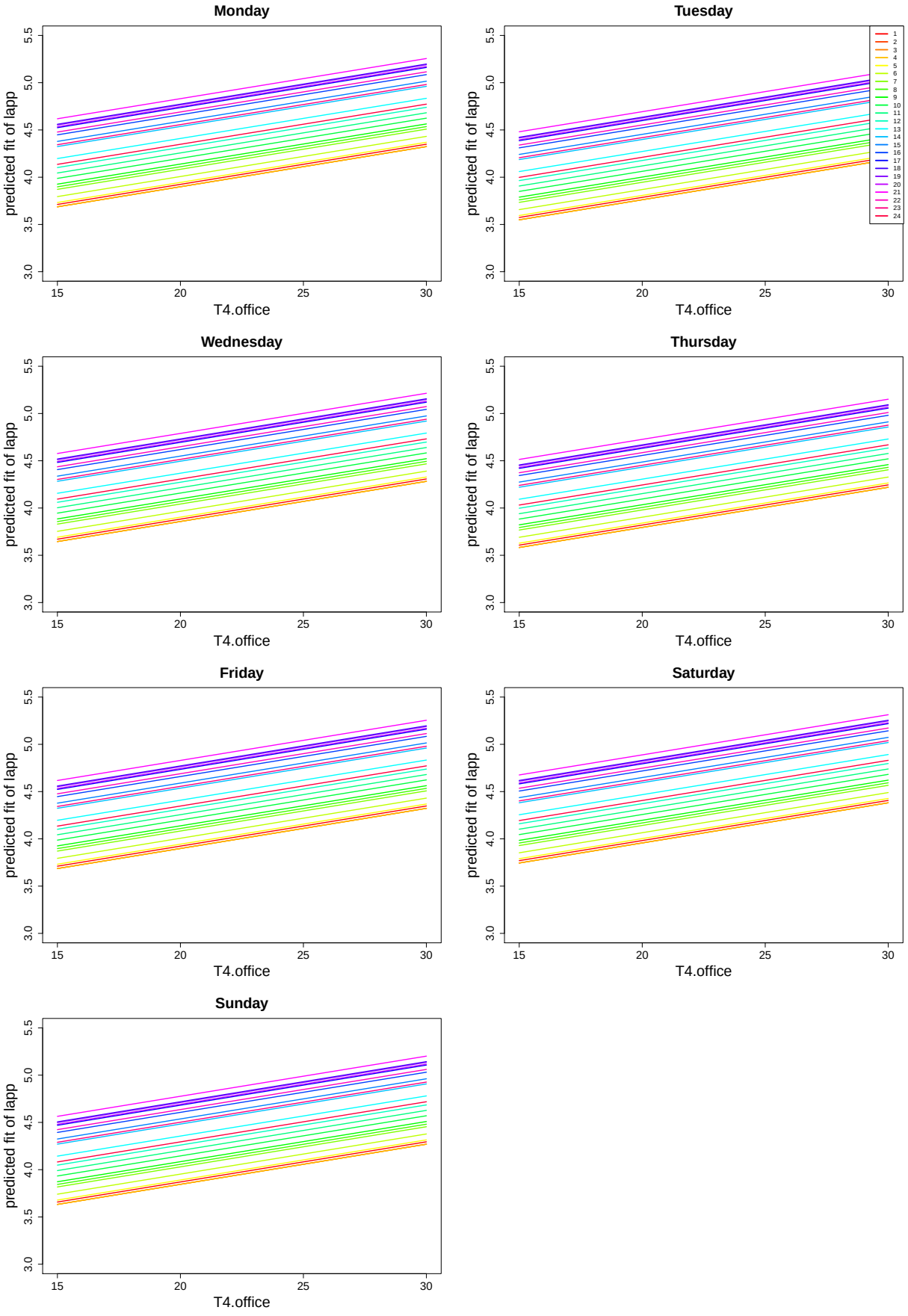



Figure A.10: Prediction of the full main model (Model 7) restricted on weekdays. The plot shows the relationship between fitted $(\widehat{\texttt{lapp}}_{i_h})_{i_h=(i\in[19735]|\texttt{hour}=h,\text{with}h=1,\ldots,24)}$ and `T4.office` to see the variation of hourly-wise pattern for all the weekdays. Condition $h$ is colored by hours 1 to 24.

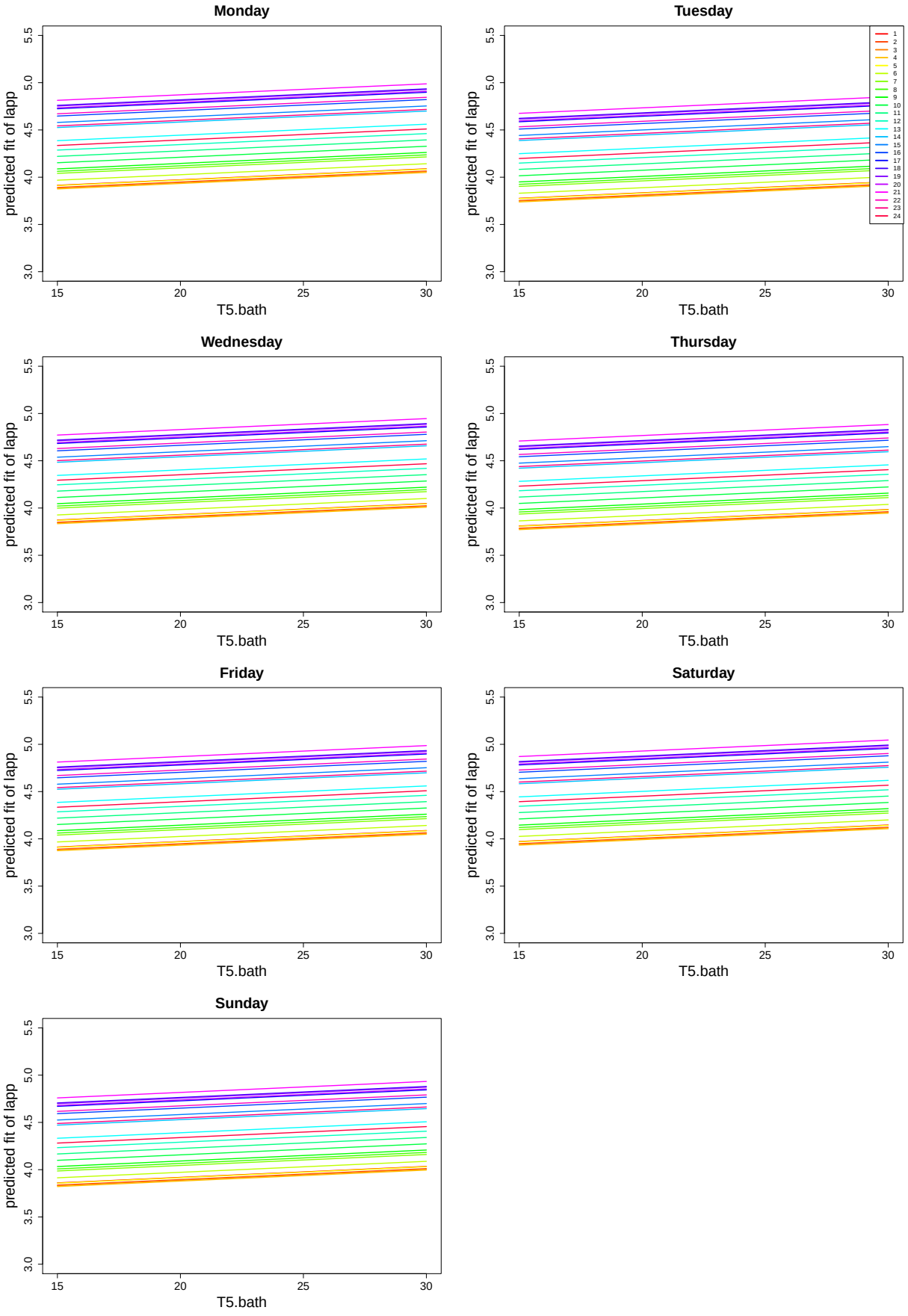



Figure A.11: Prediction of the full main model (Model 7) restricted on weekdays. The plot shows the relationship between fitted $(\widehat{\texttt{lapp}}_{i_h})_{i_h=(i\in[19735]|\texttt{hour}=h,\text{with}h=1,\ldots,24)}$ and `T5.bath` to see the variation of hourly-wise pattern for all the weekdays, while the other covariates fixed at their medians. Condition $h$ is colored by hours 1 to 24.

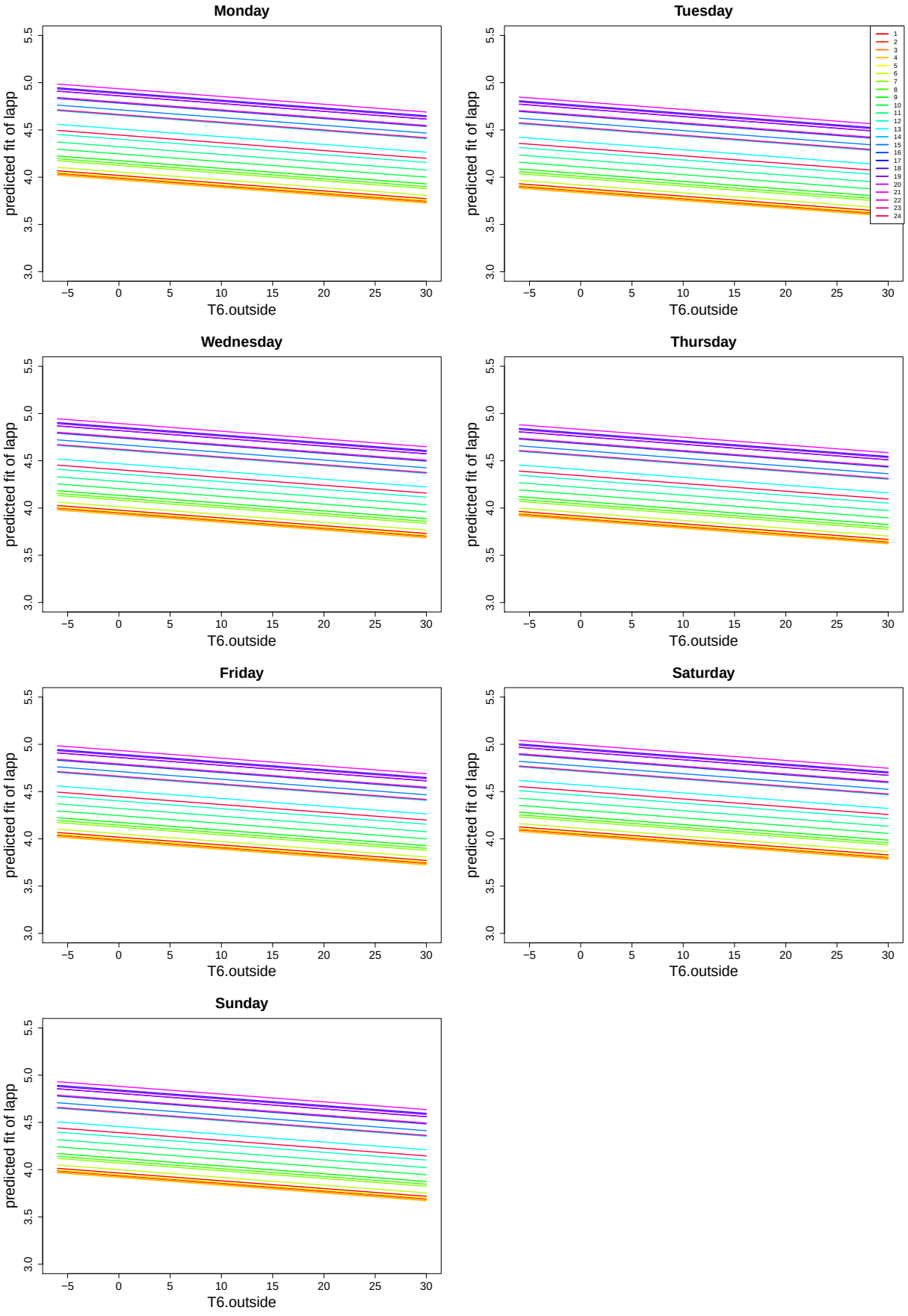



Figure A.12: Prediction of the full main model (Model 7) restricted on weekdays. The plot shows the relationship between fitted $(\widehat{\texttt{lapp}}_{i_h})_{i_h=(i\in[19735]|\texttt{hour}=h,\text{with}h=1,\ldots,24)}$ and `T6.outside` to see the variation of hourly-wise pattern for all the weekdays, while the other covariates fixed at their medians. Condition $h$ is colored by hours 1 to 24.

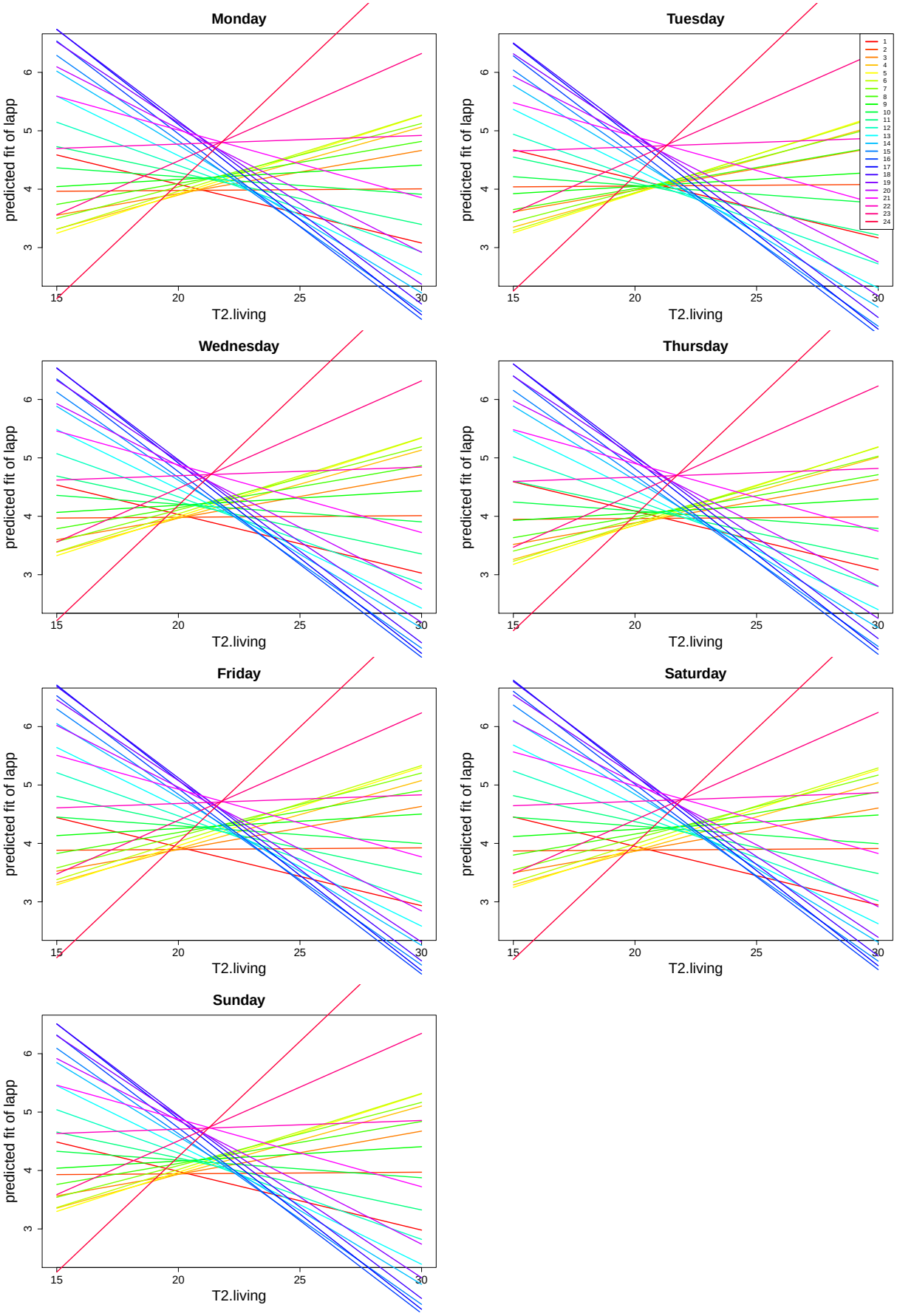



Figure A.13: Prediction of LMinter restricted on weekdays. The plot shows the relationship between fitted $(\widehat{\texttt{lapp}}_{i_h})_{i_h=(i\in[19735]|\texttt{hour}=h,\text{with}h=1,\ldots,24)}$ and `T2.living` to see the variation of hourly-wise pattern for all the weekdays, while the other covariates fixed at their medians. Condition $h$ is colored by hours 1 to 24.

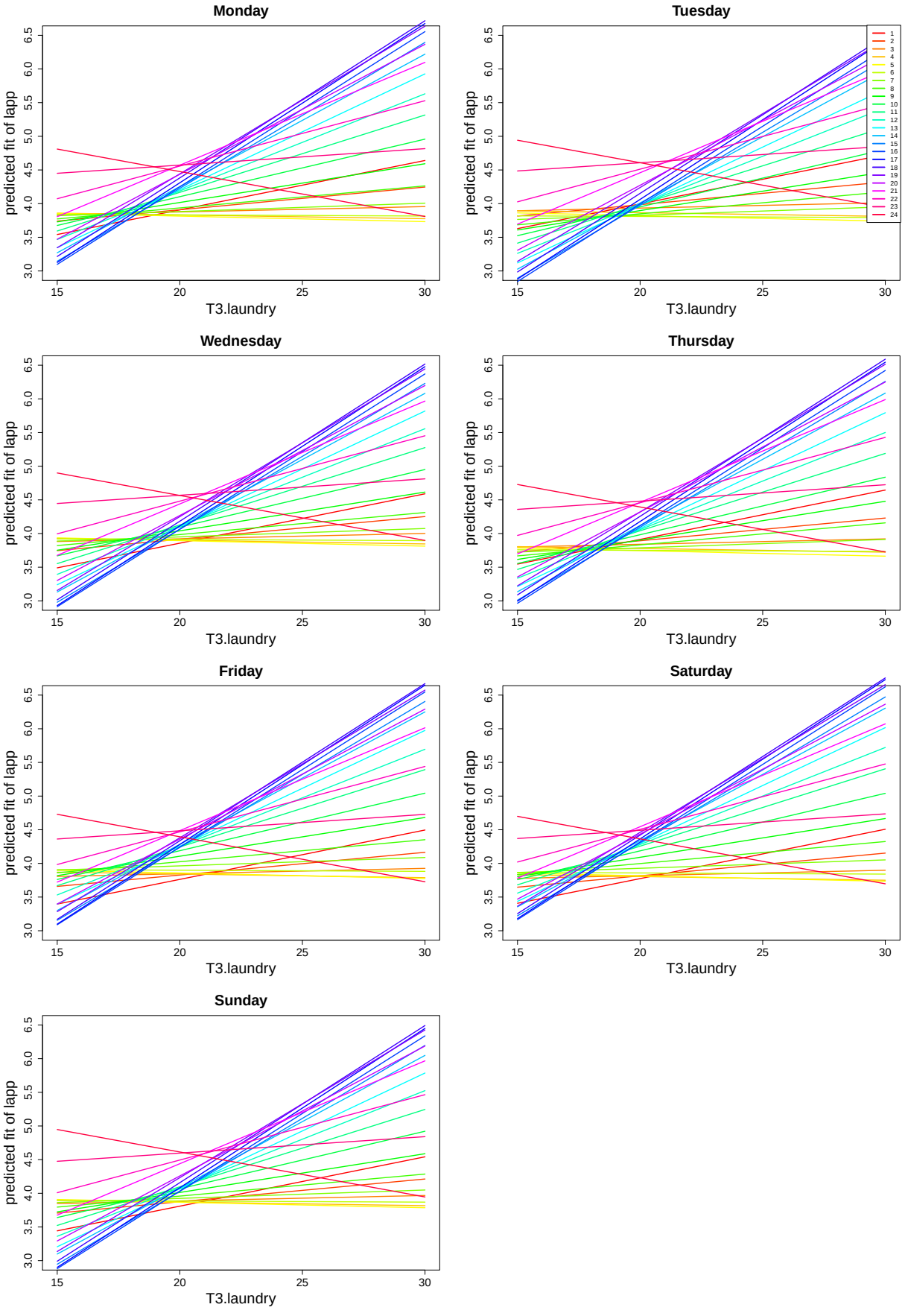



Figure A.14: Prediction of LMinter restricted on weekdays. The plot shows the relationship between fitted $(\widehat{\texttt{lapp}}_{i_h})_{i_h=(i\in[19735]|\texttt{hour}=h,\text{with}h=1,\ldots,24)}$ and `T3.laundry` to see the variation of hourly-wise pattern for all the weekdays, while the other covariates fixed at their medians. Condition $h$ is colored by hours 1 to 24.

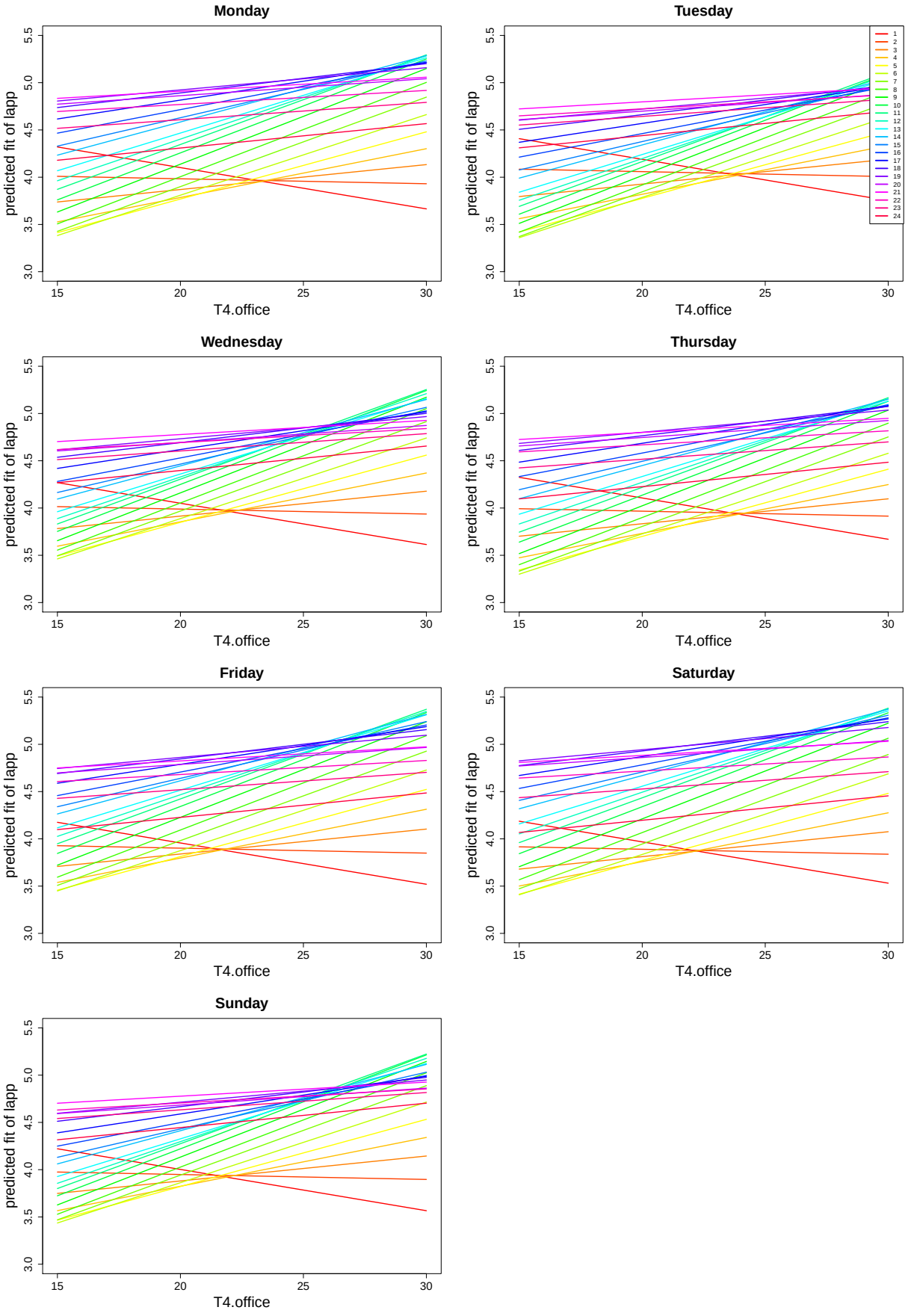



Figure A.15: Prediction of LMinter restricted on weekdays. The plot shows the relationship between fitted $(\widehat{\texttt{lapp}}_{i_h})_{i_h=(i\in[19735]|\texttt{hour}=h,\text{with}h=1,\ldots,24)}$ and `T4.office` to see the variation of hourly-wise pattern for all the weekdays, while the other covariates fixed at their medians. Condition $h$ is colored by hours 1 to 24.

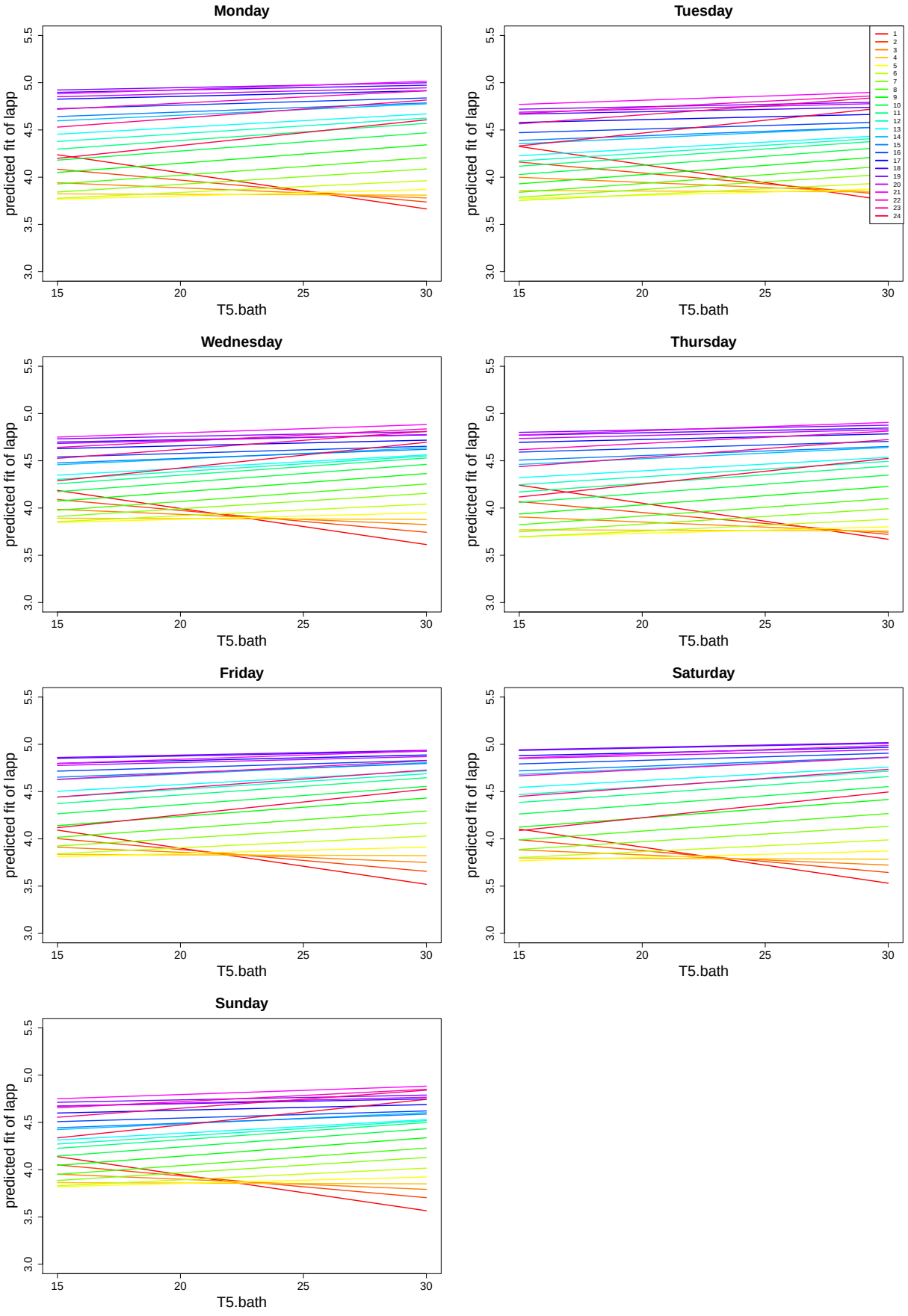



Figure A.16: Prediction of LMinter restricted on weekdays. The plot shows the relationship between fitted $(\widehat{\texttt{lapp}}_{i_h})_{i_h=(i\in[19735]|\texttt{hour}=h,\text{with}h=1,\ldots,24)}$ and `T5.bath` to see the variation of hourly-wise pattern for all the weekdays, while the other covariates fixed at their medians. Condition $h$ is colored by hours 1 to 24.

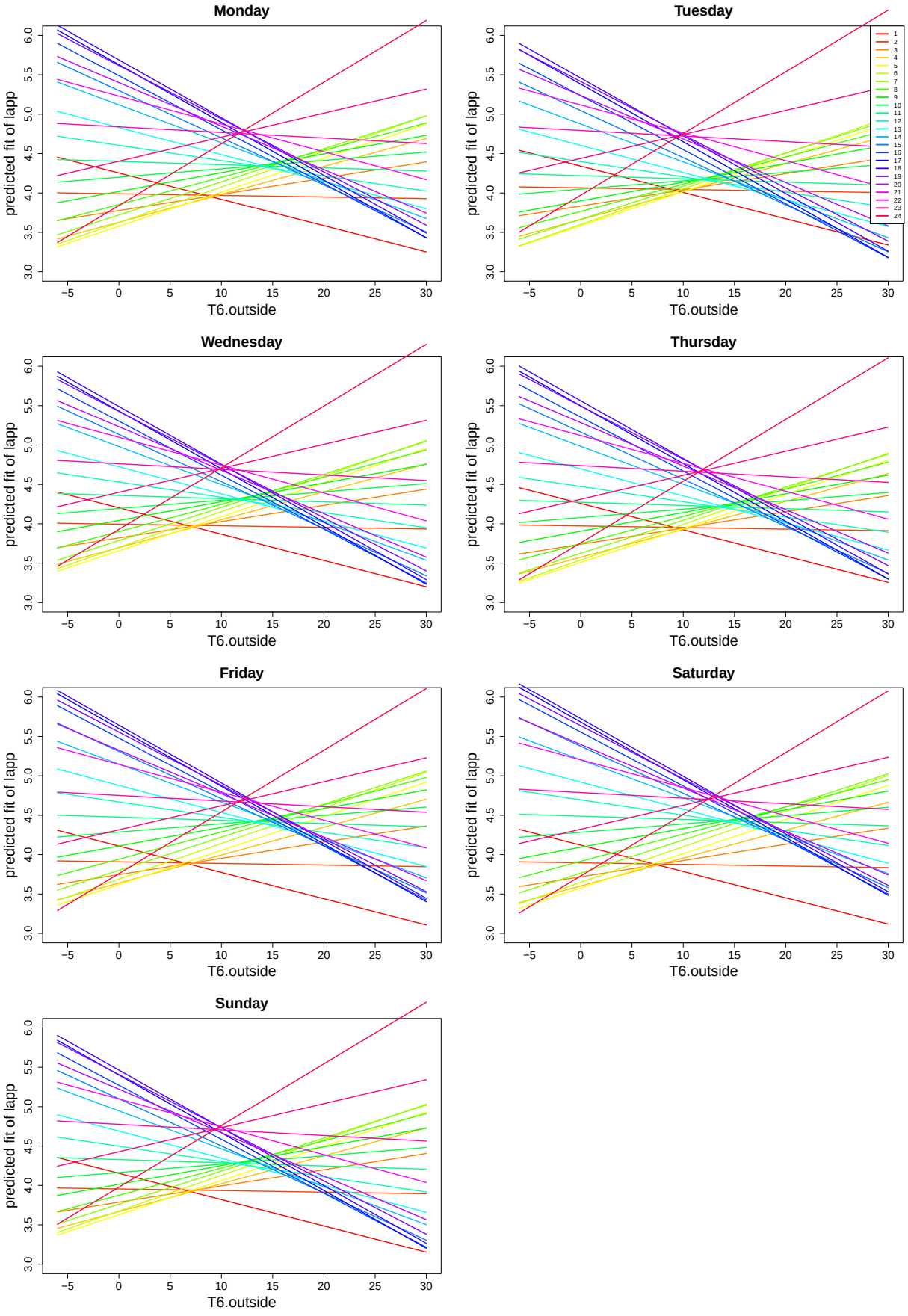



Figure A.17: Prediction of LMinter restricted on weekdays. The plot shows the relationship between fitted $(\widehat{\texttt{lapp}}_{i_h})_{i_h=(i\in[19735]|\texttt{hour}=h,\text{with}h=1,\ldots,24)}$ and `T6.outside` to see the variation of hourly-wise pattern for all the weekdays, while the other covariates fixed at their medians. Condition $h$ is colored by hours 1 to 24.

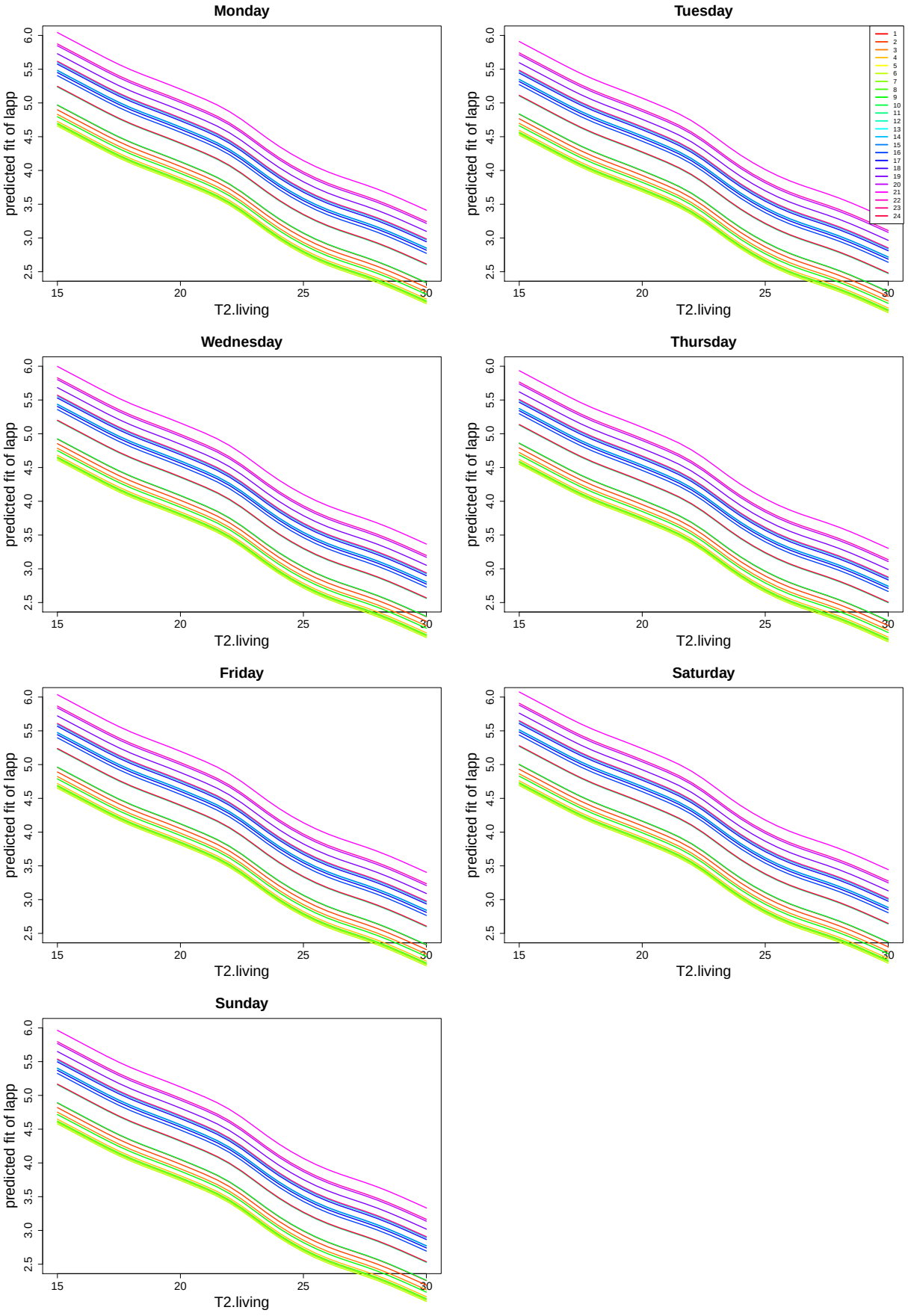



Figure A.18: Prediction of GAMmain restricted on weekdays. The plot shows the relationship between fitted $(\widehat{\texttt{lapp}}_{i_h})_{i_h=(i\in[19735]|\texttt{hour}=h,\text{with}h=1,\ldots,24)}$ and `T2.living` to see the variation of hourly-wise pattern for all the weekdays, while the other covariates fixed at their medians. Condition $h$ is colored by hours 1 to 24.

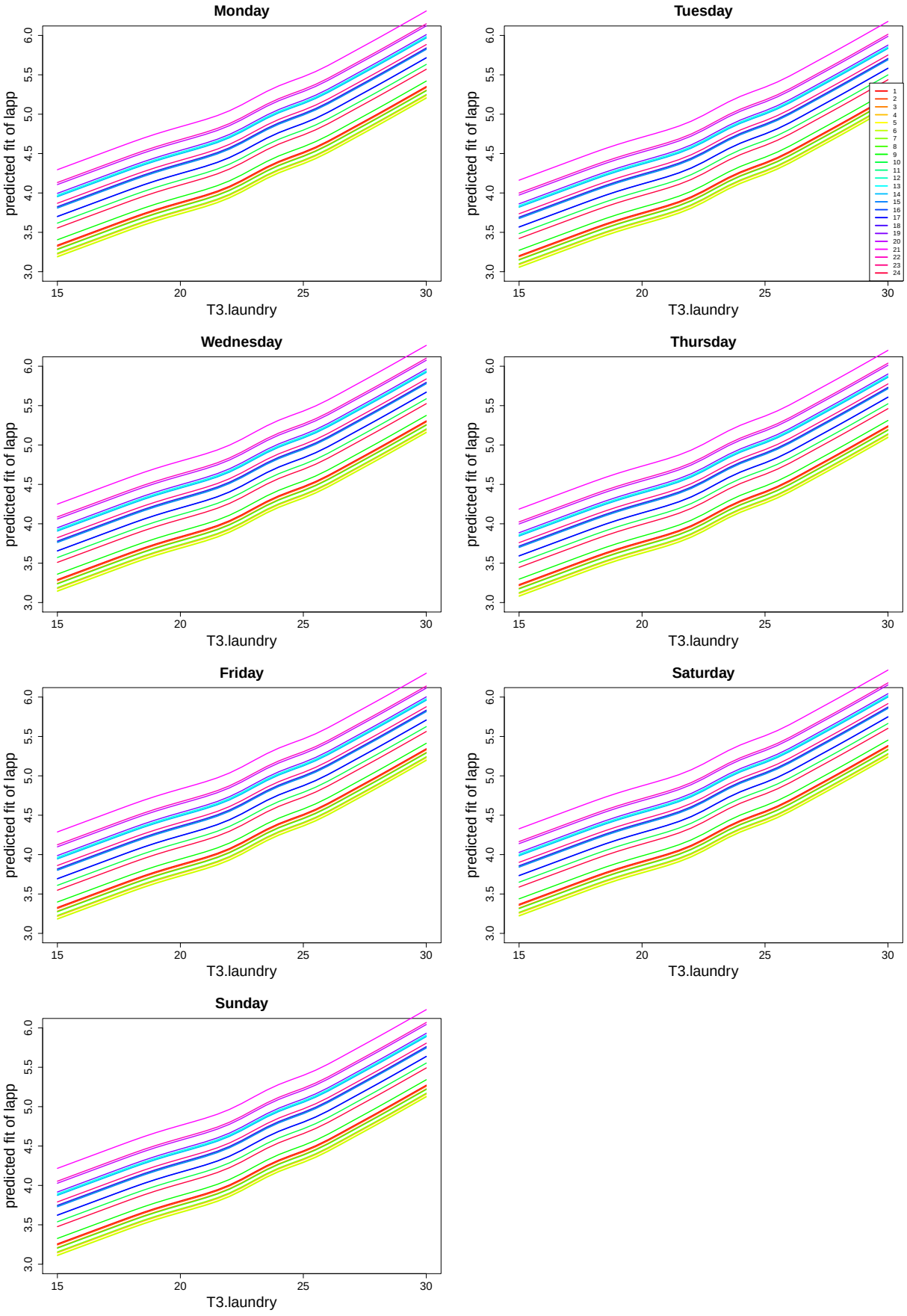



Figure A.19: Prediction of GAMmain restricted on weekdays. The plot shows the relationship between fitted $(\widehat{\texttt{lapp}}_{i_h})_{i_h=(i\in[19735]|\texttt{hour}=h,\text{with}h=1,\ldots,24)}$ and `T3.laundry` to see the variation of hourly-wise pattern for all the weekdays, while the other covariates fixed at their medians. Condition $h$ is colored by hours 1 to 24.

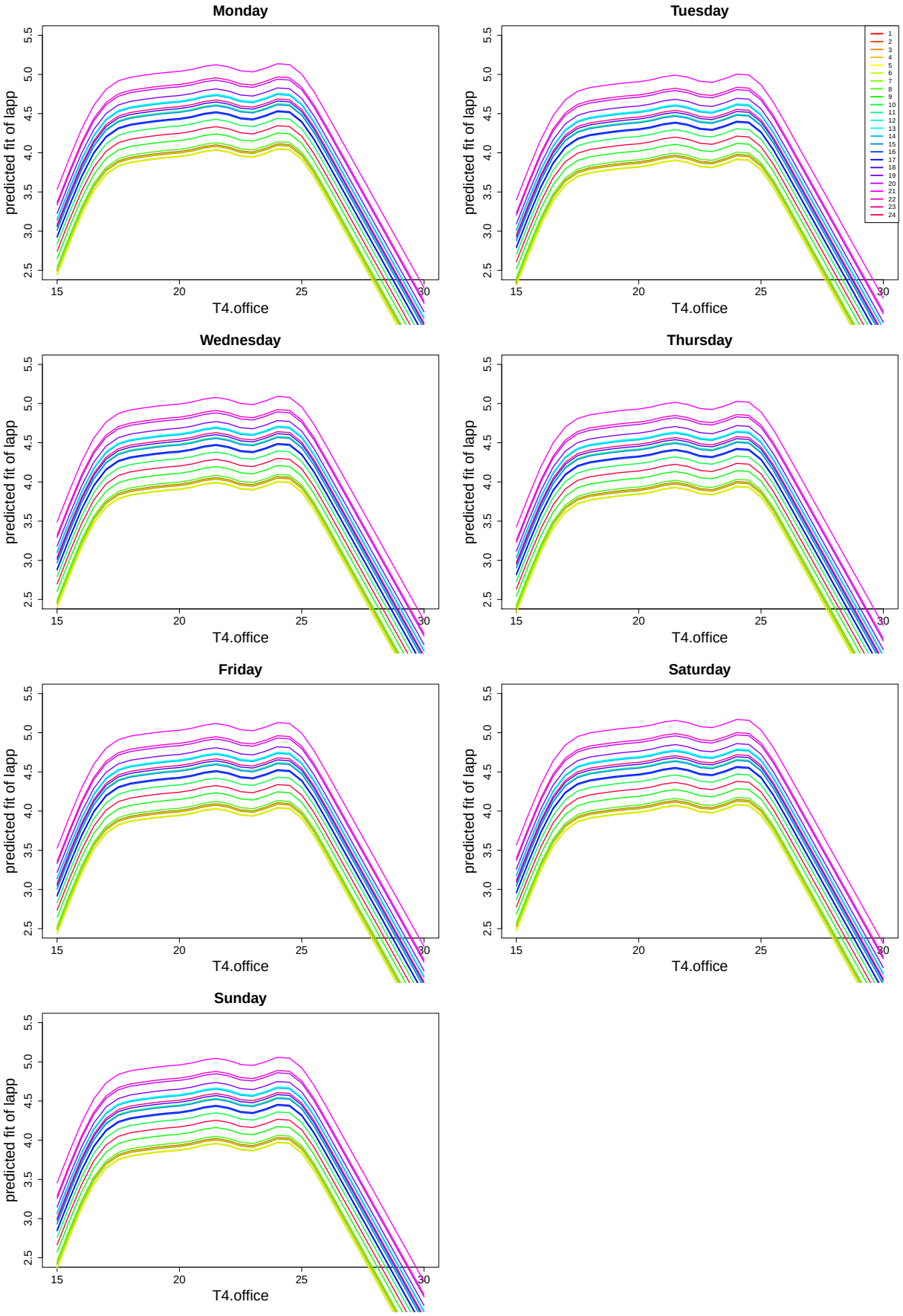



Figure A.20: Prediction of GAMmain restricted on weekdays. The plot shows the relationship between fitted $(\widehat{\texttt{lapp}}_{i_h})_{i_h=(i\in[19735]|\texttt{hour}=h,\text{with}h=1,\ldots,24)}$ and `T4.office` to see the variation of hourly-wise pattern for all the weekdays, while the other covariates fixed at their medians. Condition $h$ is colored by hours 1 to 24.

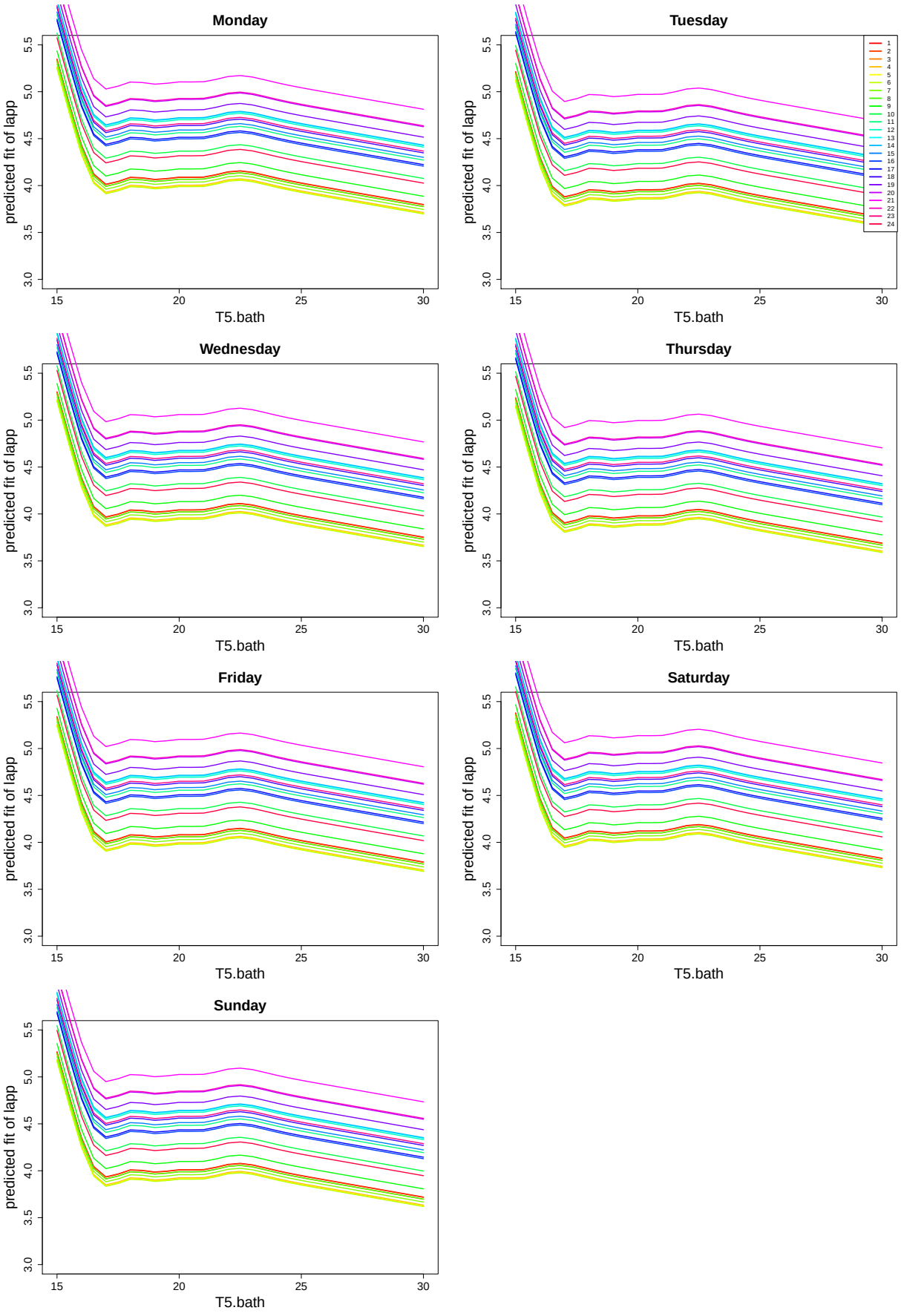



Figure A.21: Prediction of GAMmain restricted on weekdays. The plot shows the relationship between fitted $(\widehat{\texttt{lapp}}_{i_h})_{i_h=(i\in[19735]|\texttt{hour}=h,\text{with}h=1,\ldots,24)}$ and `T5.bath` to see the variation of hourly-wise pattern for all the weekdays, while the other covariates fixed at their medians. Condition $h$ is colored by hours 1 to 24.

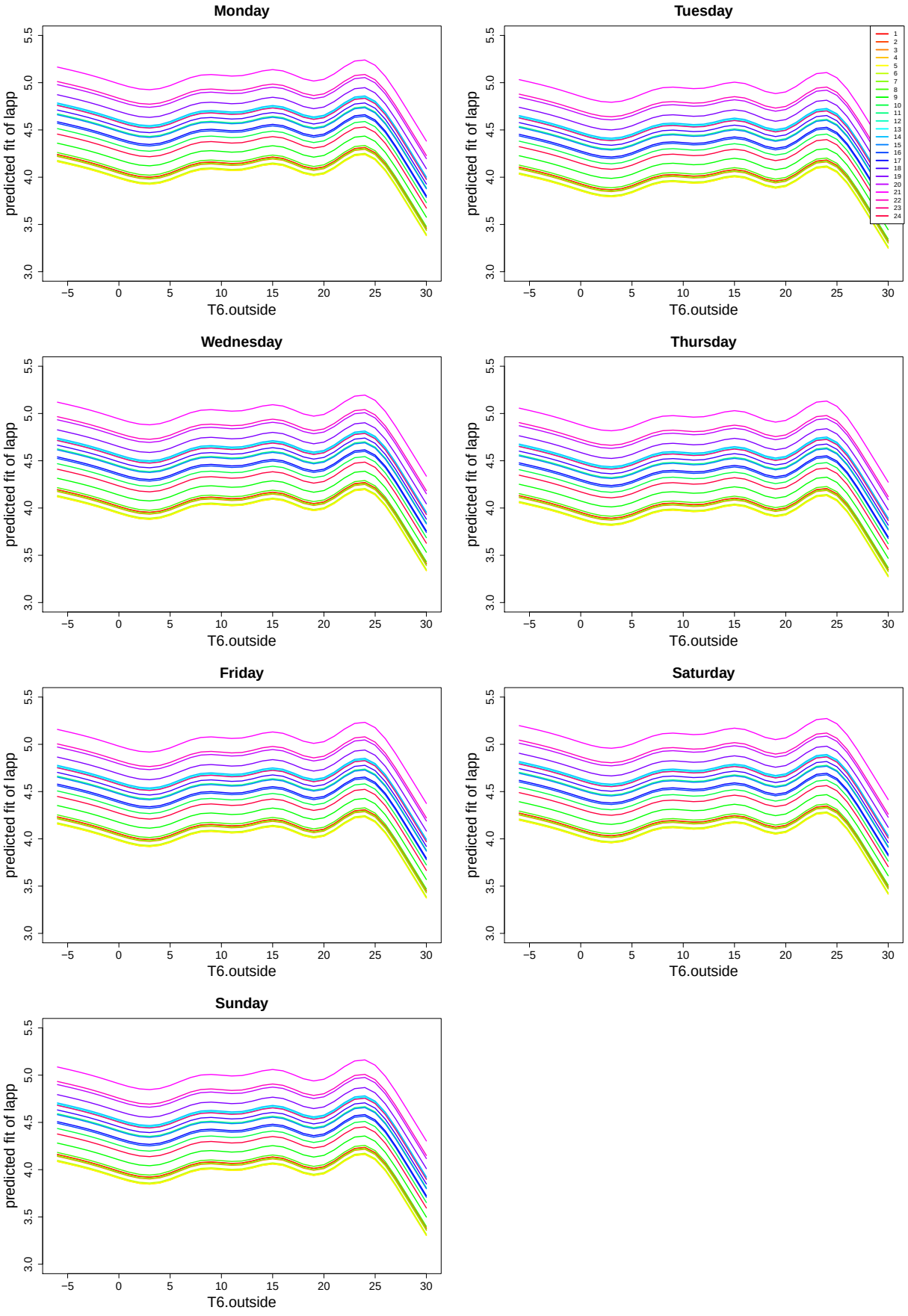



Figure A.22: Prediction of GAMmain restricted on weekdays. The plot shows the relationship between fitted $(\widehat{\texttt{lapp}}_{i_h})_{i_h=(i\in[19735]|\texttt{hour}=h,\text{with}h=1,\ldots,24)}$ and `T6.outside` to see the variation of hourly-wise pattern for all the weekdays, while the other covariates fixed at their medians. Condition $h$ is colored by hours 1 to 24.

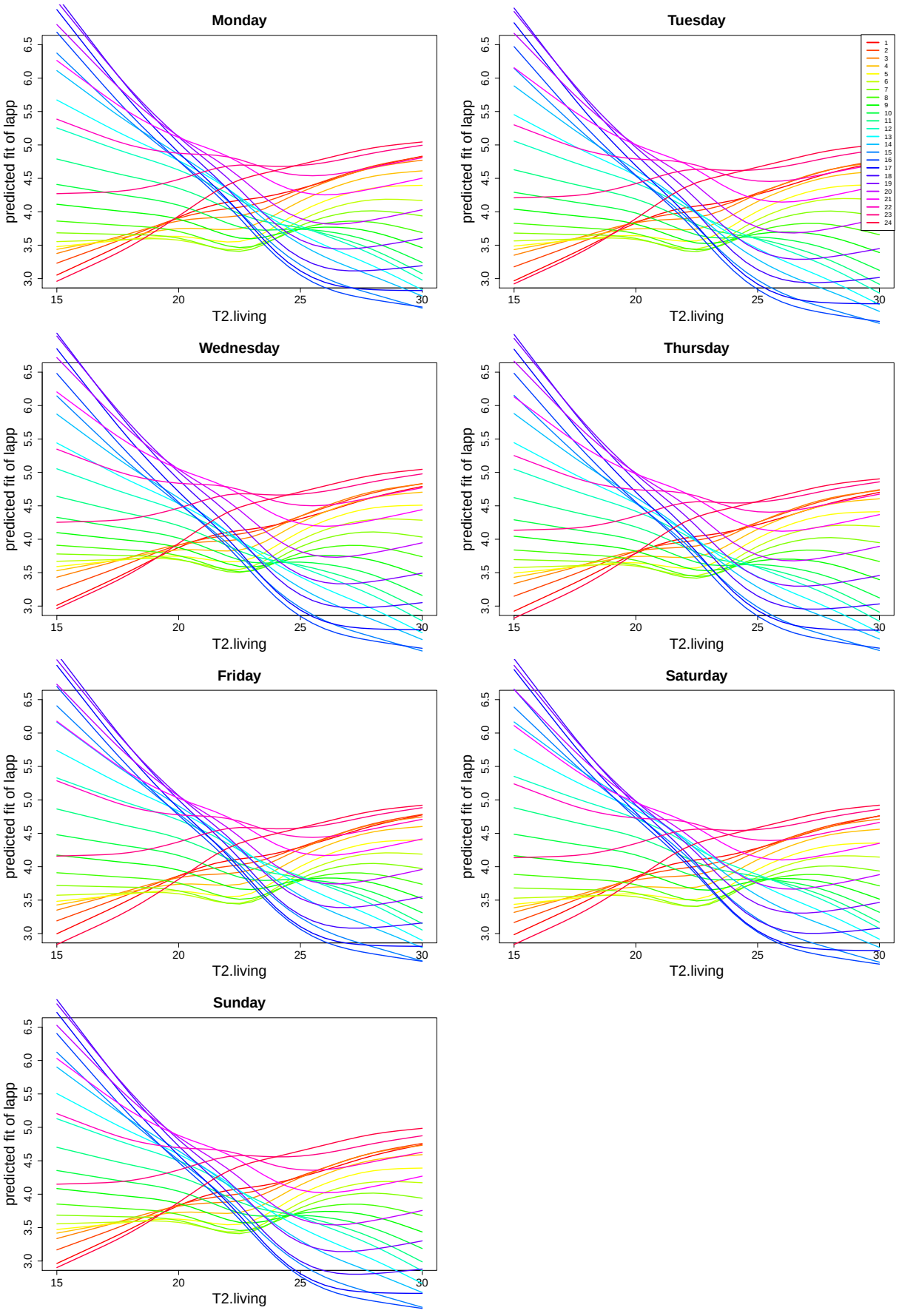



Figure A.23: Prediction of GAMinter restricted on weekdays. The plot shows the relationship between fitted $(\widehat{\texttt{lapp}}_{i_h})_{i_h=(i\in[19735]|\texttt{hour}=h,\text{with}h=1,\ldots,24)}$ and `T2.living` to see the variation of hourly-wise pattern for all the weekdays, while the other covariates fixed at their medians. Condition $h$ is colored by hours 1 to 24.

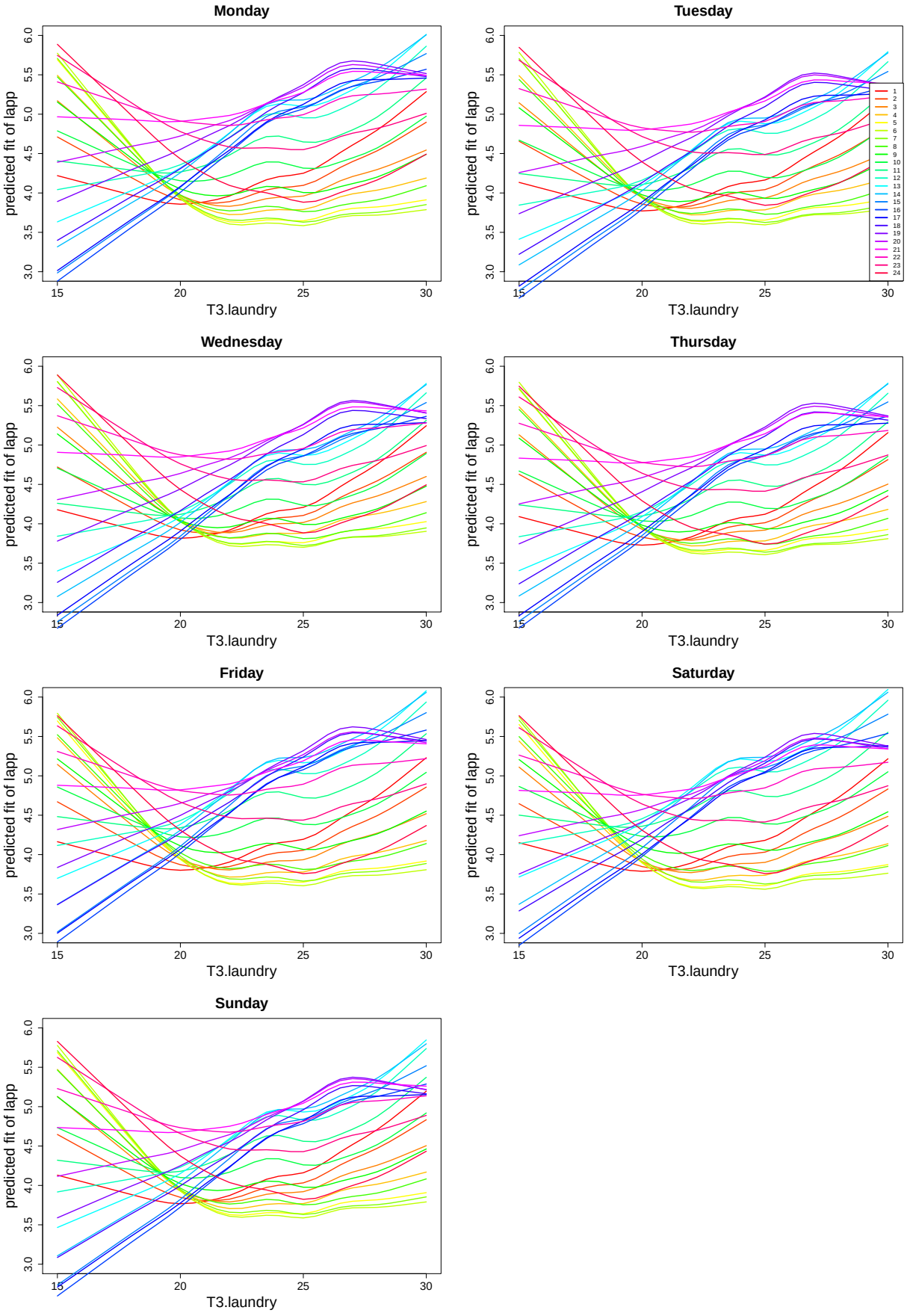



Figure A.24: Prediction of GAMinter restricted on weekdays. The plot shows the relationship between fitted $(\widehat{\texttt{lapp}}_{i_h})_{i_h=(i\in[19735]|\texttt{hour}=h,\text{with}h=1,\ldots,24)}$ and `T3.laundry` to see the variation of hourly-wise pattern for all the weekdays, while the other covariates fixed at their medians. Condition $h$ is colored by hours 1 to 24.

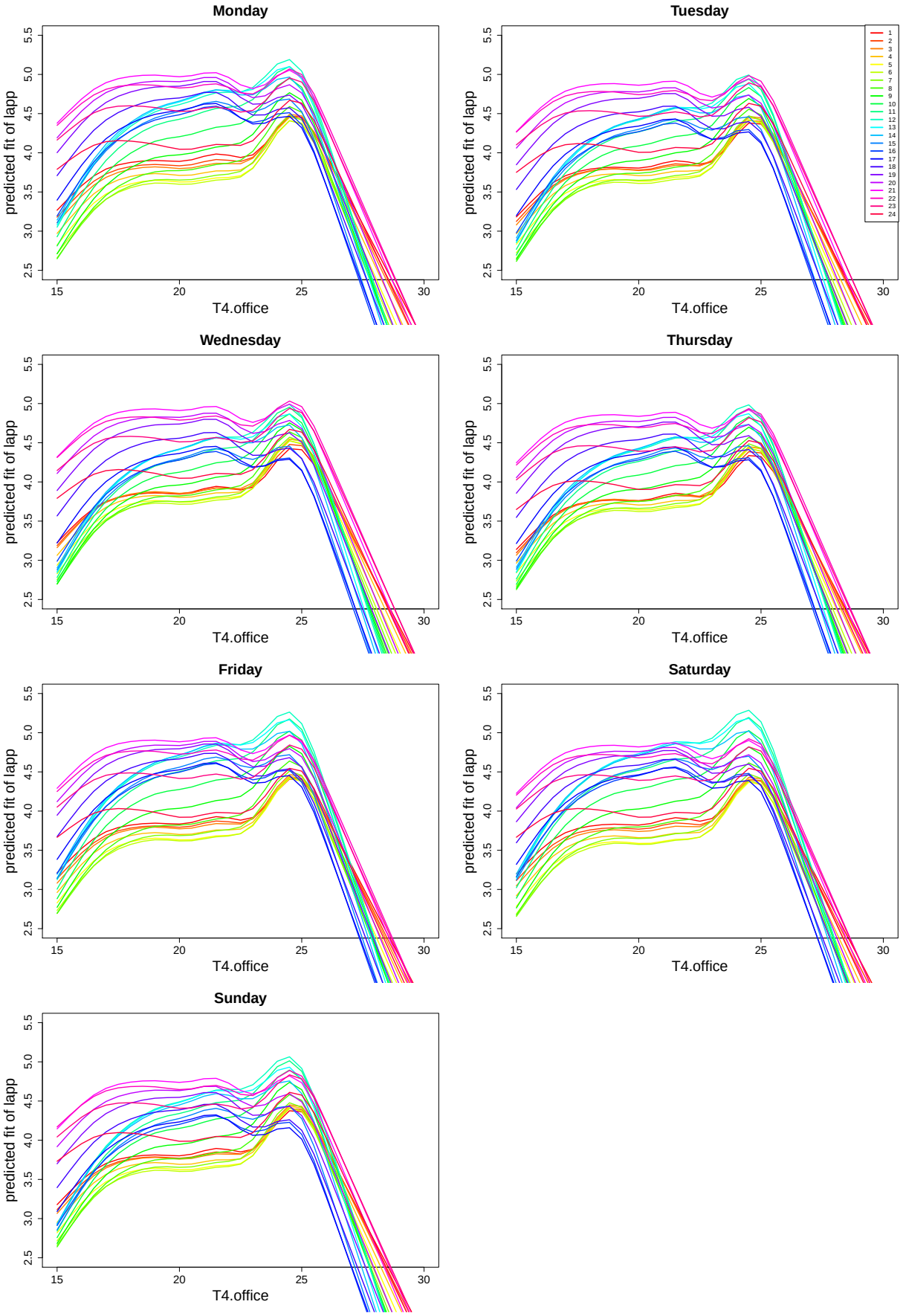



Figure A.25: Prediction of GAMinter restricted on weekdays. The plot shows the relationship between fitted $(\widehat{\texttt{lapp}}_{i_h})_{i_h=(i\in[19735]|\texttt{hour}=h,\text{with}h=1,\ldots,24)}$ and `T4.office` to see the variation of hourly-wise pattern for all the weekdays, while the other covariates fixed at their medians. Condition $h$ is colored by hours 1 to 24.

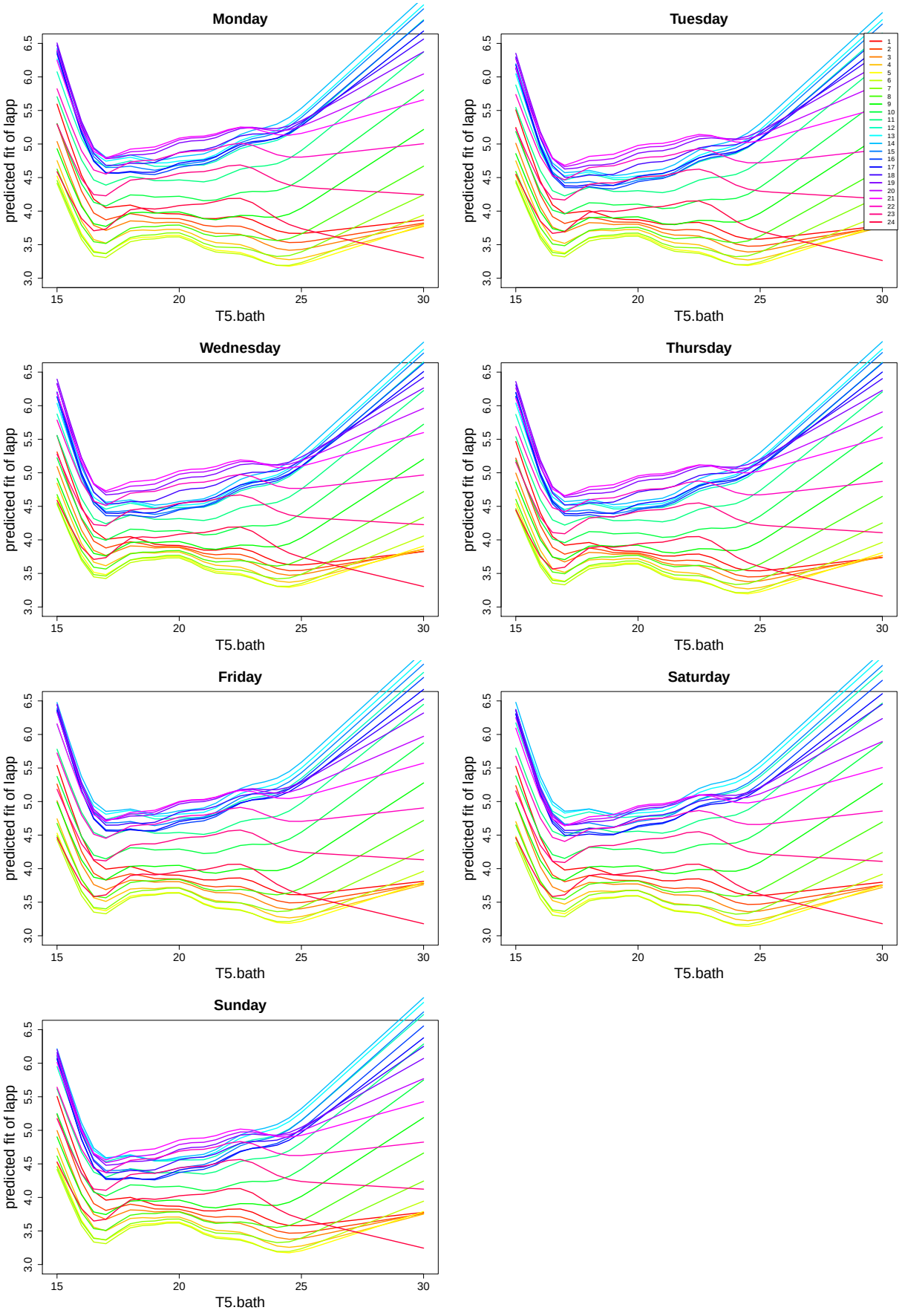



Figure A.26: Prediction of GAMinter restricted on weekdays. The plot shows the relationship between fitted $(\widehat{\texttt{lapp}}_{i_h})_{i_h=(i\in[19735]|\texttt{hour}=h,\text{with}h=1,\ldots,24)}$ and `T5.bath` to see the variation of hourly-wise pattern for all the weekdays, while the other covariates fixed at their medians. Condition $h$ is colored by hours 1 to 24.

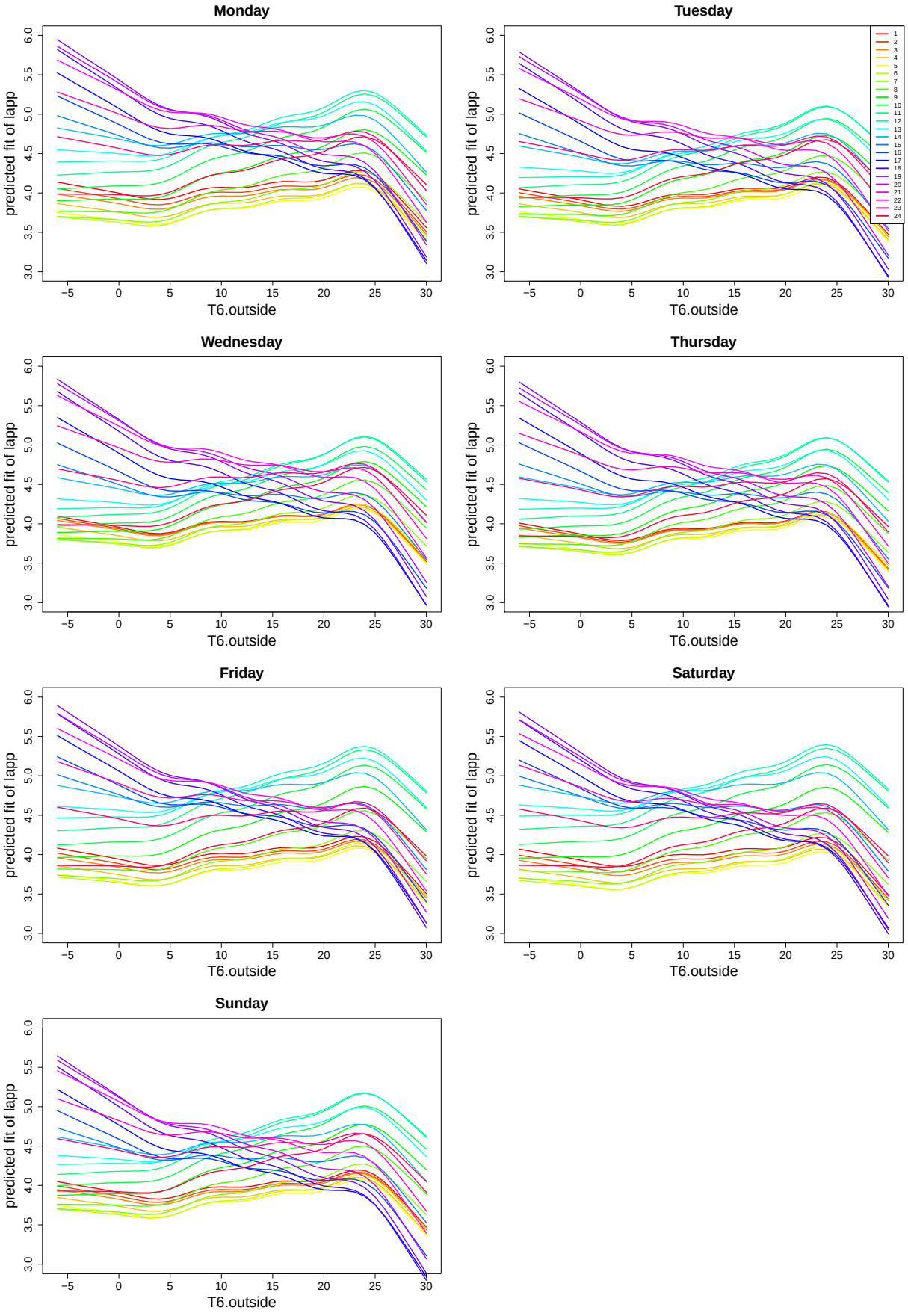



Figure A.27: Prediction of GAMinter restricted on weekdays. The plot shows the relationship between fitted $(\widehat{\texttt{lapp}}_{i_h})_{i_h=(i\in[19735]|\texttt{hour}=h,\text{with}h=1,\ldots,24)}$ and `T6.outside` to see the variation of hourly-wise pattern for all the weekdays, while the other covariates fixed at their medians. Condition $h$ is colored by hours 1 to 24.

# Bibliography


Abeare, S.
2009. Comparisons of boosted regression tree, glm and gam performance in the standardization of yellowfin tuna catch-rate data from the gulf of mexico lonline [sic] fishery.

Azevedo-Filho, A. and R. D. Shachter
1994. Laplace's method approximations for probabilistic inference in belief networks with continuous variables. In *Uncertainty Proceedings 1994*, Pp. 28–36. Elsevier.

Benjamini, Y. and Y. Hochberg
1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300.

Brockwell, P. J. and R. A. Davis
2016. *Introduction to time series and forecasting*. springer.

Brockwell, P. J., R. A. Davis, and S. E. Fienberg
1991. *Time Series: Theory and Methods: Theory and Methods*. Springer Science & Business Media.

Candanedo, J., V. Dehkordi, and M. Stylianou
2013. Model-based predictive control of an ice storage device in a building cooling system. *Applied Energy*, 111:1032–1045.

Candanedo, L. M., V. Feldheim, and D. Deramaix
2017. Data driven prediction models of energy use of appliances in a low-energy house. *Energy and buildings*, 140:81–97.

Cetin, K., P. Tabares-Velasco, and A. Novoselac
2014. Appliance daily energy use in new residential buildings: Use profiles and variation in time-of-use. *Energy and Buildings*, 84:716–726.

Cetin, K. S.
2016. Characterizing large residential appliance peak load reduction potential utilizing a probabilistic approach. *Science and Technology for the Built Environment*, 22(6):720–732.

Christensen, R.
2018. *Analysis of variance, design, and regression: linear modeling for unbalanced data*. Chapman and Hall/CRC.

Crawley, M. J.
2012. *The R book*. John Wiley & Sons.

Czado, C., J. Pfettner, S. Gschlößl, and F. Schiller
2009. Nonnested model comparison of glm and gam count regression models for life insurance data. *ASTIN Bulletin*.

Czado, C. and T. Schmidt
2011. *Mathematische Statistik*. Springer-Verlag.

Fahrmeir, L., C. Heumann, R. Künstler, I. Pigeot, and G. Tutz
2016. *Statistik: Der weg zur datenanalyse*. Springer-Verlag.

Fahrmeir, L., T. Kneib, S. Lang, and B. Marx
2013. *Regression: models, methods and applications*. Springer Science & Business Media.

Firth, S., K. Lomas, A. Wright, and R. Wall
2008. Identifying trends in the use of domestic appliances from household electricity consumption measurements. *Energy and buildings*, 40(5):926–936.

Friedman, J., T. Hastie, and R. Tibshirani
2001. *The elements of statistical learning*, volume 1. Springer series in statistics New York.

Fumo, N., P. Mago, and R. Luck
2010. Methodology to estimate building energy consumption using energyplus benchmark models. *Energy and Buildings*, 42(12):2331–2337.

Guisan, A., T. C. Edwards Jr, and T. Hastie
2002. Generalized linear and generalized additive models in studies of species distributions: setting the scene. *Ecological modelling*, 157(2-3):89–100.

Hastie, T. and R. Tibshirani
1990. Generalized additive models, volume 43 of. *Monographs on statistics and applied probability*, P. 15.

Hurvich, C. M., J. S. Simonoff, and C.-L. Tsai
1998. Smoothing parameter selection in nonparametric regression using an improved akaike information criterion. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60(2):271–293.

Karpfinger, C., T. Arens, F. Hettlich, U. Kockelkorn, K. Lichtenegger, and H. Stachel
2015. Differenzialrechnung – veraenderungen kalkulieren. In *Mathematik*, Pp. 315–372. Springer.

Kass, R. E. and D. Steffey
1989. Approximate bayesian inference in conditionally independent hierarchical models (parametric empirical bayes models). *Journal of the American Statistical Association*, 84(407):717–726.

Kavousian, A., R. Rajagopal, and M. Fischer
2015. Ranking appliance energy efficiency in households: Utilizing smart meter data and energy efficiency frontiers to estimate and identify the determinants of appliance energy efficiency in residential buildings. *Energy and Buildings*, 99:220–230.

Khouloud, T., B. Hedia, B. Nissaf, S. Marc, M. Dhafer, and C. Kouni
2017. Comparative performance analysis for generalized additive and generalized linear modeling in epidemiology. *International Journal of Advanced Computer Science and Applications*, 8:418–423.

Leon, S. J., Å. Björck, and W. Gander
2013. Gram-schmidt orthogonalization: 100 years and more. *Numerical Linear Algebra with Applications*, 20(3):492–532.

Liu, W., Y. Yang, et al.
2011. Parametric or nonparametric? a parametricness index for model selection. *The Annals of Statistics*, 39(4):2074–2102.

Luis Candanedo, U. o. M. U.
2017-02-15. Appliances energy prediction data set. Accessed: 2020-01-08.

MacKay, D. J. and D. J. Mac Kay
2003. *Information theory, inference and learning algorithms*. Cambridge university press.

Miller, D. L.
2017-04-04. Why is the default smoothing method reml rather than gcv.cp? Accessed: 2020-01-08.

Reiss, P. T. and R. Todd Ogden
2009. Smoothing parameter selection for a class of semiparametric linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(2):505–523.

Ruellan, M., H. Park, and R. Bennacer
2016. Residential building energy demand and thermal comfort: Thermal dynamics of electrical appliances and their impact. *Energy and Buildings*, 130:46–54.

Ryan, J. A. and J. M. Ulrich
2011. xts: extensible time series. *R package version 0.8-2*.

Sattelmayer, T.
2008. Technische thermodynamik: Energielehre und stoffverhalten.

Shmueli, G. et al.
2010. To explain or to predict? *Statistical science*, 25(3):289–310.

Stone, M.
1974. Cross-validation and multinomial prediction. *Biometrika*, 61(3):509–515.

Stone, M.
1977. An asymptotic equivalence of choice of model by cross-validation and akaike's criterion. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):44–47.

Swartman, R. K.
1981. Active solar cooling. In *Solar Energy Conversion II*, Pp. 469–475. Elsevier.

Wood, S. and M. S. Wood
2019. Package mgcv. *R package version*, 1.8-31:309.

Wood, S. N.
2011. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(1):3–36.

Wood, S. N.
2017. *Generalized additive models: an introduction with R*. Chapman and Hall/CRC.