Munich RE

Technische Universität München

Department of Mathematics

D-vine Regression in Insurance

Master's Thesis

by

Tamara Simjanoska

in cooperation with Munich Reinsurance Company

Supervisor:	Prof. Dr. Claudia Czado
Advisors:	Prof. Dr. Claudia Czado, M.Sc. Marija Tepegjozova
Submission Date:	11.05.2022

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

Garching, 11.05.2022

Acknowledgements

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 fruitful discussions that significantly enhanced the progress of my work.
Second, I would like to thank my advisor Marija Tepegjozova for all the time invested, her continuous support and guidance.
I would also like to thank Munich Re, especially Dr. Massimo Cavadini, for the collaboration and the support in this project.
Last but not least, I want to thank my family and friends, especially my mother and my brother, for all the encouragement throughout my life. Your unconditional love and support in each step of the way is the foundation of my personal development. I also want to thank my father, for all the guidance he provided me during my childhood, for the optimism and confidence in me.

Abstract

Due to the continuous rise in the amount of data over the recent years, insurance companies regularly seek for an improvement in their statistical analysis of the insurance claims data. Our main goal in this thesis is to investigate D-vine quantile regression, introduced by Kraus and Czado (2017), as a modelling approach for motor insurance severity rate. For that purpose we present three additional regression methods; lognormal and gamma regression which are standard approaches in modelling positive, right-skewed data, and linear quantile regression which can be easily compared to D-vine regression since both regression methods predict conditional quantiles.

After laying the necessary fundamentals and the framework of the four regression methods, we perform an extensive exploratory data analysis for lognormal and gamma regression on two real-life motor insurance claims data sets. Then, we proceed with model fitting using the different regression methods. Finally, we evaluate and compare the resulting models based on several performance measures, some of which are the log likelihood, the training and test error and the interval score.

Zusammenfassung

Aufgrund des kontinuierlichen Anstiegs der Datenmenge in den letzten Jahren suchen Versicherungsunternehmen regelmäßig nach einer Verbesserung ihrer statistischen Analyse der Versicherungsschadendaten. Unser Hauptziel in dieser Arbeit ist es, die von Kraus und Czado (2017) eingeführte D-Vine-Quantilregression als Modellierungsansatz für die Kfz-Schadensquote zu untersuchen. Zu diesem Zweck stellen wir drei weitere Regressionsmethoden vor: die Lognormal- und die Gamma-Regression, die Standardansätze für die Modellierung positiver, rechtsschiefer Daten sind, und die lineare Quantilsregression, die leicht mit der D-Vine-Regression verglichen werden kann, da beide Regressionsmethoden bedingte Quantile vorhersagen.
Nachdem wir die notwendigen Grundlagen und den Rahmen der vier Regressionsmethoden festgelegt haben, führen wir eine ausführliche explorative Datenanalyse für Lognormal- und Gamma-Regression an zwei realen Datensätzen von Kfz-Versicherungsansprüchen durch. Dann fahren wir mit der Modellanpassung unter Verwendung der verschiedenen Regressionsmethoden fort. Schließlich bewerten und vergleichen wir die resultierenden Modelle auf der Grundlage verschiedener Leistungsmaße, darunter die Log-Likelihood, den Trainings- und Testfehler und den Intervallwert.

Contents

1 Introduction

Modelling insurance claim severity is characterized by a response which is a positive and right-skewed random variable. One of the common approaches in severity risk modelling is the lognormal regression model, which is a subclass of the classical linear regression model. The transformation of the response using the logarithmic function ensures positive predicted values of the response, while the relationship between the transformed independent variable and the predictors is linear. The advantages of the linear regression models are that they are easy to interpret and have low computational cost, which is why they are applied in almost every discipline.
Another common approach in severity risk modelling is the gamma regression model, which falls into the class of generalized linear models (GLMs), first introduced by Nelder and Wedderburn (1972). Compared to linear regression models, this class of models allows for non-normally distributed responses, while still keeping the linear relationship via the link function. The gamma regression model relaxes the assumption of constant variance in the linear regression model. In particular, it allows for increase of the variance of the response, as the mean of the response increases. Therefore, the gamma regression model offers more flexibility, while maintaining low computational cost and complexity. However, both the lognormal and gamma regression are limited to prediction of the mean of the response. An alternative approach which predicts conditional quantiles and is more robust against outliers is the quantile regression, first introduced by Koenker and Bassett (1978).
Linear quantile regression (Koenker and Bassett (1978)) complements linear regression by providing a more accurate modelling of the relationship between the variables, especially in the tails. In our case this is particularly useful, since insurance companies are interested in the extreme claim severity observations in the tails. However, this method can lead to issues like quantile crossings, transformations, interactions and collinearity. Additionally, Bernand and Czado (2015) show that the linearity assumption is strong and almost never fulfilled.
One of the ways to overcome the shortfalls of linear quantile regression is to use vine copula quantile regression. This method was introduced by Kraus and Czado (2017) and models multivariate data using bivariate building blocks, a procedure called pair copula construction (PCC), by sequentially adding variables in the model based on the maximum conditional log likelihood. The D-vine quantile regression results in a highly flexible model with easily extractable conditional quantiles. In addition to Kraus and Czado (2017), we also refer to the work of Tepegjozova, Zhou, Claeskens and Czado (2022) where a fully nonparametric D-vine quantile regression is discussed. In particular, our main goal in this thesis is to analyse the performance of this regression approach compared to the lognormal, gamma and linear quantile regression.
The remainder of the thesis is organized as follows. Chapter 2 gives a review of the statistical concepts our analyisis builds on. In Chapter 3, we introduce the third party liability motor claims data set and the division of this data set to two different data sets, based on the bonus malus class of the policy holders, and we present the exploratory data analysis performed on the training data sets which is a necessary preprocessing step for lognormal and gamma regression. In Chapter 5, we present the fitted regression models on the data set where the policy holders belong to the best bonus malus class and we compare the

models using the performance measures presented in Chapter 2. Similarly, we repeat this procedure more extensively on the second data set in Chapter 5, since the R^2_{adj} of the lognormal models on this data set was slightly higher than the lognormal models fitted on the first data set. Chapter 6 concludes.

2 Theoretical background

In this chapter we will present the necessary theoretical background for the remainder of this thesis. First, we define some univariate and multivariate distributions. Then, we discuss four different statistical models that we fit to our data. Finally, we explain the performance measures used to compare the models.
We denote vectors with bold letters, random variables with capital letters and observed values with small letters.

2.1 Univariate and multivariate distributions

We present all probability distributions we use in this thesis. We denote by f the *probability density function (pdf)* and by $\mathbb{R}_+$ the set of all positive real numbers. For this section we consult Czado and Schmidt (2011), Ahsanullah (2017), Basso, Lachos, Cabral and Ghosh (2010) and Czado (2019).

Definition 2.1.1 *(Uniform Distribution)*
Let $X \in \mathbb{R}$ be a random variable following the uniform distribution. Then the probability function of X at x is defined as

$$f(x) := \begin{cases} \frac{1}{b-a} & \text{for } a \leq x \leq b \\ 0 & \text{for } a > x \text{ or } x > b \end{cases} \tag{2.1.1}$$

and we write $X \sim \mathcal{U}(a, b)$. If $a = 0$ and $b = 1$ i.e. $X \sim \mathcal{U}(0, 1)$, X follows a standard uniform distribution.

Definition 2.1.2 *(Normal Distribution)*
Let $X \in \mathbb{R}$ be a random variable following the normal distribution with mean $\mu \in \mathbb{R}$ and variance $\sigma^2 > 0$. Then the probability function of X at x is defined as

$$f(x) := \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big\{ -\frac{1}{2\sigma^2}(x-\mu)^2 \Big\} \tag{2.1.2}$$

and we write $X \sim \mathcal{N}(\mu, \sigma^2)$. If $\mu = 0$ and $\sigma^2 = 1$ i.e. $X \sim \mathcal{N}(0, 1)$, X follows a standard normal distribution and its probability function is presented by ϕ, whereas its cumulative distribution function by Φ.

Definition 2.1.3 *(Lognormal Distribution)*
Let $X \in \mathbb{R}_+$ be a random variable following the lognormal distribution with location parameter $\mu > 0$ and scale parameter $\sigma^2 > 0$. Then the probability function of X at x is defined as

$$f(x) := \frac{1}{x\sqrt{2\pi\sigma^2}} \exp\Big\{ -\frac{1}{2}\Big(\frac{\ln(x-\mu)}{\sigma}\Big)^2 \Big\}, \quad x > 0 \tag{2.1.3}$$

and we write $X \sim LN(\mu, \sigma^2)$. If $X \sim \mathcal{N}(\mu, \sigma^2)$, then for the random variable $Y := e^X$ it holds that $Y \sim LN(\mu, \sigma^2)$.

Definition 2.1.4 *(Skew-Normal Distribution)*
Let $X \in \mathbb{R}$ be a random variable following the skew-normal distribution with location parameter $\mu \in \mathbb{R}$, scale parameter $\sigma^2 > 0$ and skewness parameter $\lambda \in \mathbb{R}$. Then the probability function of X at x is defined as

$$f(x) := 2\phi(x|\mu, \sigma^2)\Phi(\lambda\frac{y-\mu}{\sigma}), \tag{2.1.4}$$

where $\phi(\cdot|\mu, \sigma^2)$ denotes the density of an $\mathcal{N}(\mu, \sigma^2)$ random variable and Φ denotes the distribution function of the standard normal distribution, and we write $X \sim SN(\mu, \sigma^2, \lambda)$.

Definition 2.1.5 *(Gamma Distribution)*
Let $X \in \mathbb{R}_+$ be a random variable following the gamma distribution with shape $\alpha > 0$ and rate $\beta > 0$. Then the probability function of X at x is defined as

$$f(x) := \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}, \quad x > 0 \tag{2.1.5}$$

and we write $X \sim Gamma(\alpha, \beta)$, where Γ is the gamma function defined as

$$\Gamma(a) := \int_0^\infty t^{a-1} e^{-t} dt, \quad a > 0. \tag{2.1.6}$$

Definition 2.1.6 *(Chi-squared Distribution)*
Let $X \in \mathbb{R}_+$ be a random variable following the chi-squared distribution with degrees of freedom $n > 0$. Then the probability function of X at x is defined as

$$f(x) := \frac{1}{2^{n/2}\Gamma(\frac{n}{2})} x^{\frac{n}{2}-1} e^{-\frac{x}{2}}, \quad x > 0 \tag{2.1.7}$$

and we write $X \sim \chi^2_n$, where Γ is the gamma function defined with Equation (2.1.6).
If $X_1, \ldots, X_n$ are i.i.d. $\mathcal{N}(0,1)$ random variables, then $Y_n = \sum_{i=1}^n X_i^2$ is χ^2_n-distributed.

Definition 2.1.7 *(t-Distribution)*
Let $X \in \mathbb{R}$ be a random variable following the t-distribution with $\nu > 0$ degrees of freedom. Then the probability function of X at x is defined as

$$f_\nu(x) := \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\Gamma(\frac{1}{2})\sqrt{\nu}} \Big(1 + \frac{x^2}{\nu}\Big)^{-\frac{\nu+1}{2}} \tag{2.1.8}$$

and we write $X \sim t_\nu$, where Γ is the gamma function defined with Equation (2.1.6).
If $X \sim \mathcal{N}(0,1)$ 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. $\mathcal{N}(\mu, \sigma^2)$ 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 2.1.8 *(F-Distribution)*
Let $X \in \mathbb{R}_+$ be a random variable following the F-distribution with degrees of freedom $n, m > 0$. Then the probability function of X at x is defined as

$$f(x) := \frac{n^{n/2} m^{m/2}}{B(n/2, m/2)} \frac{x^{\frac{n}{2}-1}}{(m+nx)^{(n+m)/2}}, \quad x > 0 \tag{2.1.9}$$

and we write $X \sim F_{n,m}$*, where* B *is the beta function defined as*

$$B(a,b) = \int_0^1 t^{a-1}(1-t)^{b-1}dt, \quad a,b>0. \tag{2.1.10}$$

If $V \sim \chi^2_n$ *and* $W \sim \chi^2_m$ *are independent random variables, then for* $F := \frac{V/n}{W/m}$ *it holds that* $F \sim F_{n,m}$.
If $Y \sim t_m$ *then* $X = Y^2 \sim F_{1,m}$.

Definition 2.1.9 *(Multivariate Normal Distribution)*
A continuous p-dimensional vector $\boldsymbol{X} = (X_1, \ldots, X_p)^\top$ *is said to have a multivariate normal distribution if it has probability density function*

$$f(\boldsymbol{x}) := (2\pi)^{-\frac{p}{2}} |\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp\Big\{ -\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\Big\}, \tag{2.1.11}$$

with $\boldsymbol{\mu} \in \mathbb{R}^p$ *and positive definite* $(p \times p)$*-matrix* $\boldsymbol{\Sigma}$*. We write* $\boldsymbol{X} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

Definition 2.1.10 *(Multivariate t-Distribution)*
A continuous p-dimensional vector $\boldsymbol{X} = (X_1, \ldots, X_p)^\top$ *is said to have a multivariate t-distribution with* ν *degrees of freedom, location parameter* $\boldsymbol{\mu}$*, and (positive definite) dispersion matrix* $\boldsymbol{\Sigma}$*, if it has probability density function*

$$f(\boldsymbol{x}) := |\boldsymbol{\Sigma}|^{-\frac{1}{2}} (\nu\pi)^{-\frac{p}{2}} \frac{\Gamma((\nu+p)/2)}{\Gamma(\nu/2)} \Big(1 + \frac{(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}{\nu}\Big)^{-(\nu+p)/2} \tag{2.1.12}$$

and we write $\boldsymbol{X} \sim t_\nu(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

2.2 Data preprocessing

In this section we describe the transformed response variable used as a response variable for the lognormal, linear quantile and D-vine quantile regression models, as well as the split of a data set to a training and a test data set.

Data transformation

Throughout the thesis, we are interested in predicting a response variable using several predictors. In case of a positive response variable with wide range, an useful transformation of the response can be the natural logarithm (Fahrmeir, Kneib, Lang and Marx (2013)). This guarantees that our predicted values will be positive. We denote the natural logarithm with the abbreviation $\ln(\cdot)$. If $Y \in \mathbb{R}_+$ is the response variable then $\ln(Y)$ will be the transformed response variable.

Data splitting

As discussed by Hastie, Tibshirani and Friedman (2009), in a data-rich situation the best approach for model selection and model assessment is to randomly divide the data set into training and test data sets. The training data set is used to fit the models and

the test data set is used for assessment of the error of the final chosen model. The test data set should be brought out only at the end of the analysis, when we perform model evaluation. It is difficult to give a general rule on how to choose the number of observations in each of the two parts, as that depends on the data itself as well as on the sample size. In this thesis, as we have large data sets, we split them to roughly 90% training sample and 10% test sample. More details are given in Chapter 3.

2.3 Lognormal Regression

Linear regression is one of the most commonly used statistical methods. As a first type of regression analysis to be studied, it offers a lot of advantages. It can be implemented easily and the results are straightforwardly interpretable. As a response variable we use the transformed response variable $\ln(Y)$ defined in the previous section. The use of transformed response variable implies that the original response variable Y follows a lognormal distribution. Therefore, this model is called *lognormal model*. After the fitting, we can easily bring the predictions on the original scale using the exponential function. Following Olive (2017) and Fahrmeir, Kneib, Lang and Marx (2013) we provide a basis for this model.

Definition 2.3.1 *(Lognormal Regression Model)*
The lognormal regression model is defined as

$$\ln(Y_i) = \beta_0 + x_{i1}\beta_1 + \cdots + x_{ik}\beta_k + \varepsilon_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} + \varepsilon_i \tag{2.3.1}$$

with strictly positive random variables $Y_i, i = 1, \ldots, n$. Here n is the sample size and the random variable ε_i is the i^{th} error term . The error terms are independent and normally distributed random variables, with $\mathbb{E}[\varepsilon_i] = 0$ and constant variance $Var[\varepsilon_i] = \sigma^2$.

In matrix notation, these n equations become

$$\boldsymbol{W} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{2.3.2}$$

where $\boldsymbol{W} = (\ln(Y_1), \ldots, \ln(Y_n))^\top$ is an $n \times 1$ vector of dependent variables, $\boldsymbol{X}$ is an $n \times p$ matrix of predictors also called *design matrix*, $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown coefficients and $\boldsymbol{\varepsilon} \sim \mathcal{N}_n(\boldsymbol{0}, \boldsymbol{I}_n\sigma^2)$ is an $n \times 1$ vector of unknown error terms, where $\boldsymbol{I}_n$ is an $n \times p$ square matrix with ones on the main diagonal and zeros elsewhere, and $p = k + 1$. Equivalently,

$$\begin{bmatrix} \ln(Y_1) \\ \ln(Y_2) \\ \vdots \\ \ln(Y_n) \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots & \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}. \tag{2.3.3}$$

The $p = k + 1$ regression parameters are unknown and have to be estimated from n observations $(\boldsymbol{x}_i^\top, \ln(y_i)) = (1, x_{i1}, x_{i2}, \cdots, x_{ik}, \ln(y_i)), i = 1, \cdots, n$, where $\ln(y_i)$ are the observed values of the random variables $\ln(Y_i)$.

Parameter estimation

There are two different methods that can be used for parameter estimation for the linear regression model: least squares and maximum likelihood estimation. If the assumptions of independence, homogeneity and normality are fulfilled i.e. $\boldsymbol{\varepsilon} \sim \mathcal{N}_n(\mathbf{0}, \boldsymbol{I}_n\sigma^2)$, these two estimation techniques yield the same estimate. Here we present the *maximum likelihood (ML) estimation* (Fahrmeir, Kneib, Lang and Marx (2013)).
Assuming normally distributed error terms we have $\boldsymbol{W} \sim \mathcal{N}_n(\boldsymbol{X\beta}, \boldsymbol{I}_n\sigma^2)$, which yields the likelihood

$$L(\boldsymbol{\beta}, \sigma^2|\boldsymbol{w}) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp(-\frac{1}{2\sigma^2}(\boldsymbol{w} - \boldsymbol{X\beta})^\top(\boldsymbol{w} - \boldsymbol{X\beta})), \tag{2.3.4}$$

where $\boldsymbol{w} = (\ln(y_1), \ldots, \ln(y_n))^\top$. The log likelihood is thus given by

$$l(\boldsymbol{\beta}, \sigma^2|\boldsymbol{w}) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}(\boldsymbol{w} - \boldsymbol{X\beta})^\top(\boldsymbol{w} - \boldsymbol{X\beta}). \tag{2.3.5}$$

When maximizing the log likelihood with respect to $\boldsymbol{\beta}$, we can ignore the first two terms in Equation (2.3.5), because they are independent of $\boldsymbol{\beta}$. Maximizing $-\frac{1}{2\sigma^2}(\boldsymbol{w}-\boldsymbol{X\beta})^\top(\boldsymbol{w}-\boldsymbol{X\beta})$ is equivalent to minimizing $(\boldsymbol{w} - \boldsymbol{X\beta})^\top(\boldsymbol{w} - \boldsymbol{X\beta})$, which results in the equations

$$\boldsymbol{\beta}^\top(\boldsymbol{w} - \boldsymbol{X\beta}) = \mathbf{0}. \tag{2.3.6}$$

If $\boldsymbol{X}^\top\boldsymbol{X}$ is nonsingular, then the unique solution is given by

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{w}, \tag{2.3.7}$$

from which we immediately have

$$\hat{\boldsymbol{\beta}} \sim \mathcal{N}_n(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}^\top\boldsymbol{X})^{-1}). \tag{2.3.8}$$

The vector of fitted values is $\hat{\boldsymbol{w}} = (\widehat{\ln(y_1)}, \ldots, \widehat{\ln(y_n)})^\top = \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{w}$, where the $n \times n$ matrix $\boldsymbol{H} := \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top$ is also called *hat matrix*. Therefore, $\hat{\boldsymbol{W}} = (\widehat{\ln(Y_1)}, \ldots, \widehat{\ln(Y_n)})^\top = \boldsymbol{HW} \sim \mathcal{N}_n(\boldsymbol{X\beta}, \sigma^2 H)$. Differentiation of the log likelihood with respect to σ^2 and setting to zero, yields

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(\ln(y_i) - \widehat{\ln(y_i)})^2. \tag{2.3.9}$$

However, from Equation (2.3.9) an unbiased estimator of σ^2 can be derived:

$$s^2 = \frac{1}{n-p}\sum_{i=1}^{n}(\ln(Y_i) - \widehat{\ln(Y_i)})^2 = \frac{n}{n-p}\hat{\sigma}^2. \tag{2.3.10}$$

Under the normality assumption, it follows that

$$\frac{(n-p)s^2}{\sigma^2} = \frac{\sum_{i=1}^{n}(\ln(Y_i) - \widehat{\ln(Y_i)})^2}{\sigma^2} \sim \chi^2_{n-p}. \tag{2.3.11}$$

Hypothesis testing

To check whether a particular covariate has a significant influence on the model, we use hypothesis testing. In particular, we present our problem as

$$H_0 : \beta_j = 0 \text{ and } H_1 : \beta_j \neq 0.$$

Under the assumption that H_0 is true, from Equations (2.3.8), (2.3.11) and the Definition of t-distribution we obtain the test statistic (Fahrmeir, Kneib, Lang and Marx (2013))

$$T_j = \frac{\hat{\beta}_j}{\sqrt{s^2(\boldsymbol{X}^\top \boldsymbol{X})^{-1}_{jj}}} \overset{H_0}{\sim} t_{n-p}, \tag{2.3.12}$$

where $(\boldsymbol{X}^\top \boldsymbol{X})^{-1}_{jj}$ is the j*th* diagonal element of $(\boldsymbol{X}^\top \boldsymbol{X})^{-1}$. We define the statistical hypothesis test as follows.

Definition 2.3.2 *(Test of significance (t-test))*
Suppose we have a regression model as defined in Definition (2.3.1) and let t_j be the observed value of T_j. Then, testing the significance of one particular coefficient can be done as follows

$$\textit{Reject } H_0 : \beta_j = 0 \textit{ vs } H_1 : \beta_j \neq 0 \textit{ at level } \alpha \Leftrightarrow |t_j| > t_{n-p,1-\frac{\alpha}{2}},$$

where $t_{n-p,1-\frac{\alpha}{2}}$ is the $(1-\frac{\alpha}{2})$-quantile of the t-distribution with $n-p$ degrees of freedom.

Additionally, we are interested in testing the significance of multiple covariates simultaneously, or comparing two different models. In particular, we have the following hypothesis testing

$$H_0 : \boldsymbol{\beta}_2 = \mathbf{0} \text{ and } H_1 : \boldsymbol{\beta}_2 \neq \mathbf{0}$$

where $\boldsymbol{\beta}_2$ is a vector of length p_2 of the coefficients whose significance we want to test. We can write the design matrix as $\boldsymbol{X} = (\boldsymbol{X}_1, \boldsymbol{X}_2)$ where $\boldsymbol{X}_1$ is an $n \times (p - p_2)$ matrix and $\boldsymbol{X}_2$ is a $n \times p_2$ matrix. Therefore, we introduce the reduced model $\boldsymbol{W}_R := \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{\varepsilon}$, where $\boldsymbol{\beta}_1$ is the vector of the parameters who are not components of $\boldsymbol{\beta}_2$. Let $\boldsymbol{H}_R := \boldsymbol{X}_1(\boldsymbol{X}_1^\top \boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^\top$ be the hat matrix of the reduced model and we define the random vector of fitted values of the reduced model as $\hat{\boldsymbol{W}}_R = (\widehat{\ln(Y_1)}_R, \ldots, \widehat{\ln(Y_n)}_R)^\top = \boldsymbol{H}_R\boldsymbol{W}_R$. For the purpose of this hypothesis testing we define the test statistic (Fahrmeir, Kneib, Lang and Marx (2013))

$$F = \frac{n-p}{p_2} \frac{\sum_{i=1}^n (\ln(Y_i) - \widehat{\ln(Y_i)}_R)^2 - \sum_{i=1}^n (\ln(Y_i) - \widehat{\ln(Y_i)})^2}{\sum_{i=1}^n (\ln(Y_i) - \widehat{\ln(Y_i)})^2} \overset{H_0}{\sim} F_{p_2,n-p}. \tag{2.3.13}$$

For more details about the derivation of this test statistic, one can look in Section 3.3 in Fahrmeir (2013).

Definition 2.3.3 *(F-test)*
Suppose we have a full lognormal regression model as defined in Definition (2.3.1), and reduced lognormal regression model as defined above. Let $F(\boldsymbol{w})$ be the observed value of the test statistic F. Then, testing the significance of the coeficients $\boldsymbol{\beta}_2$ can be done as follows

$$H_0 : \boldsymbol{\beta}_2 = \mathbf{0} \text{ vs } H_1 : \boldsymbol{\beta}_2 \neq \mathbf{0} \text{ at level } \alpha \Leftrightarrow F(\boldsymbol{w}) > F_{p_2,n-p,1-\alpha},$$

where $F_{p_2,n-p,1-\alpha}$ is the $(1-\alpha)$-quantile of the F-distribution with p_2 and $n-p$ degrees of freedom.

Consequently, using the F-test we can compare the two models, the full one and reduced one. When the null hypothesis is rejected, we can say that at least one of the covariates which are not included in the reduced model is significant.

Handling of covariates

Before we proceed with model diagnostics and evaluation, we want to give a deeper explanation on how we handle different types of covariates, using Section 3.3 of Olive (2017) and Section 3.1 of Fahrmeir, Kneib, Lang and Marx (2013). In general, we differentiate between two different types of covariates, quantitative and qualitative. Let a qualitative also called categorical covariate has c different categories $a_1, \ldots, a_c$. Then this factor is incorporated into the regression model by using c-1 indicator variables

$$x_j = \begin{cases} 1, & \text{if the observation is in category } a_j \\ 0, & \text{if the observation is not in category } a_j \end{cases}$$

for $j = 1, \ldots, c-1$, where one of the levels a_j is ommited.
On the other hand, if the quantitative also called continuous covariate shows nonlinear relationship with the response variable, we can either transform it or do polynomial regression. Common transformation functions are $f(x) = \ln(x)$, $f(x) = \sqrt{x}$ in case of a positive covariate, as well as $f(x) = \frac{1}{x}$. A good indicator of a proper transformation for a continuous covariate is when the scatterplot of the transformed covariate against the response shows linear relationship. If the nonlinear relationship of the continuous covariate x and the response variable looks polynomial, then we fit polynomial regression i.e. we include the covariates $x, x^2, \ldots, x^l$ in the model instead of x, where l is the degree of the polynomial we choose. In this thesis, we use orthogonal polynomials, since the fitting of these polynomials has proven more stable numerically.
Apart from the quantitative and qualitative covariates which are also called main effects, we also define interactions. An interaction between covariates exists if the effect of a covariate on the response depends on the value of other covariate. In the following we present the modelling of interaction in three different cases:

- Let x and z be two categorical variables with c and m categories respectively. Let us choose the last category as the reference category for both x and z and we denote the respective dummy variables as $x_j, j = 1, \ldots, c-1$ and $z_k, k = 1, \ldots, m-1$. For modelling the interaction effect we have to consider all possible combinations of the values of x and z (with the exception of the reference categories), specifically $x_1z_1, \ldots, x_1z_{m-1}, x_2z_1, \ldots, x_{c-1}z_{m-1}$.

- Let x be continuous and z be categorical variable, with m categories. As before, we choose the last category of z as reference category and denote the dummy variables $z_k, k = 1, \ldots, m-1$. In this case, for modelling the interaction effect of x and z, we need to include $xz_1, \ldots, xz_{m-1}$ in our model.

- Let x and z be continuous variables. For modelling the interaction effect of x and z, we need to include xz in our model.

Important remark is that when the lognormal regression model contains a covariate with power l, then all lower degrees of the covariate should be included in the model. Additionally, when the lognormal regression model contains interactions, then all corresponding main effects should be in the model as well.

Model selection criteria

When fitting a lognormal regression model we can have a lot of variables to choose from. What we want is our chosen model to contain all influential covariates, but to discard all insignificant variables. In this section, we define our model selection criteria used in the thesis, some of which are more rigid towards the complexity of the model. All of the defined criteria can be found in Fahrmeir, Kneib, Lang and Marx (2013). First, we introduce the *coefficient of determination* R^2.

Definition 2.3.4 *(Coefficient of determination)*
Let us have a regression model as defined in Definition 2.3.1. Then the coefficient of determination R^2 is defined as

$$R^2 := \frac{\sum_{i=1}^{n}(\widehat{\ln(y_i)} - \bar{w})^2}{\sum_{i=1}^{n}(\ln(y_i) - \bar{w})^2}, \tag{2.3.14}$$

where $\ln(y_i), i = 1, \ldots, n$ are the observed values of the random variables $\ln(Y_i), i = 1, \ldots, n$ respectively, $\widehat{\ln(y_i)}, i = 1, \ldots, n$ are the fitted values and $\bar{w} = \frac{1}{n}\sum_{i=1}^{n}\ln(y_i)$. It holds that $0 \leq R^2 \leq 1$.

R^2 is usually used as a goodness of fit measure. We can interpret it the following way: the closer R^2 is to 1, the fit to the data is better. What R^2 tells us, is what proportion of the variability of the data is explained by our model. However, the R^2 is not appropriate for model comparison, since the coefficient of determination will always increase with the addition of a new covariate into the model. The *adjusted coefficient of determination* is an alternative criterion, which accounts for the number of parameters in the model.

Definition 2.3.5 *(Adjusted coefficient of determination)*
Let us have a regression model as defined in Definition 2.3.1. Then the adjusted coefficient of determination R^2_{adj} is defined as

$$R^2_{adj} := 1 - \frac{n-1}{n-p}(1 - R^2). \tag{2.3.15}$$

In the following we introduce two additional model selection criteria, which are widely used for model choice within the scope of likelihood based inference.

Definition 2.3.6 *(Akaike Information Criterion)*
We define the Akaike information criterion AIC as

$$AIC := -2l(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\sigma}}^2 | \boldsymbol{w}) + 2q, \tag{2.3.16}$$

where $l(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\sigma}}^2|\boldsymbol{w})$ is the maximum value of the log likelihood, since the estimated parameters are inserted and q is the number of parameters in the model. For the lognormal model $q = p + 1$.

Definition 2.3.7 *(Bayesian Information Criterion)*
We define the Bayesian information criterion BIC as

$$BIC := -2l(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\sigma}}^2|\boldsymbol{w}) + \ln(n)q, \tag{2.3.17}$$

where $l(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\sigma}}^2|\boldsymbol{w})$ is the maximum value of the log likelihood, since the estimated parameters are inserted and q is the number of parameters in the model. For the lognormal model $q = p + 1$.

For both these criteria, AIC and BIC, smaller value indicates a better model fit. The main difference between them is that BIC penalizes the complex models much more than AIC. Addittionally, for variable subset selection we use *backward elimination*. This procedure starts with the full model containing all potential covariates. Subsequently, in every iteration, the covariate which provides the greatest reduction of the model choice criteria (AIC or BIC) is eliminated from the model. The algorithm stops when no further reduction is possible.

Model diagnostics and evaluation

Once we have our chosen model, we can use several graphical tools for assessment of the assumptions of the lognormal regression model. For that purpose, using Fahrmeir, Kneib, Lang and Marx (2013) we introduce the following definitions.

Definition 2.3.8 *(Raw residuals)*
Let $\ln(y_i), i = 1, \ldots, n$ be the observed values and $\widehat{\ln(y_i)}, i = 1, \ldots, n$ be the fitted values of $\ln(Y_i), i = 1, \ldots, n$ respectively. Then, the raw residuals are defined as

$$r_i := \ln(y_i) - \widehat{\ln(y_i)}, \quad i = 1, \ldots, n. \tag{2.3.18}$$

Definition 2.3.9 *(Internally studentized residuals)*
We define internally studentized residuals as

$$s_i := \frac{r_i}{s\sqrt{1 - h_{ii}}}, \quad i = 1, \ldots, n, \tag{2.3.19}$$

where r_i is given in Equation (2.3.18), s in Equation (2.3.10) and h_{ii} is the ith diagonal element of the hat matrix $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top$.

To check the model assumptions we can plot the internally studentized residuals $s_i, i = 1, \ldots, n$ against the observation number. This plot should show random fluctuation around the zero with constant variance, since under the model assumptions these residuls are $\mathcal{N}(0, 1)$ distributed. Additionally, we can check the distributional assumptions using Q-Q plot, where the empirical quantiles are compared to the quantiles of the theoretical distribution. If the data follows the distribution, the points should closely follow the 45° bisecting line.
Now we define leverage and Cook's distance which are helpful for detecting outliers and influential observations.

Definition 2.3.10 *(Leverage)*
The ith leverage

$$h_{ii} := \boldsymbol{H}_{ii} \tag{2.3.20}$$

is the ith diagonal element of the hat matrix $\boldsymbol{H}$.

It is shown that one can use $h_{ii} > \frac{2p}{n}, i = 1, \ldots, n$ as a benchmark for high leverage points. We introduce the Cook's distance as defined in Olive (2017), p.131.

Definition 2.3.11 *(Cook's Distance)*
We define Cook's distance as

$$D_i := \frac{s_i^2}{p} \frac{h_{ii}}{1 - h_{ii}}, \quad i = 1, \ldots, n, \tag{2.3.21}$$

where s_i *are defined with Equation (2.3.19) and* h_{ii} *with Equation (2.3.20).*

Observations with Cook's distance larger than 1 should be studied and if necessary, removed.

2.4 Gamma regression

After introducing lognormal regression, we present another useful regression for modelling strongly positive response. Namely, the Gamma regression model falls into a group of models called *generalized linear models (GLMs)* (first introduced by Nelder and Wedderburn (1972)), which are regression models for non-normal response variables with certain common properties. This class of models is extensively studied in McCullagh and Nelder (1989).

Gamma regression model as a GLM

Definition 2.4.1 *(Components of a generalized linear model)*

1. *Random Component: Responses* $Y_i, i = 1, \ldots, n$ *are independent with probability density function or probability mass fuction from the exponential family with parameter* θ *and* $\phi > 0$ *given by*

$$f(y|\theta, \phi) = \exp\left\{\frac{\theta y - b(\theta)}{a(\phi)} + c(y, \phi)\right\}, \tag{2.4.1}$$

   where ϕ *is a dispersion parameter,* θ *is called canonical parameter and the functions* $b(\cdot), a(\cdot)$ *and* $c(\cdot, \cdot)$ *are known.*

2. *Systematic Component: The quantity*

$$\eta(\boldsymbol{\beta}) := \boldsymbol{x}_i^\top \boldsymbol{\beta} = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} \tag{2.4.2}$$

   is called the linear predictor and $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_k)^\top$ *are p unknown regression parameters.*

3. Parametric Link Component: The link function

$$g(\mu_i) = \eta(\boldsymbol{\beta}) = \boldsymbol{x}_i^\top \boldsymbol{\beta} \tag{2.4.3}$$

defines the relationship between the linear predictor η_i and the mean μ_i of Y_i.

Many known distributions are members of exponential family class, one of which is the gamma distribution. For the moments of a random variable Y with exponential family distribution, it holds that

$$\begin{aligned} \mathbb{E}(Y) &= b'(\theta) \\ Var(Y) &= b''(\theta)a(\phi), \end{aligned} \tag{2.4.4}$$

where the functions $b(\cdot), a(\cdot)$ are defined in Definition 2.4.1. To show that the gamma distribution indeed belongs to the exponential family class, let us have a random variable $Y \sim Gamma(\nu, \nu/\mu)$, whose density is defined with Equation (2.1.5).
Then $\theta = -\frac{1}{\mu},\ b(\theta) = -\ln(-\theta), \phi = \frac{1}{\nu}, a(\phi) = \phi$ and $c(y,\phi) = \frac{1-\phi}{\phi}\ln(y) - \frac{\ln(\phi)}{\phi} - \ln\left(\Gamma(\frac{1}{\phi})\right)$.
Using the Equations (2.4.4) we derive

$$\mathbb{E}(Y) = b'(\theta) = -\frac{1}{\theta} = \mu, \quad Var(Y) = b''(\theta)a(\phi) = \frac{1}{\theta^2}\frac{1}{\nu} = \frac{\mu^2}{\nu}.$$

For more details regarding this derivation, one can look at De Jong and Heller (2008). As a link function in gamma regression, we use the *log link function* i.e. $g(\mu) = \ln(\mu)$. Therefore, for exploratory data analysis in gamma regression we plot x_{ij} versus $\ln(y_i)$ for $i = 1, \ldots, n$. If the resulting plot looks linear for all $j = 1, \ldots, k$, then the link function is appropriate. We handle covariates the same way as in lognormal regression.

Parameter estimation

Given n observations of n independent random variables $Y_i \sim Gamma(\nu, \nu/\mu_i)$ and a $n \times p$ matrix of predictors $\boldsymbol{X} = (x_{ij})_{i=1,\ldots,n, j=1,\ldots,p}$, we define the *likelihood* with

$$L(\boldsymbol{y}|\boldsymbol{\theta}, \phi) = \prod_{i=1}^{n} f_i(y_i|\theta_i, \phi), \tag{2.4.5}$$

where the functions $f_i, i = 1, \ldots, n$ are defined with Equation (2.4.1). Therefore, the *log likelihood* is defined with

$$l(\boldsymbol{\mu}, \phi|\boldsymbol{y}) = \sum_{i=1}^{n} \ln f_i(y_i|\theta_i, \phi), \tag{2.4.6}$$

where $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)^\top$. The *maximum likelihood estimator* $\hat{\boldsymbol{\beta}}$ maximizes the log likelihood. The maximum-likelihood equations for $\beta_j, j = 1, \ldots, p$ are given by

$$\sum_{i=1}^{n} w_i(y_i - \mu_i)\frac{d\eta_i}{d\mu_i}x_{ij} = 0, \quad j = 1, \ldots, p, \tag{2.4.7}$$

where $w_i = \left[V_i(\frac{d\eta_i}{d\mu_i})^2\right]^{-1}$ with $V_i = b''(\theta_i)$ for $i = 1, \ldots, n$. For more details one can refer to McCullagh and Nelder (1989). However, an important realization is that when

$a(\phi) = \phi$, the dispersion parameter disappears, as is the case in the gamma regression model. *Fisher's scoring method* uses the gradient vector $\frac{\partial l}{\partial \boldsymbol{\beta}} = \boldsymbol{u}$, where l denotes the log likelihood, and minus the expected value of the Hessian matrix

$$-\mathbb{E}\Big(\frac{\partial^2 l}{\partial \beta_r \partial \beta_s}\Big)_{r,s\in\{1,\dots,p\}} = \boldsymbol{A}. \tag{2.4.8}$$

Now given the current estimate $\boldsymbol{b}$ of $\boldsymbol{\beta}$, the new estimate $\boldsymbol{b}^\star = \boldsymbol{b} + \delta\boldsymbol{b}$ of $\boldsymbol{\beta}$ satisfies the equation

$$\boldsymbol{A}\boldsymbol{b}^\star = \boldsymbol{A}\boldsymbol{b} + \boldsymbol{u} \tag{2.4.9}$$

Since $(\boldsymbol{A}\boldsymbol{b})_j = \sum_{i=1}^n w_i(\boldsymbol{b}) x_{ij}\eta_i$, we obtain that the new estimate $\boldsymbol{b}^\star$ satisfies

$$(\boldsymbol{A}\boldsymbol{b}^\star)_j = \sum_{i=1}^n w_i(\boldsymbol{b}) x_{ij}\{\eta_i + (y_i - \mu_i)\frac{d\eta_i}{d\mu_i}\}, \tag{2.4.10}$$

where $w_i(\boldsymbol{b}), i = 1,\dots,n$ are defined as in Equation (2.4.7). These equations have the form of linear weighted least-squares equations with weight $\boldsymbol{W}(\boldsymbol{b}) = diag(w_1(\boldsymbol{b}),\dots,w_n(\boldsymbol{b}))$ and dependent variable $\boldsymbol{z}(\boldsymbol{b}) = (z_1(\boldsymbol{b}),\dots,z_n(\boldsymbol{b}))^\top$ where

$$z_i(\boldsymbol{b}) = \eta_i + (y_i - \mu_i)\frac{d\eta_i}{d\mu_i}. \tag{2.4.11}$$

According to the stopping criterion, the algorithm typically converges close to a maximum after a number of iterative steps.
After we obtain the maximum likelihood estimator $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$, we can estimate the dispersion parameter ϕ using moments, as suggested by McCullagh and Nelder (1989). Namely, from $Var(Y_i) = \phi b''(\theta_i)$ and denoting by $v(\mu_i) = b''(\theta_i)$ the so-called variance function, as well as using the implicit dependence of $b''(\theta_i)$ on μ_i since $b'(\theta_i) = \mu_i$ for $i = 1,\dots,n$ we can derive a consistent estimator for the dispersion parameter in gamma regression by

$$\hat{\phi} = \frac{1}{n-p}\sum_{i=1}^n \frac{(Y_i - \hat{\mu}_i)^2}{v(\hat{\mu}_i)} = \frac{1}{n-p}\sum_{i=1}^n \Big(\frac{Y_i - \hat{\mu}_i}{\hat{\mu}_i}\Big)^2, \tag{2.4.12}$$

where $\hat{\mu}_i = g^{-1}(\boldsymbol{x}_i^\top\hat{\boldsymbol{\beta}}), i = 1,\dots,n$.

Measuring the goodness of fit

The main goodness of fit we will be concerned with is that formed from the logarithm of a ratio of likelihoods, called the *deviance*. For the theory in this section we consult McCullagh and Nelder (1989). Given n observations we can fit a model to them containing up to n parameters. The full model has n parameters, one per observation, thus the $\mu_i, i = 1,\dots,n$ derived from it match the data exactly. Although this model is uninformative since it only repeats the data, it gives us a baseline for a goodness of a fit measure for a model with p parameters. Let $l(\hat{\boldsymbol{\mu}}, \phi; \boldsymbol{y})$ be the log likelihood maximized over $\boldsymbol{\beta}$ with length p for a fixed value of the dispersion parameter ϕ. The maximum log likelihood achievable in a full model with n parameters is $l(\boldsymbol{y}, \phi; \boldsymbol{y})$. Then we define the *scaled deviance* as

$$D(\hat{\boldsymbol{\mu}}, \boldsymbol{y}; \phi) := -2\big[l(\hat{\boldsymbol{\mu}}, \phi; \boldsymbol{y}) - l(\boldsymbol{y}, \phi; \boldsymbol{y})\big] \tag{2.4.13}$$

and the *deviance* as

$$D(\hat{\boldsymbol{\mu}}, \boldsymbol{y}) := \phi D(\hat{\boldsymbol{\mu}}, \boldsymbol{y}; \phi). \tag{2.4.14}$$

In particular, the deviance of a gamma regression model is given by

$$D(\hat{\boldsymbol{\mu}}, \boldsymbol{y}) = -2 \sum_{i=1}^{n} \Big[\ln\big(\frac{y_i}{\hat{\mu}_i}\big) - \frac{y_i - \hat{\mu}_i}{\hat{\mu}_i} \Big]. \tag{2.4.15}$$

Criteria for model selection

Following De Jong and Heller (2008), we define two criteria for model selection for generalized linear model. To assess a model fit, we can use the *residual deviance test*. We can use the *deviance statistic* to test the null hypothesis H_0 that the model assumptions of the specified generalized linear model are satisfied, i.e. $\eta_i = \boldsymbol{x}_i^\top \hat{\beta}_i = g(\mu_i), i = 1, \ldots, n$ for a specified link function g. The alternative hypothesis is that the specified GLM is not a good fit for the data. The asymptotic distribution of the deviance is

$$\frac{D(\hat{\boldsymbol{\mu}}, \boldsymbol{Y})}{\phi} \xrightarrow{\mathcal{D}} \chi^2_{n-p} \text{ as } n \to \infty \tag{2.4.16}$$

when $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^\top$ follows the generalized linear model with deviance $D(\hat{\boldsymbol{\mu}}, \boldsymbol{Y})$ and dispersion parameter ϕ. Then we obtain

$$\text{Reject } H_0 \text{ at level } \alpha \Leftrightarrow \tfrac{D(\hat{\boldsymbol{\mu}}, \boldsymbol{Y})}{\hat{\phi}} > \chi^2_{n-p,1-\alpha}$$

where $\hat{\phi}$ is an estimate of ϕ and $\chi^2_{n-p,1-\alpha}$ is the $(1-\alpha)$-quantile of χ^2 distribution with $n-p$ degrees of freedom.
To compare two nested models i.e. to test the hypothesis $H_0 : \boldsymbol{\beta}_2 = \boldsymbol{0}$ against $H_1 : \boldsymbol{\beta}_2 \neq \boldsymbol{0}$ where $\boldsymbol{\beta}_2$ is a subvector of $\boldsymbol{\beta}$ we can use the *partial deviance test*. This corresponds to comparing two models, namely, the reduced one defined by $\boldsymbol{\beta}_2 = \boldsymbol{0}$ and the full model. We will use the subscripts F and R to denote the estimates of the full and reduced model respectively. Then we obtain

$$\text{Reject } H_0 \text{ at level } \alpha \Leftrightarrow \tfrac{D(\hat{\boldsymbol{\mu}}_R, \boldsymbol{y}) - D(\hat{\boldsymbol{\mu}}_F, \boldsymbol{y})}{\hat{\phi}_F} > \chi^2_{p_2,1-\alpha}$$

where p_2 is the length of $\boldsymbol{\beta}_2$ and $\chi^2_{p_2,1-\alpha}$ is the $(1-\alpha)$-quantile of χ^2 distribution with p_2 degrees of freedom.
Additionally, as a model criteria we can use AIC and BIC defined with Equations (2.3.16) and (2.3.17) where we can plug in the maximized log likelihood of the gamma regression model and $q = p + 1$.

2.5 Copulas

In this section we introduce copulas, a popular model class for analysing multivariate data (Czado (2019)). Their popularity rose due to the separation of margins and dependence by copula approach, tail asymmetries and separate multivariate component modelling. We want to characterize the dependence between random variables with common marginal distribution given by the uniform distribution. Thus, copulas are useful tool to separate the dependence between the components from the marginal distributions. Referring to Czado (2019), we build a foundation for the theory of copulas.

Concept of Copulas

Definition 2.5.1 *(Probability integral transform (PIT))*
If $X \sim F$ is a continuous random variable and x is an observed value of X, then the transformation $u := F(x)$ is called the probability integral transform (PIT) at x.

Moreover, if $X \sim F$ then $U := F(X)$ is uniformly distributed, since

$$P(U \leq u) = P(F(X) \leq u) = P(X \leq F^{-1}(u)) = F(F^{-1}(u)) = u$$

holds for every $u \in [0,1]$. Therefore, we can apply the PIT on our vector of interest $\boldsymbol{X} = (X_1, \cdots, X_d)^\top$ and obtain uniformly distributed data $\boldsymbol{U} = (U_1, \ldots, U_d)^\top$, where

$$U_i = F_i(X_i) \quad \text{for} \;\; i = 1, ..., d. \tag{2.5.1}$$

$\boldsymbol{U} = (U_1, \cdots, U_d)^\top$ is called *u-scale* or *copula scale* data.

Definition 2.5.2 *(Copula)*
A d-dimensional copula C is a multivariate distribution function on the d-dimensional hypercube $[0,1]^d$ with uniformly distributed marginals.

Definition 2.5.3 *(Copula density)*
The corresponding copula density for an absolutely continuous copula is denoted by c and can be obtained by partial differentiation, i.e.

$$c(u_1, \ldots, u_d) := \frac{\partial^d}{\partial u_1 \ldots \partial u_d} C(u_1, \ldots, u_d) \tag{2.5.2}$$

for all $\boldsymbol{u}$ in $[0,1]^d$.

One of the fundamental theorems is proven by Sklar (1959). It allows representation of multivariate distributions in terms of their marginal distributions and a corresponding copula.

Theorem 2.5.1 *(Sklar's Theorem)*
Let $\boldsymbol{X}$ be a d-dimensional random vector with joint distribution function F and marginal distribution functions F_i, $i = 1, \ldots, d$, then the joint distribution function can be expressed as

$$F(x_1, \ldots, x_d) = C(F_1(x_1), \ldots, F_d(x_d)), \tag{2.5.3}$$

with associated density or probability mass function

$$f(x_1, \ldots, x_d) = c(F_1(x_1), \ldots, F_d(x_d)) f_1(x_1) \ldots f_d(x_d) \tag{2.5.4}$$

for some d-dimensional copula C with copula density c. For absolutely continuous distributions, the copula C is unique.
The inverse also holds: the copula corresponding to a multivariate distribution function F with marginal distribution functions F_i, $i = 1, \ldots, d$ can be expressed as

$$C(u_1, \ldots, u_d) = F(F_1^{-1}(u_1), \ldots, F_d^{-1}(u_d)) \tag{2.5.5}$$

and its copula density or probability mass function is determined by

$$c(u_1, \ldots, u_d) = \frac{f(F_1^{-1}(u_1), \ldots, F_d^{-1}(u_d))}{f_1(F_1^{-1}(u_1)) \ldots f_d(F_d^{-1}(u_d))}. \tag{2.5.6}$$

Proof. The proof can be found in the book by Nelsen (2006) in Section 2.3.

Lemma 2.5.1 *(Conditional densities and distribution functions of bivariate distributions in terms of their copula)*
The conditional density and distribution function can be rewritten as

$$f_{1|2}(x_1|x_2) = c_{12}(F_1(x_1), F_2(x_2))f_2(x_2) \tag{2.5.7}$$

$$\begin{aligned} F_{1|2}(x_1|x_2) &= \frac{\partial}{\partial u_2} C_{12}(F_1(x_1), u_2)|_{u_2=F_2(x_2)} \\ &=: \frac{\partial}{\partial F_2(x_2)} C_{12}(F_1(x_1), F_2(x_2)). \end{aligned} \tag{2.5.8}$$

Proof. The proof can be found in the book by Czado (2019) on page 20.

Remark 2.5.1 *Lemma 2.5.1 can also be applied to the bivariate copula distribution C_{12}. In particular, it follows that*

$$C_{1|2}(u_1|u_2) = \frac{\partial}{\partial u_2} C_{12}(u_1, u_2), \ \forall u_1 \in [0,1]. \tag{2.5.9}$$

The relationship between $F_{1|2}$ and $C_{1|2}$ using Lemma 2.5.1 is therefore given by

$$F_{1|2}(x_1|x_2) = \frac{\partial}{\partial u_2} C_{12}(F_1(x_1), u_2)|_{u_2=F_2(x_2)} = C_{1|2}(F_1(x_1)|F_2(x_2)). \tag{2.5.10}$$

Applying (2.5.10) yields a relationship among the inverse function of the conditional distribution functions:

$$F_{1|2}^{-1}(u_1|x_2) = F_1^{-1}(C_{1|2}^{-1}(u_1|F_2(x_2)). \tag{2.5.11}$$

The conditional distribution function $C_{1|2}$ associated with a copula is denoted as an *h-function.*

Definition 2.5.4 *(h-functions of bivariate copulas)*
The h functions corresponding to a bivariate copula C_{12} are defined for all $(u_1, u_2) \in [0,1]^2$ as

$$h_{1|2}(u_1|u_2) := \frac{\partial}{\partial u_2} C_{12}(u_1, u_2) \tag{2.5.12}$$

$$h_{2|1}(u_2|u_1) := \frac{\partial}{\partial u_1} C_{12}(u_1, u_2). \tag{2.5.13}$$

Dependence measures

To capture dependence between two random variables we introduce the measure *Kendall's tau* and the notion of *tail dependence* (Czado (2019)). The advantage of the Kendall's tau measure is that it is rank-based and therefore invariant with monotone transformations of the marginals. Additionally, it can be expressed solely in terms of the associated copula, so its value does not depend on the marginal distributions.

Definition 2.5.5 *(Kendall's tau)*
The Kendall's τ between the continuous random variables X_1 and X_2 is defined as

$$\tau(X_1, X_2) = P((X_{11} - X_{21})(X_{12} - X_{22}) > 0) - P((X_{11} - X_{21})(X_{12} - X_{22}) < 0), \quad (2.5.14)$$

where (X_{11}, X_{12}) and (X_{21}, X_{22}) are independent and identically distributed copies of (X_1, X_2).

For the estimation of Kendall's τ we define concordance and discordance.

Definition 2.5.6 *(Concordant discordant, and extra pairs)*
The pair $(\boldsymbol{x}_i, \boldsymbol{x}_j)$ where $\boldsymbol{x}_i = (x_{i1}, x_{i2})$ and $\boldsymbol{x}_j = (x_{j1}, x_{j2})$ is called

- *concordant if the ordering in $\boldsymbol{x}^1 := (x_{i1}, x_{j1})$ is the same as in $\boldsymbol{x}^2 := (x_{i2}, x_{j2})$ i.e., $x_{i1} < x_{j1}$ and $x_{i2} < x_{j2}$ holds or $x_{i1} > x_{j1}$ and $x_{i2} > x_{j2}$ holds,*
- *discordant if the ordering in x^1 is opposite to the ordering of x^2, i.e. $x_{i1} < x_{j1}$ and $x_{i2} > x_{j2}$ holds or $x_{i1} > x_{j1}$ and $x_{i2} < x_{j2}$ holds,*
- *extra x_1 pair if $x_{i1} = x_{j1}$ holds,*
- *extra x_2 pair if $x_{i2} = x_{j2}$ holds.*

Definition 2.5.7 *(Estimate of Kendall's τ)*
Let N_c be the number of the concordant pairs, N_d be the number of the discordant pairs, N_1 be the number of extra x_1 pairs, and N_1 be the number of extra x_2 pairs of random sample $x_{i1}, x_{i2}, i = 1, \ldots, n$ from the joint distribution of (X_1, X_2). Then an estimate of Kendall's τ is given by

$$\hat{\tau}_n^\star := \frac{N_c - N_d}{\sqrt{N_c + N_d + N_1} \times \sqrt{N_c + N_d + N_2}}. \quad (2.5.15)$$

Theorem 2.5.2 *(Kendall's τ expressed in terms of the copula)*
Let X_1 and X_2 be two continuous random variables, then Kendall's τ can be expressed as

$$\tau = 4 \int_{[0,1]^2} C(u_1, u_2) dC(u_1, u_2) - 1. \quad (2.5.16)$$

Proof. The proof can be found in the book by Czado (2019) on page 32.
Finally, we define tail dependence coefficients, which are helpful for analysis of joint extreme events.

Definition 2.5.8 *(Upper and lower tail dependence coefficients)*
The upper tail dependence coefficient of a bivariate distribution with copula C is defined as

$$\begin{aligned} \lambda^{upper} &= \lim_{t \to 1^-} P(X_2 > F_2^{-1}(t) | X_1 > F_1^{-1}(t)) \\ &= \lim_{t \to 1^-} \frac{1 - 2t + C(t, t)}{1 - t}, \end{aligned} \quad (2.5.17)$$

while the lower tail dependence coefficient is

$$\begin{aligned} \lambda^{lower} &= \lim_{t \to 0^+} P(X_2 \leq F_2^{-1}(t) | X_1 \leq F_1^{-1}(t)) \\ &= \lim_{t \to 0^+} \frac{C(t, t)}{t}. \end{aligned} \quad (2.5.18)$$

Bivariate copulas

After defining copulas, we introduce different construction approaches of bivariate copulas. First we define *parametric copulas* using Czado (2019) and Joe (2014). One way of constructing parametric copulas is applying the elliptical distributions to each margin, which yields the class of *elliptical copulas* (Czado (2019)). Another class of parametric copulas is *Archimedian copulas* (Czado (2019)), which is obtained with help of generator functions. Additionally, we define the *BB copulas* (Joe (2014)), which depend on two parameters.

Example 2.5.1 *(Elliptical copulas)*

- *Bivariate Gaussian copula*
  We denote by $\Phi(\cdot)$ the distribution function of a univariate standard normal distribution with mean 0 and variance 1, and by $\Phi_2(\cdot,\cdot;\rho)$ the distribution function of a bivariate standard normal distribution with zero means, unit variances and correlation ρ. By applying Equation (2.5.5), the inverse statement of Sklar's Theorem, we obtain the bivariate Gaussian copula:

$$C(u_1,u_2;\rho)=\Phi_2(\Phi^{-1}(u_1),\Phi^{-1}(u_2);\rho). \tag{2.5.19}$$

- *Bivariate Student t copula*
  We denote by $T_\upsilon(\cdot)$ the distribution function of a univariate Student's t distribution with mean 0 and $\upsilon>0$ degrees of freedom, and by $T_{2,\upsilon}(\cdot,\cdot;\rho)$ the distribution function of a bivariate Student's t distribution with zero means, $\upsilon>0$ degrees of freedom and correlation ρ. Again, by applying Equation (2.5.5), the inverse statement of Sklar's Theorem, we obtain the bivariate Student t copula:

$$C(u_1,u_2;\upsilon,\rho)=T_{2,\upsilon}(T_\upsilon^{-1}(u_1),T_\upsilon^{-1}(u_2);\rho). \tag{2.5.20}$$

Example 2.5.2 *(Bivariate Archimedean copulas with a single parameter)*

- *Clayton copula*

$$C(u_1,u_2)=(u_1^{-\delta}+u_2^{-\delta}-1)^{-\frac{1}{\delta}}, \tag{2.5.21}$$

  where $0<\delta<\infty$ control the degree of dependence. Full dependence is obtained when $\delta\to\infty$. Independence corresponds to $\delta\to 0$.

- *Gumbel copula*

$$C(u_1,u_2)=\exp[-\{(-\ln u_1)^\delta+(-\ln u_2)^\delta\}^{\frac{1}{\delta}}], \tag{2.5.22}$$

  where $\delta\geq 1$ is the parameter of dependence. For $\delta\to\infty$ we have full dependence, while $\delta=1$ corresponds to independence.

- *Frank copula*

$$C(u_1,u_2)=-\frac{1}{\delta}\ln\left(\frac{1}{1-e^{-\delta}}[(1-e^{-\delta})-(1-e^{-\delta u_1})(1-e^{-\delta u_2})]\right), \tag{2.5.23}$$

  where the parameter $\delta\in[-\infty,\infty]\setminus\{0\}$. When $\delta\to 0^+$ we obtain the independence copula.

- *Joe copula*

$$C(u_1, u_2) = 1 - \Big((1-u_1)^\delta + (1-u_2)^\delta - (1-u_1)^\delta(1-u_2)^\delta\Big)^{\frac{1}{\delta}}, \tag{2.5.24}$$

where $\delta \geq 1$. The independence copula corresponds to $\delta = 1$.

Example 2.5.3 *(Independence copula)*
From the definitions of different Archemedean bivariate copulas, we obtain the independence copula:

$$C(u_1, u_2) = u_1 u_2. \tag{2.5.25}$$

Example 2.5.4 *(Bivariate BB copulas)*

- *BB1 copula*

$$C(u, v; \theta, \delta) = \big\{1 + [(u^{-\theta} - 1)^\delta + (v^{-\theta} - 1)^\delta]^{\frac{1}{\delta}}\big\}^{-\frac{1}{\theta}}, \tag{2.5.26}$$

where $\theta > 0$ and $\delta \geq 1$. For $\theta \to 0^+$ and $\delta \to 1^+$ the independence copula arises.

- *BB6 copula*

$$C(u, v; \theta, \delta) = 1 - \Big(1 - \exp\Big\{ - [(-\ln(1-\bar{u}^\theta))^\delta + (-\ln(1-\bar{v}^\theta))^\delta]^{\frac{1}{\delta}}\Big\}\Big)^{\frac{1}{\theta}}, \tag{2.5.27}$$

where $\bar{u} = 1 - u$, $\bar{v} = 1 - v$, $\theta \geq 1$ and $\delta \geq 1$.

- *BB7 copula*

$$C(u, v; \theta, \delta) = 1 - \Big(1 - [(1-(1-u)^\theta)^{-\delta} + (1-(1-v)^\theta)^{-\delta} - 1]^{-\frac{1}{\delta}}\Big)^{\frac{1}{\theta}}, \tag{2.5.28}$$

where $\theta \geq 1$ and $\delta > 0$. The independence copula corresponds to $\theta = 1$ and $\delta = 0$.

- *BB8 copula*

$$C(u, v; \theta, \delta) = \delta^{-1}\Big(1 - \{1 - \eta^{-1}[1-(1-\delta u)^\theta][1-(1-\delta v)^\theta]\}^{\frac{1}{\theta}}\Big), \tag{2.5.29}$$

where $\eta = 1 - (1-\delta)^\theta$, $\theta \geq 1$ and $0 < \delta \leq 1$.

Example 2.5.5 *(Rotated copulas)*
In order to extend the range of dependence, using Czado (2019) we introduce clockwise rotations of the copula density $c(\cdot, \cdot)$ by

- $90° : c_{90}(u_1, u_2) := c(1 - u_1, u_2)$,
- $180° : c_{180}(u_1, u_2) := c(1 - u_1, 1 - u_2)$, *and*
- $270° : c_{270}(u_1, u_2) := c(u_1, 1 - u_2)$.

Next, we introduce a *nonparametric* construction approach of bivariate copulas (Nagler (2014)).

Example 2.5.6 *(Local likelihood transformation estimator)*
Let $U_i, V_i, i = 1, \dots, n$ be pseudo data. We define $\boldsymbol{W}_i = (U_i, V_i)$ and $\triangle_{k_n}(x, y)$ as the euclidean distance between (x, y) and its k_nth closest observation amongst all $\big(\Phi^{-1}(\boldsymbol{W}_i)\big)_{i=1,\dots,n} := \big(\Phi^{-1}(U_i), \Phi^{-1}(V_i)\big)_{i=1,\dots,n}$. For all $(u, v) \in [0, 1]^2$, the local likelihood transformation estimator of a copula density $c(u, v)$ with nearest-neighbor factor $\triangle_{k_n}$ and bandwidth matrix B is given by

$$\hat{c}^{(TLL)}(u, v) = \frac{\exp\big\{\hat{a}_1\big(\Phi^{-1}(u), \Phi^{-1}(v)\big)\big\}}{\phi(\Phi^{-1}(u))\phi(\Phi^{-1}(v))}, \tag{2.5.30}$$

where $\hat{a}_1\big(\Phi^{-1}(u), \Phi^{-1}(v)\big)$ can be found via

$$\begin{aligned}
&\hat{\boldsymbol{a}}(x, y) = \\
&arg\max_{\boldsymbol{a}\in\mathbb{R}^6}\Bigg\{ \sum_{i=1}^{n} K_{\triangle_{k_n}(x,y)B}\big((x, y) - \Phi^{-1}(\boldsymbol{W}_i)\big) P_{\boldsymbol{a},2}\big((x, y) - \Phi^{-1}(\boldsymbol{W}_i)\big) \\
&\qquad - n\int_{\mathbb{R}^2} K_{\triangle_{k_n}(x,y)B}\big((x, y) - \Phi^{-1}(\boldsymbol{z})\big) \exp\big\{P_{\boldsymbol{a},2}\big((x, y) - \Phi^{-1}(\boldsymbol{z})\big)\big\} d\boldsymbol{z}\Bigg\}
\end{aligned} \tag{2.5.31}$$

with

$$\begin{aligned}
P_{\boldsymbol{a}(x,y),2}(x', y') =& a_1(x, y) + a_2(x, y)(x - x') + a_3(y - y') \\
&+ a_4(x - x')^2 + a_5(x - x')(y - y') + a_6(y - y')^2
\end{aligned} \tag{2.5.32}$$

and $K_B(\boldsymbol{x}) := K\big(\big(B^{-1}(\boldsymbol{x})\big)_1\big) K\big(\big(B^{-1}(\boldsymbol{x})\big)_2\big)$, where K is called kernel function and it denotes some symmetric probability density.

The most commonly used kernel function is the density of the standard normal distribution, which is called *Gaussian kernel*. For more details we refer to Nagler (2014).

Exploratory Visualization

Copula visualisation is useful for analysing the bivarate parametric copulas we introduced so far. One useful tool is a scatterplot. However, as the support of the copula is the unit square, the copula densities for the different classes discussed before are not easy to interpret. Solution to this problem is *normalized bivariate copula contour plot* (Czado (2019)), which is the contour plot of a bivariate density obtained from a copula density transformed to achieve standard normal margins. We look at the transformed density

$$g(z_1, z_2) = c(\Phi(z_1), \Phi(z_2))\phi(z_1)\phi(z_2) \tag{2.5.33}$$

of (Z_1, Z_2), where $Z_i := \Phi^{-1}(U_i) = \Phi^{-1}(F_i(X_i))$ for $i = 1, 2$. Here, $\Phi(\cdot)$ and $\phi(\cdot)$ are the distribution and density function of a $\mathcal{N}(0, 1)$ variable.

Bivariate parametric copula estimation

Now we proceed with estimation of bivariate parametric copulas and their parameter. Let us have bivariate pseudo-copula data $\boldsymbol{u} = \{(u_{i1}, u_{i2}), i = 1, \dots, n\}$, which we will use for estimation of the copula parameter of the bivariate copula family. For parameter

estimation in bivariate copula models we define the *maximum likelihood method* (Czado (2019)).
Then the (pseudo) maximum likelihood estimator $\hat{\boldsymbol{\theta}}_{ML}$ maximizes the likelihood, which is given by

$$\ell(\boldsymbol{\theta}; \boldsymbol{u}) := \prod_{i=1}^{n} c(u_{i1}, u_{i2}; \boldsymbol{\theta}), \tag{2.5.34}$$

where $c(\cdot, \cdot; \boldsymbol{\theta})$ is the respective copula density of a chosen bivariate copula family.

2.6 Linear quantile regression

Before introducing D-vine regression model, we would like to define a simpler quantile regression model, based on the work of Koenker (2005). The linear quantile regression model was first introduced by Koenker and Bassett (1978) as an extention of the ordinary least square regression method, as a class of more flexible statistical models which provide more complete picture of the stochastic relationships among the random variables. As a regression method which is more robust against outliers of the response variable, it is nowadays widely used in almost every discipline.
For the definition of quantile regression, we first introduce the concept of quantiles. Given an univariate ramdon variable X, characterized by its distribution function

$$F(x) = P(X \leq x),$$

we define the αth quantile of X as the "inverse" function of F,

$$F^{-1}(\alpha) = \inf\{x : F(x) \geq \alpha\} := q_\alpha(x), \quad 0 < \alpha < 1. \tag{2.6.1}$$

Namely, quantiles arise from a simple optimization problem that is fundamental to quantile regression. Assume we want to solve a simple decision theoretic problem: a point estimate for a random variable X with a distribution function F. We define the loss function as a piecewise linear function,

$$\rho_\alpha(x) = x(\alpha - I(x < 0)), \quad \alpha \in (0, 1) \tag{2.6.2}$$

where $I(\cdot)$ denotes the indicator function. Our aim is to find $\hat{x}$ which minimizes the expected loss function

$$\underset{\hat{x}}{\operatorname{argmin}} E[\rho_\alpha(X - \hat{x})], \tag{2.6.3}$$

which can be written down as

$$E[\rho_\alpha(X - \hat{x})] = (\alpha - 1) \int_{-\infty}^{\hat{x}} (x - \hat{x}) dF(x) + \alpha \int_{\hat{x}}^{\infty} (x - \hat{x}) dF(x). \tag{2.6.4}$$

By differentiating Equation (2.6.4) with respect to $\hat{x}$ we obtain

$$0 = (1 - \alpha) \int_{-\infty}^{\hat{x}} dF(x) - \alpha \int_{\hat{x}}^{\infty} dF(x) = F(\hat{x}) - \alpha. \tag{2.6.5}$$

F is monotone function, therefore any element of $\{x : F(x) = \alpha\}$ minimizes the expected loss function. If the solution is unique, then $\hat{x} = F^{-1}(\alpha)$, otherwise we obtain an interval

of αth quantiles, from which the smallest element is chosen, which leads to the definition of our target function $q_\alpha(x) = \inf\{x : F(x) \geq \alpha\}$.
For the purpose of quantile regression we first introduce the conditional distribution function, as defined by Bernand and Czado (2015).

Definition 2.6.1 *(Conditional distribution function for a continuous random variable) Given a continuous response variable Y and predictor variables $X_1, \ldots, X_k$, the conditional quantile function for $\alpha \in (0,1)$ is defined as*

$$\begin{aligned} q_\alpha(x_1, \ldots, x_k) :&= F^{-1}_{Y|X_1,\ldots,X_k}(\alpha | X_1 = x_1, \ldots, X_k = x_k) \\ &= \inf\{y \in \mathbb{R} | F_{Y|X_1,\ldots,X_k}(y) \geq \alpha\}. \end{aligned} \tag{2.6.6}$$

Now we proceed with definition of the first quantile regression model introduced in Koenker and Bassett (1978), where the conditional quantiles depend linearly on the predictors.

Definition 2.6.2 *(Linear quantile regression model)*
Let Y be a continuous response variable depending on set of predictors $X_1, \ldots, X_k, k \geq 1$. For every α the conditional quantiles of Y given $X_1 = x_1, \ldots, X_k = x_k$ are given as

$$Q_{Y|X_1,\ldots,X_k}(\alpha | x_1, \ldots, x_k) := \beta_0(\alpha) + \sum_{j=1}^{k} \beta_j(\alpha) x_j. \tag{2.6.7}$$

As proposed in Koenker and Bassett (1978), given n observations of the response variable Y and the predictors $X_1, \ldots, X_k$, we estimate the coefficients $\hat{\boldsymbol{\beta}}(\alpha) = (\hat{\beta}_0(\alpha), \hat{\beta}_1(\alpha), \ldots, \hat{\beta}_k(\alpha)) \in \mathbb{R}^{k+1}$ by solving the minimization problem

$$\min_{\boldsymbol{\beta}(\alpha) \in \mathbb{R}^{k+1}} \sum_{i=1}^{n} \rho_\alpha(y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}(\alpha)) \tag{2.6.8}$$

where ρ_α is the loss function defined with Equation (2.6.2) and $\boldsymbol{x}_i^\top = (1, x_{i1}, \ldots, x_{ik})$ is the ith observation of the predictors $X_1, \ldots, X_k$, including the intercept. This optimization problem might be solved by a linear programming technique such as the simplex method, and its solution can be found in Section 6.2 of Koenker (2005).
Linear quantile regression has however few drawbacks. One of them, as pointed out by Koenker (2005), is quantile crossing, where the regression lines of different quantiles levels may cross due to different slopes. This may be problematic when we use predicted quantiles as upper and lower bounds of prediction intervals. Namely, for central 90% prediction interval we can use the predicted conditional 0.05- and 0.95-quantiles. If these two quantiles cross, i.e. $\hat{q}_{0.05} > \hat{q}_{0.95}$, we do not obtain reasonable prediction interval. Another drawback, which is pointed out by Czado and Bernard (2015), is that the assumption of linearity of the conditional quantiles is strong and almost never fulfilled. Additionally, this regression method is prone to typical issues like multicollinearity, necessary variable transformations and interactions. D-vine quantile regression takes care of these issues and allows for more flexible models, as presented in the following section.

2.7 D-vine quantile regression

In this section we want to introduce a method of modelling *n-dimensional data* using only bivariate building blocks (Kraus and Czado (2017)). Such a tool is called *pair copula construction (PCC)*, which allows us to decompose multidimensional density to a bivariate copulas and conditional distribution functions. D-vine copulas allow for representation of the data through a graph theoretical model given by a sequence of trees. Moreover, D-vine quantile regression overcomes issues like collinearity, need for transformation and interaction of variables (Kraus and Czado (2017)).

Illustration of PCC in three dimensions

Before defining regular vines and D-vine based quantile regression, we want to present an example of pair copula construction of three-dimensional joint density (Czado (2019)). Let X_1,X_2 and X_3 be random variables. Using recursion we can rewrite their joint density as

$$f_{123}(x_1, x_2, x_3) = f_{3|12}(x_3|x_1, x_2) f_{2|1}(x_2|x_1) f_1(x_1). \tag{2.7.1}$$

Now we continue with decomposing each part separately. For $f_{3|12}(x_3|x_1, x_2)$ we consider the bivariate conditional density $f_{13|2}(x_1, x_3|x_2)$, which by Equation (2.5.4) in Sklar's Theorem can be written as:

$$f_{13|2}(x_1, x_3|x_2) = c_{13;2}(F_{1|2}(x_1|x_2), F_{3|2}(x_3|x_2); x_2) f_{1|2}(x_1|x_2) f_{3|2}(x_3|x_2), \tag{2.7.2}$$

where $c_{13;2}(\cdot, \cdot; x_2)$ denotes the copula density associated with the conditional distribution of (X_1, X_3) given $X_2 = x_2$. Now, applying Lemma 2.5.1 to Equation (2.7.2) we get:

$$\begin{aligned} f_{3|12}(x_3|x_1, x_2) &= \frac{f_{123}(x_1, x_2, x_3)}{f_{12}(x_1, x_2)} = \frac{f_{13|2}(x_1, x_3|x_2) f_2(x_2)}{f_{1|2}(x_1|x_2) f_2(x_2)} \\ &= \frac{c_{13;2}(F_{1|2}(x_1|x_2), F_{3|2}(x_3|x_2); x_2) f_{1|2}(x_1|x_2) f_{3|2}(x_3|x_2)}{f_{1|2}(x_1|x_2)} \\ &= c_{13;2}(F_{1|2}(x_1|x_2), F_{3|2}(x_3|x_2); x_2) f_{3|2}(x_3|x_2) \\ &= c_{13;2}(F_{1|2}(x_1|x_2), F_{3|2}(x_3|x_2); x_2) c_{23}(F_2(x_2), F_3(x_3)) f_3(x_3) \end{aligned} \tag{2.7.3}$$

Again, directly applying Lemma 2.5.1 to $f_{2|1}(x_2|x_1)$ we can rewrite Equation (2.7.1) in a form of pair copula decomposition:

$$\begin{aligned} f_{123}(x_1, x_2, x_3) = {} & c_{13;2}(F_{1|2}(x_1|x_2), F_{3|2}(x_3|x_2); x_2) c_{23}(F_2(x_2), F_3(x_3)) \\ & \times c_{12}(F_1(x_1), F_2(x_2)) f_3(x_3) f_2(x_2) f_1(x_1). \end{aligned} \tag{2.7.4}$$

Important note is that this decomposition is not unique, since

$$\begin{aligned} f_{123}(x_1, x_2, x_3) &= f_{2|13}(x_2|x_1, x_3) f_{1|3}(x_1|x_3) f_3(x_3) \\ &= c_{12;3}(F_{1|3}(x_1|x_3), F_{2|3}(x_2|x_3); x_3) c_{13}(F_1(x_1), F_3(x_3)) \\ &\times c_{23}(F_2(x_2), F_3(x_3)) f_1(x_1) f_2(x_2) f_3(x_3) \end{aligned} \tag{2.7.5}$$

and

$$\begin{aligned} f_{123}(x_1, x_2, x_3) &= f_{1|23}(x_1|x_2, x_3) f_{3|2}(x_3|x_2) f_2(x_2) \\ &= c_{23;1}(F_{2|1}(x_2|x_1), F_{3|1}(x_3|x_1); x_1) c_{13}(F_1(x_1), F_3(x_3)) \\ &\times c_{12}(F_1(x_1), F_2(x_2)) f_1(x_1) f_2(x_2) f_3(x_3). \end{aligned} \tag{2.7.6}$$

In Equation (2.7.4) the copula density $c_{13;2}(\cdot,\cdot;x_2)$ depends on the value x_2. It is common one to make the assumption that this dependence can be ignored. In such a case we talk about making the *simplifying assumption* (Czado (2019)). In this thesis, we assume that the simplifying assumption holds. In three dimensions, this means that for any $x_2 \in \mathbb{R}$,

$$c_{13;2}(u_1u_2;x_2) = c_{13;2}(u_1u_2), \quad \text{for} \quad u_1,u_2 \in [0,1] \tag{2.7.7}$$

holds. However, we are often interested in considering the dependence structure characterized by the copula on its own. Therefore, a three-dimensional parametric copula family with parameter vector $\boldsymbol{\theta} = (\theta_{12},\theta_{23},\theta_{13;2})$ can be defined as

$$\begin{aligned} c_{123}(u_1,u_2,u_3;\boldsymbol{\theta}) := c_{13;2}(C_{1|2}(u_1|u_2),C_{3|2}(u_3|u_2);\theta_{13;2})c_{23}(u_2,u_3;\theta_{23}) \\ \times c_{12}(u_1,u_2;\theta_{12}), \end{aligned} \tag{2.7.8}$$

where $C_{1|2}(\cdot|u_2)$ and $C_{3|2}(\cdot|u_2)$ are the conditional distribution functions of U_1 given $U_2 = u_2$ and U_3 given $U_2 = u_2$ respectively, as Equation (2.5.9) shows.
Using the density expression from Sklar's Theorem (Theorem 2.5.1) gives the copula density of (U_1,U_3) given $U_2 = u_2$:

$$c_{13;2}(u_1u_3;u_2) = \frac{c_{123}(C^{-1}_{1|2}(u_1|u_2),u_2,C^{-1}_{3|2}(u_3|u_2))}{c_{12}(C^{-1}_{1|2}(u_1|u_2),u_2)c_{23}(u_2,C^{-1}_{3|2}(u_3|u_2))}, \tag{2.7.9}$$

where the inverse conditional distribution functions as shown with Equation (2.5.11) are given by

$$C^{-1}_{1|2}(u_1|u_2) = F_1^{-1}(F^{-1}_{1|2}(u_1|x_2)) \tag{2.7.10}$$

$$C^{-1}_{3|2}(u_3|u_2) = F_3^{-1}(F^{-1}_{3|2}(u_3|x_2)) \tag{2.7.11}$$

and $u_2 = F_2(x_2)$.

Discrete PCCs

We also define PCCs for discrete ordinal margins $Y_1,\dots,Y_m$, as introduced in Panagiotelis, Czado and Joe (2012). Similarly to Equation (2.7.1), we can decompose the probability mass function as

$$\begin{aligned} P(Y_1 = y_1,\dots,Y_m = y_m) = P(Y_1 = y_1|Y_2 = y_2,\dots,Y_m = y_m) \\ \times P(Y_2 = y_2|Y_3 = y_3,\dots,Y_m = y_m)\dots P(Y_m = y_m). \end{aligned} \tag{2.7.12}$$

The terms on the right-hand side of Equation (2.7.12) have the form $P(Y_j = y_j|\boldsymbol{V} = \boldsymbol{v})$, where y_j is a scalar element of $\boldsymbol{y}$ and $\boldsymbol{v}$ is a subset of $\boldsymbol{y}$ which does not contain y_j. We choose a single element of $\boldsymbol{V}$, V_h, and let $\boldsymbol{V}_{\setminus h}$ be its complement. Then

$$\begin{aligned} P(Y_j = y_j|\boldsymbol{V} = \boldsymbol{v}) &= \frac{P(Y_j = y_j,V_h = v_h|\boldsymbol{V}_{\setminus h} = \boldsymbol{v}_{\setminus h})}{P(V_h = v_h|\boldsymbol{V}_{\setminus h} = \boldsymbol{v}_{\setminus h})} \\ &= \frac{\sum_{i_j=0,1}\sum_{i_h=0,1}(-1)^{i_j+i_h}P(Y_j \le y_j - i_j,V_h \le v_h - i_h|\boldsymbol{V}_{\setminus h} = \boldsymbol{v}_{\setminus h})}{P(V_h = v_h|\boldsymbol{V}_{\setminus h} = \boldsymbol{v}_{\setminus h})} \\ &= \frac{\sum_{i_j=0,1}\sum_{i_h=0,1}(-1)^{i_j+i_h}C_{Y_j,V_h|\boldsymbol{V}_{\setminus h}}\big(F_{Y_j|\boldsymbol{V}_{\setminus h}}(y_j - i_j|\boldsymbol{v}_{\setminus h}),F_{V_h|\boldsymbol{V}_{\setminus h}}(v_h - i_h|\boldsymbol{v}_{\setminus h})\big)}{P(V_h = v_h|\boldsymbol{V}_{\setminus h} = \boldsymbol{v}_{\setminus h})}, \end{aligned} \tag{2.7.13}$$

where $F_{A|B}(a|b)$ is notation for the distribution function $P(A \leq a|B = b)$. Equation (2.7.13) can be recursively applied to Equation (2.7.12) which results in decomposition of the multivariate probability mass function into bivariate copula building blocks. The arguments of the copula functions in Equation (2.7.13) have the form

$$\begin{aligned} F_{Y_j|V_h,\boldsymbol{V}_{\setminus h}}(y_j|v_h,\boldsymbol{v}_{\setminus h}) = &\frac{C_{Y_j,V_h|\boldsymbol{V}_{\setminus h}}\big(F_{Y_j|\boldsymbol{V}_{\setminus h}}(y_j|\boldsymbol{v}_{\setminus h}), F_{V_h|\boldsymbol{V}_{\setminus h}}(v_h|\boldsymbol{v}_{\setminus h})\big)}{P(V_h = v_h|\boldsymbol{V}_{\setminus h} = \boldsymbol{v}_{\setminus h})} \\ &- \frac{C_{Y_j,V_h|\boldsymbol{V}_{\setminus h}}\big(F_{Y_j|\boldsymbol{V}_{\setminus h}}(y_j|\boldsymbol{v}_{\setminus h}), F_{V_h|\boldsymbol{V}_{\setminus h}}(v_h-1|\boldsymbol{v}_{\setminus h})\big)}{P(V_h = v_h|\boldsymbol{V}_{\setminus h} = \boldsymbol{v}_{\setminus h})}. \end{aligned} \tag{2.7.14}$$

Now we want to investigate Sklar's theorem in a discrete case. For $\boldsymbol{V}_{\setminus h} = \boldsymbol{v}_{\setminus h}$, Sklar's theorem defines a unique copula with a domain given by the Cartesian product of the ranges of the cdf's of $Y_j|\boldsymbol{V}_{\setminus h} = \boldsymbol{v}_{\setminus h}$ and the range of the cdf's of $V_h|\boldsymbol{V}_{\setminus h} = \boldsymbol{v}_{\setminus h}$. Given that a copula denoted by $C_{Y_j,V_h|\boldsymbol{V}_{\setminus h}}$ exists over all possible values of $\boldsymbol{V}_{\setminus h} = \boldsymbol{v}_{\setminus h}$, it holds that the copula will be unique over the union of these domains. Let $a_{(1)} < a_{(2)} < \ldots < a_{(\kappa_1)}$ be the unique points corresponding to the ranges of $F_{Y_j|\boldsymbol{V}_{\setminus h}}$ in increasing order, and let $b_{(1)} < b_{(2)} < \ldots < b_{(\kappa_2)}$ be the ranges of $F_{V_h|\boldsymbol{V}_{\setminus h}}$. Let $a_{(0)} = b_{(0)} = 0$ and $a_{(\kappa_1+1)} = b_{(\kappa_2+1)} = 1$. We denote $a_{y_j|\boldsymbol{v}_{\setminus h}} := P(Y_j \leq y_j|\boldsymbol{V}_{\setminus h} = \boldsymbol{v}_{\setminus h})$, $b_{v_h|\boldsymbol{v}_{\setminus h}} := P(V_h \leq v_h|\boldsymbol{V}_{\setminus h} = \boldsymbol{v}_{\setminus h})$ and $p_{y_j,v_h|\boldsymbol{v}_{\setminus h}} := P(Y_j \leq y_j, V_h \leq v_h|\boldsymbol{V}_{\setminus h} = \boldsymbol{v}_{\setminus h})$. Then the constraints

$$C_{Y_j,V_h|\boldsymbol{V}_{\setminus h}}(a_{y_j|\boldsymbol{v}_{\setminus h}}, b_{v_h|\boldsymbol{v}_{\setminus h}}) = p_{y_j,v_h|\boldsymbol{v}_{\setminus h}}, \tag{2.7.15}$$

$$C_{Y_j,V_h|\boldsymbol{V}_{\setminus h}}(a_{y_j|\boldsymbol{v}_{\setminus h}}, 1) = a_{y_j|\boldsymbol{v}_{\setminus h}}, \;\; C_{Y_j,V_h|\boldsymbol{V}_{\setminus h}}(1, b_{v_h|\boldsymbol{v}_{\setminus h}}) = b_{v_h|\boldsymbol{v}_{\setminus h}}, \tag{2.7.16}$$

must be satisfied for all y_j, v_h and $\boldsymbol{v}_{\setminus h}$, which leads to 3κ constraints, where κ is the product of the cardinalities of the sets of possible values for Y_j, V_h and $\boldsymbol{V}_{\setminus h}$. Therefore, a bivariate copula $C = C_{Y_j,V_h|\boldsymbol{V}_{\setminus h}}$, constant over $\boldsymbol{v}_{\setminus h}$ exists if a solution to these constraints exists, such that all $(\kappa_1 + 1)(\kappa_2 + 1)$ unknowns $R_{jk} := C(a_{(j)}, b_{(k)}) + C(a_{(j-1)}, b_{(k)}) - C(a_{(j)}, b_{(k-1)}) + C(a_{(j-1)}, b_{(k-1)})$ are nonnegative.
More details regarding discrete PCCs can be found in Panagiotelis, Czado and Joe (2012).

D-vine copulas

The theory for D-vine copulas can be found in Chapters 4 and 5 in Czado (2019). Before defining D-vine tree sequence, we first present a necessary background from graph theory.

Definition 2.7.1 *(Graph, node, edge, degree, path, connected graph, cycle)*

- *A graph is a pair $G = (N, E)$ of sets such that $E \subseteq \{\{x, y\} : x, y \in N\}$.*
- *The elements of E are called edges of the graph G, whereas the elements of N are called nodes.*
- *The number of neighbours of a node $\upsilon \in N$ is the degree of υ denoted by $d(\upsilon)$.*
- *A path is a graph $P = (N, E)$ with node set $N = \{\upsilon_0, \upsilon_1, \ldots, \upsilon_k\}$ and edges $E = \{\{\upsilon_0, \upsilon_1\}, \{\upsilon_1, \upsilon_2\}, \ldots, \{\upsilon_{k-1}, \upsilon_k\}\}$.*

- *A graph G is called connected if any two of its nodes are linked by a path in G.*
- *A cycle is a path with $\upsilon_0 = \upsilon_k$.*

We define the term *tree* through the following theorem:

Theorem 2.7.1 *(Characterization of trees)*
The following statements are equivalent for a graph $T = (N, E)$:

1. *T is a tree.*
2. *Any two nodes of T are connected by a unique path in T.*
3. *T is minimally connected, i.e. T is connected but $T - e$ is disconnected for every edge $e \in E$.*
4. *T is maximally acyclic i.e. T contains no cycle but $T + \{x, y\}$ does for any two nodes $x, y \in N$ that are not connected by an edge in T.*

Now, we can finally define *regular (R-) vine tree sequence.*

Definition 2.7.2 *(Regular (R-) vine tree sequence)*
The set of trees $\mathcal{V} = (T_1, \dots, T_{d-1})$ is a regular vine tree sequence on d elements if

1. *Each tree $T_j = (N_j, E_j)$ is connected.*
2. *T_1 is a tree with node set $N_1 = \{1, \dots, d\}$ and edge set E_1.*
3. *For $j \leq 2$, T_j is a tree with node set $N_j = E_{j-1}$ and edge set E_j.*
4. *For $j = 2, \dots, d-1$ and $\{a, b\} \in E_j$ it must hold that $|a \cap b| = 1$.*

Remark 2.7.1 *(Proximity condition) The property 4. is called proximity condition. It ensures that if there is an edge e connecting a and b in tree T_j, $j \leq 2$, then a and b (which are edges in T_{j-1}) must share a common node in T_{j-1}.*

Definition 2.7.3 *(Complete union and conditioned sets)*
For any edge $e \in E_i$ we define the set called complete union:

$$A_e := \{j \in N_1 | \exists e_1 \in E_1, \dots, e_{i-1} \in E_{i-1} \text{ such that } j \in e_1 \in \dots e_{i-1} \in e\}. \tag{2.7.17}$$

The conditioning set D_e of an edge $e = \{a, b\}$ is defined by

$$D_e := A_a \cap A_b \tag{2.7.18}$$

and the conditioned sets $\mathcal{C}_{e,a}$ and $\mathcal{C}_{e,b}$ are given by

$$\begin{aligned} \mathcal{C}_{e,a} &:= A_a \setminus D_e \\ \mathcal{C}_{e,b} &:= A_b \setminus D_e \\ \mathcal{C}_e &:= \mathcal{C}_{e,a} \cup \mathcal{C}_{e,b}. \end{aligned} \tag{2.7.19}$$

We often abbreviate each edge $e = (\mathcal{C}_{e,a}, \mathcal{C}_{e,b}; D_e)$ in the vine tree sequence by

$$e = (e_a, e_b; D_e). \tag{2.7.20}$$

In the following part we add stochastic component to the regular (R-) vine tree sequence.

Definition 2.7.4 *(Regular vine distribution)*
The joint distribution F for the d dimensional random vector $\boldsymbol{X} = (X_1, \ldots, X_d)$ has a regular vine distribution if we can specify a triplet $(\mathcal{F}, \mathcal{V}, \mathcal{B})$ such that:

1. ***Marginal distributions:*** *$\mathcal{F} = (F_1, \ldots, F_d)$ is a vector of continuous invertible marginal distribution functions, representing the marginal distribution functions of the random variable X_i, i=1, ..., d.*
2. ***Regular vine tree sequence:*** *$\mathcal{V}$ is an R-vine tree sequence on d elements.*
3. ***Bivariate copulas:*** *The set $\mathcal{B} = \{C_e | e \in E_i; i = 1, \ldots, d-1\}$, where C_e is a symmetric bivariate copula with density. Here E_i is the edge set of tree T_i in the R-vine tree sequence $\mathcal{V}$.*
4. ***Relationship between R-vine tree sequence $\mathcal{V}$ and the set $\mathcal{B}$ of bivariate copulas:*** *For each $e \in E_i$, i=1,..., d-1, $e = \{a, b\}$, C_e is the copula associated with the conditional distribution of $X_{C_{e,a}}$ and $X_{C_{e,b}}$ given $\boldsymbol{X}_{D_e} = \boldsymbol{x}_{D_e}$. Further $C_e(\cdot, \cdot)$ does not depend on the specific value of $\boldsymbol{x}_{D_e}$, which is also called simplifying assumption.*

Definition 2.7.5 *(Pair copula and copula density associated with edge e)*
We will denote the copula C_e corresponding to the edge e by $C_{\mathcal{C}_{e,a}\mathcal{C}_{e,b};D_e}$ and the corresponding density by $c_{\mathcal{C}_{e,a}\mathcal{C}_{e,b};D_e}$ respectively. This copula is also called a pair copula.

In this thesis we will focus on D-vine tree sequence, which we define next.

Definition 2.7.6 *(D-vine tree sequence)*
A regular vine tree sequence $\mathcal{V} = (T_1, \ldots, T_{d-1})$ is called D-vine tree sequence if for each node $n \in N_i$ we have $|\{e \in E_i | n \in e\}| \leq 2$.

Remark 2.7.2 *For a D-vine tree sequence the proximity condition of Definition 2.7.2 induces that once tree T_1 is fixed all other trees T_2 to T_d are determined.*

Given a set of random variables $(X_1, \ldots, X_d)$, distinct indices i,j and $D := \{i_1, \ldots, i_k\}$ with $i < j$ and $i_1 < \ldots < i_k$ in the following theorem we use the abbreviation

$$c_{i,j;D} := c_{i,j;D}(F_{i|D}(x_i|\boldsymbol{x}_D), F_{j|D}(x_j|\boldsymbol{x}_D); \boldsymbol{x}_D). \tag{2.7.21}$$

Theorem 2.7.2 *(Drawable vine (D-vine) density)*
The joint density $f_{1,\ldots,d}$ of continuously distributed random vector $\boldsymbol{X} = (X_1, \ldots, X_d)$ can be decomposed as

$$f_{1,\ldots,d}(x_1, \ldots, x_d) = \left[\prod_{j=1}^{d-1}\prod_{i=1}^{d-j} c_{i,(i+j);(i+1)\ldots(i+j-1)}\right] \cdot \left[\prod_{k=1}^{d} f_k(x_k)\right], \tag{2.7.22}$$

where we used the abbreviation from Equation (2.7.21). The distribution associated with this density decomposition is called drawable vine (D-vine).

The distribution associated with the density decomposition (2.7.22), if the marginals are uniformly distributed, is called *D-vine copula*.

Definition 2.7.7 *(Simplifying assumption for D-vines)*
If it holds that

$$c_{i,j;D}(F_{i|D}(x_i|\boldsymbol{x}_D), F_{j|D}(x_j|\boldsymbol{x}_D); \boldsymbol{x}_D) = c_{i,j;D}(F_{i|D}(x_i|\boldsymbol{x}_D), F_{j|D}(x_j|\boldsymbol{x}_D)) \tag{2.7.23}$$

for all $\boldsymbol{x}_D$, then the corresponding D-vine distribution is called simpified.

Next we present an Example of simplified five dimensional D-vine density given by Kraus and Czado (2017).

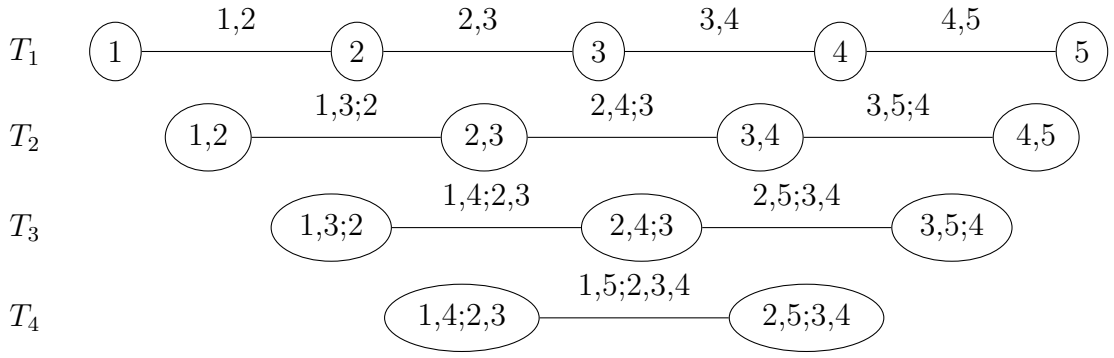

Figure 1: D-vine tree sequence in five dimensions.

Example 2.7.1 *(Simplified five dimensional D-vine density)*
As illustrated in Figure 1 and stated in Theorem 2.7.2, for d=5 the simplified D-vine density has the following form

$$\begin{aligned} f_{12345}(x_1,x_2,x_3,x_4,x_5) &= c_{12}(F_1(x_1),F_2(x_2))c_{23}(F_2(x_2),F_3(x_3))c_{34}(F_3(x_3),F_4(x_4)) \\ &\times c_{45}(F_4(x_4),F_5(x_5))c_{13;2}(F_{1|2}(x_1|x_2),F_{3|2}(x_3|x_2)) \\ &\times c_{24;3}(F_{2|3}(x_2|x_3),F_{4|3}(x_4|x_3))c_{35;4}(F_{3|4}(x_3|x_4),F_{5|4}(x_5|x_4)) \\ &\times c_{14;23}(F_{1|23}(x_1|x_2,x_3),F_{4|23}(x_4|x_2,x_3)) \\ &\times c_{25;34}(F_{2|34}(x_2|x_3,x_4),F_{5|34}(x_5|x_3,x_4)) \\ &\times c_{15;234}(F_{1|234}(x_1|x_2,x_3,x_4),F_{5|234}(x_5|x_2,x_3,x_4))[\prod_{i=1}^{5} f_i(x_i)]. \end{aligned}$$

As presented in Kraus and Czado (2017), the conditional distributions $F_{i|D}(x_i|\boldsymbol{x}_D)$ appearing in the PCC can be evaluated using only pair-copulas specified for D-vine from lower trees by applying the following recursion. Let $l \in D$ and $D_{-l} := D \setminus \{l\}$, then

$$F_{i|D}(x_i|\boldsymbol{x}_D) = h_{i|l;D_{-l}}(F_{i|D_{-l}}(x_i|\boldsymbol{x}_{D_{-l}})|F_{l|D_{-l}}(x_l|\boldsymbol{x}_{D_{-l}})), \tag{2.7.24}$$

where for $i,j \notin D, i<j, h_{i|j;D}(u|v) = \partial C_{ij;D}(u,v)/\partial v = C_{i|j;D}(u|v)$ and analogously $h_{j|i;D}(u|v) = \partial C_{ij;D}(u,v)/\partial u = C_{j|i;D}(u|v)$ are the h-functions associated with the pair-copula $C_{ij;D}$.

Using this formula on the first argument of $c_{14;23}$ from Tree 3 in Example 2.7.1, we can evaluate $F_{1|23}(x_1|x_2, x_3)$ using the h-functions associated with $C_{13;2}$, C_{12} and C_{23} from the first two trees:

$$\begin{aligned} F_{1|23}(x_1|x_2, x_3) &= h_{1|3;2}(F_{1|2}(x_1|x_2)|F_{3|2}(x_3|x_2)) \\ &= h_{1|3;2}(h_{1|2}(F_1(x_1)|F_2(x_2))|h_{3|2}(F_3(x_3)|F_2(x_2))). \end{aligned}$$

D-vine based quantile regression

D-vine quantile regression was first introduced by Kraus and Czado (2017). In this section we follow their work. The focus of interest is modelling the response variable Y given the outcome of some predictor variables $X_1, \ldots, X_d,\ d \geq 1$, where $Y \sim F_Y$ and $X_j \sim F_j,\ j = 1, \ldots, d$. Therefore, we want to predict the $\alpha \in (0, 1)$ quantile of the response variable Y given $\boldsymbol{X}$, which can be achieved by a joint modelling of $(Y, \boldsymbol{X})^\top$ and using the conditional quantile function

$$q_\alpha(x_1, \ldots, x_d) := F^{-1}_{Y|X_1,\ldots,X_d}(\alpha|x_1, \ldots, x_d). \tag{2.7.25}$$

As shown by Kraus and Czado (2017), we can estimate the conditional distribution

$$F_{Y|X_1,\ldots,X_d}(y|x_1, \ldots, x_d) = P(Y \leq y|X_1 \leq x_1, \ldots, X_d \leq x_d)$$

using D-vine copulas. First we transform all the variables to the u-scale using the probability integral transforms $V := F_Y(Y)$ and $U_j := F_j(X_j)$, with corresponding PIT values $v := F_Y(y)$ and $u_j := F_j(x_j)$. It follows that

$$\begin{aligned} F_{Y|X_1,\ldots,X_d}(y|x_1, \ldots, x_d) &= P(Y \leq y|X_1 = x_1, \ldots, X_d = x_d) \\ &= P(F_Y(y) \leq v|F_1(x_1) = u_1, \ldots F_d(x_d) = u_d) \\ &= C_{V|U_1,\ldots,U_d}(v|u_1, \ldots, u_d). \end{aligned} \tag{2.7.26}$$

Therefore,

$$F^{-1}_{Y|X_1,\ldots,X_d}(\alpha|x_1, \ldots, x_d) = F^{-1}_Y(C^{-1}_{V|U_1,\ldots,U_d}(\alpha|u_1, \ldots, u_d)). \tag{2.7.27}$$

Under the assumption that the margins $F_Y,\ F_j$, for $j = 1, \ldots, d$ are known, we can obtain an estimate of the conditional quantile function $F^{-1}_{Y|X_1,\ldots,X_d}$ by only estimating the conditional distribution function $C_{V|U_1,\ldots,U_d}$. We denote the estimated inverses of the marginal distributions by $\hat{F}^{-1}_Y,\ \hat{F}^{-1}_j, j = 1, \ldots, d$ and of the conditional distribution function by $\hat{C}_{V|U_1,\ldots,U_d}$ such that

$$\hat{q}_\alpha(x_1, \ldots, x_d) := \hat{F}^{-1}_Y(\hat{C}^{-1}_{V|U_1,\ldots,U_d}(\alpha|\hat{u}_1, \ldots, \hat{u}_d)), \tag{2.7.28}$$

where $\hat{u}_j = \hat{F}_j(x_j)$ for $j = 1, \ldots, d$. Let $C_{V,U_1,\ldots,U_d}$ be the (d+1)-dimensional copula associated with the joint distribution of $(Y, \boldsymbol{X})^\top$. The problem of estimating this multivariate copula can be solved in sequential way using D-vine pair copula constructions. Therefore, we fit a D-vine copula to $(V, U_1, \ldots, U_d)^\top$, with V as the first node in the first tree and order $V - U_{l_1} - \ldots - U_{l_d}$ where $(l_1, \ldots, l_d)^\top$ is an arbitrary permutation of $(1, \ldots, d)^\top$. To obtain the conditional distribution function $C_{V|U_1,\ldots,U_d}$ from the copula $C_{V,U_1,\ldots,U_d}$ we will use Equation (2.7.24) to express the conditional distribution function in terms of nested h-functions and consequently, $C^{-1}_{V|U_1,\ldots,U_d}(\alpha|u_1, \ldots, u_d)$ in terms of inverse h-functions (Kraus and Czado (2017)).

Example 2.7.2 *(Expressing $C_{V|U_1,\dots,U_d}$ using h-functions and $C^{-1}_{V|U_1,\dots,U_d}$ using inverse h-functions in four dimensions)*
Given a D-vine with order $V-U_1-U_2-U_3$, using the Equation (2.7.24) we can recursively express the conditional distribution of V given $(U_1,U_2,U_3)^\top$ as

$$\begin{aligned}
&C_{V|U_1,U_2,U_3}(v|u_1,u_2,u_3) = h_{V|U_3;U_1,U_2}(C_{V|U_1,U_2}(v|u_1,u_2)|C_{U_3|U_1,U_2}(u_3|u_1,u_2))\\
&= h_{V|U_3;U_1,U_2}(h_{V|U_2;U_1}(C_{V|U_1}(v|u_1)|C_{U_2|U_1}(u_2|u_1))|h_{U_3|U_1;U_2}(C_{U_3|U_2}(u_3|u_2)|C_{U_1|U_2}(u_1|u_2)))\\
&= h_{V|U_3;U_1,U_2}(h_{V|U_2;U_1}(h_{V|U_1}(v|u_1)|h_{U_2|U_1}(u_2|u_1))|h_{U_3|U_1;U_2}(h_{U_3|U_2}(u_3|u_2)|h_{U_1|U_2}(u_1|u_2))).
\end{aligned}$$

Inversion yields

$$\begin{aligned}
&C^{-1}_{V|U_1,U_2,U_3}(\alpha|u_1,u_2,u_3) =\\
&h^{-1}_{V|U_1}\Big[h^{-1}_{V|U_2;U_1}\Big\{h^{-1}_{V|U_3;U_1,U_2}\Big(\alpha|h_{U_3|U_1;U_2}(h_{U_3|U_2}(u_3|u_2)|h_{U_1|U_2}(u_1|u_2))\Big)\Big|h_{U_2|U_1}(u_2|u_1)\Big\}\Big|u_1\Big].
\end{aligned}$$

This example can be easily expanded to higher dimensions. Now we proceed with the estimation process of $q_\alpha(\boldsymbol{x})$, which is a two step estimation procedure. In the first step we estimate the marginal distribution functions $F_Y, F_j, j = 1,\dots,d$ and in the second step we estimate the D-vine that specifies the pair copulas needed to evaluate $\hat{C}^{-1}_{V|U_1,\dots,U_d}(\alpha|\hat{u}_1,\dots,\hat{u}_d)$.
Let $\boldsymbol{y} := (y^{(i)})_{i=1,\dots,n}, \boldsymbol{\mathcal{X}} := (x_j^{(i)})_{j=1,\dots,d,i=1,\dots,n}$ be n independent and identically distributed observations of the random vector $(Y,X_1,\dots,X_d)^\top$. In the following two sections we introduce the steps of the estimation procedure.

Estimation of the marginals

There are two ways we can fit the marginal distributions, either parametrically or non-parametrically. As a *parametric approach* we use finite mixture of distributions, which are beneficial for multimodal data. We use their estimation implemented in the R package `mixsmsn` (Prates, Lachos and Cabral (2013)). Let X be a continuous random variable which has (skewed) density function with more than one mode. We approximate its density using the density of a finite mixture of distributions, which has the form

$$g(x|\boldsymbol{\Psi}) := \sum_{i=1}^{g} p_i f_i(x|\boldsymbol{\psi}_i), \tag{2.7.29}$$

where $p_i \geq 0, i = 1,\dots,g$ with $\sum_{i=1}^g p_i = 1$ are called mixing weights, the density $f_i(\cdot|\boldsymbol{\psi}_i)$ is the ith component of the mixture, which is indexed by the (possibly multivariate) parameter $\boldsymbol{\psi}_i, i = 1,\dots,g$ and $\boldsymbol{\Psi} = ((p_1,\dots,p_g)^\top, \boldsymbol{\psi}_1^\top,\dots,\boldsymbol{\psi}_i^\top)^\top$. Furthermore, we assume that the components of the mixture belong to the class of the **scale mixtures of the skew-normal distributions (SMSN)**, which is a rich class of flexible distributions. However,the components of the mixtures that we use are either normal or skew-normal distributions, which are accommodated by the SMSN family of distributions, as shown below the following definition (Prates, Lachos and Cabral (2013)).

Definition 2.7.8 *(Univariate SMSN distribution family)*
The distribution of a random variable Y belongs to the univariate SMSN family when $Y = \mu + U^{-1/2}Z$, where $\mu \in \mathbb{R}$ is a location parameter, $Z \sim SN(0, \sigma^2, \lambda)$ is a skew-normal random variable and U is a positive random variable, independent of Z, with distribution function $H(\cdot|\boldsymbol{v})$. The parameters $\sigma^2 > 0$ and $\lambda \in \mathbb{R}$ are called scale and shape parameters respectively, and $H(\cdot|\boldsymbol{v})$ is known as the mixing scale distribution, indexed by (possibly multivariate) parameter $\boldsymbol{v}$. The density of Y at y has the form

$$SMSN(y|\mu, \sigma^2, \lambda, \boldsymbol{v}) = 2\int_0^\infty \phi(y|\mu, u^{-1}\sigma^2)\Phi(u^{\frac{1}{2}}\lambda\sigma^{-1}(y-\mu))dH(u|\boldsymbol{v}). \tag{2.7.30}$$

When $U = 1$ and $\lambda = 0$ we obtain

$$SMSN(y|\mu, \sigma^2) = \phi(y|\mu, \sigma^2),$$

which is the univariate normal distribution. Similarly when $U = 1$,

$$SMSN(y|\mu, \sigma^2, \lambda) = 2\phi(y|\mu, \sigma^2)\Phi(\lambda\sigma^{-1}(y-\mu))$$

the distribution of Y results in univariate skew-normal distribution. The estimation procedure is maximum likelihood via an EM-type algorithm, which is explained in detail in Basso, Lachos, Cabral and Ghosh (2010) and Cabral, Lachos and Prates (2012). We choose appropriate mixture of distributions for each marginal using the AIC and BIC criteria. In the final step we integrate the fitted density functions $\hat{g}_Y(\cdot; \hat{\boldsymbol{\Psi}}_Y), \hat{g}_j(\cdot; \hat{\boldsymbol{\Psi}}_j), j = 1, \ldots, d$ to obtain the cumulative distribution functions of the marginals, $\hat{F}_Y(\cdot; \hat{\boldsymbol{\Psi}}_Y)$ and $\hat{F}_j(\cdot; \hat{\boldsymbol{\Psi}}_j), j = 1, \ldots, d$. We use these to transform the observed data to pseudo copula data $\hat{v}^{(i)} := \hat{F}_Y(y^{(i)}; \hat{\boldsymbol{\Psi}}_Y)$ and $\hat{u}_j^{(i)} := \hat{F}_j(x_j^{(i)}; \hat{\boldsymbol{\Psi}}_j), j = 1, \ldots, d, i = 1, \ldots, n$.
Modelling the marginals as well as the copula parametrically might result is biased and inconsistent parametric estimator if one of the parametric models is misspecified. Therefore, we introduce the kernel smoothing estimator as a *nonparametric approach* for estimation of marginals (Geenens (2014)). Given a sample $(x^{(i)})_{i=1,\ldots,n}$ we define the kernel smoothing estimator as

$$\hat{F}(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - x^{(i)}}{h}\right), \quad x \in \mathbb{R}, \tag{2.7.31}$$

where $K(x) := \int_{-\infty}^{x} k(t)dt$ with $k(\cdot)$ being a symmetric probability density function and $h > 0$ a bandwidth parameter. This estimator is implemented in the R package `kde1d` (Nagler and Vatter 2022). Hence, we obtain $\hat{F}_Y$ and $\hat{F}_j$ as estimates for the marginal distibution functions. We use these to transform the observed data to pseudo copula data $\hat{v}^{(i)} := \hat{F}_Y(y^{(i)})$ and $\hat{u}_j^{(i)} := \hat{F}_j(x_j^{(i)}), j = 1, \ldots, d, i = 1, \ldots, n$.

Estimation of the D-vine

The pseudo copula data $(\hat{\boldsymbol{v}}, \hat{\boldsymbol{U}})$, where $\hat{\boldsymbol{v}} := (\hat{v}^{(i)})_{i=1,\ldots,n}, \hat{\boldsymbol{U}} := ((\hat{\boldsymbol{u}}^{(1)})^\top, \ldots, (\hat{\boldsymbol{u}}^{(n)})^\top)^\top = (\hat{u}_j^{(i)})_{j=1,\ldots,d, i=1,\ldots,n}$ obtained by either of the two approaches presented in the previous section is an approximately i.i.d. sample from the PIT random vector $(V, U_1, \ldots, U_d)^\top$ and is used by Kraus and Czado (2017) to estimate the D-vine copula. For the estimation of the D-vine with order $V - U_{l_1} - \ldots - U_{l_d}$, the ordering $\boldsymbol{l} = (l_1, \ldots, l_d)^\top$ can be generally

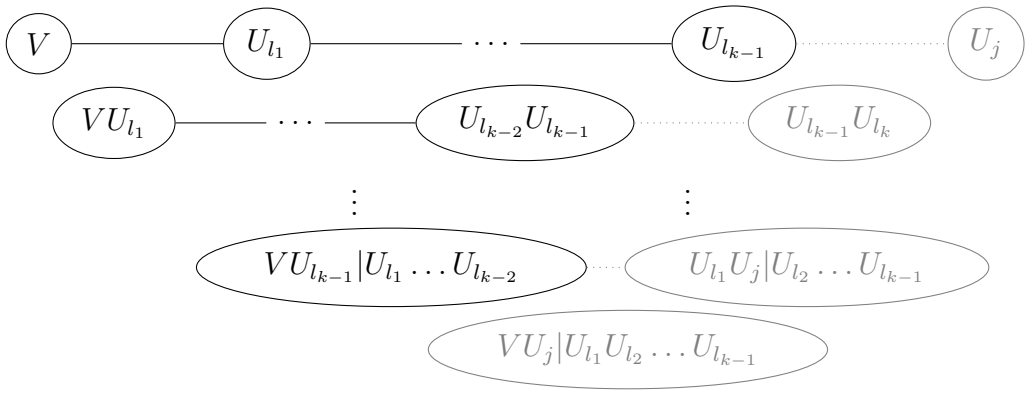

Figure 2: Extending the current D-vine (black) by adding U_j in the k-th step of the algorithm. The gray pair-copulas need to be estimated.

chosen arbitrary. However, we want to choose an ordering such that the model has the highest explanatory power. One option is to compare all $d!$ possible orderings, but this is infeasible solution. Therefore, Kraus and Czado (2017) introduce an algorithm which sequentially adds the most influential covariates. In each step, we add the covariate to the model that improves the model's fit the most. As a goodness of fit measure we define the *conditional log likelihood (cll)* of an estimated D-vine copula with ordering $\boldsymbol{l}$, estimated parametric pair-copula families $\hat{\boldsymbol{\mathcal{F}}}$ and corresponding copula parameters $\hat{\boldsymbol{\theta}}$ given pseudo data $(\hat{\boldsymbol{v}},\hat{\boldsymbol{\mathcal{U}}})$ as

$$cll(\boldsymbol{l},\hat{\boldsymbol{\mathcal{F}}},\hat{\boldsymbol{\theta}},\hat{\boldsymbol{\mathcal{U}}}) := \sum_{i=1}^{n} \ln c_{V|U}\Big(\hat{v}^{(i)}|\hat{\boldsymbol{u}}^{(i)};\boldsymbol{l},\hat{\boldsymbol{\mathcal{F}}},\hat{\boldsymbol{\theta}}\Big). \tag{2.7.32}$$

The conditional copula density $c_{V|U}$ can be expressed by the product of all pair copulas of the D-vine that contain V:

$$\begin{aligned} c_{V|U}\Big(\hat{v}^{(i)}|\hat{\boldsymbol{u}}^{(i)};\boldsymbol{l},\hat{\boldsymbol{\mathcal{F}}},\hat{\boldsymbol{\theta}}\Big) &= c_{VU_{l_1}}(\hat{v}^{(i)},\hat{u}_{l_1}^{(i)};\hat{\mathcal{F}}_{VU_{l_1}},\hat{\theta}_{VU_{l_1}})\times \\ &\prod_{j=2}^{d} c_{VU_{l_j};U_{l_1}\dots U_{l_{j-1}}}\Big(\hat{C}_{V|U_{l_1}\dots U_{l_{j-1}}}(\hat{v}^{(i)}|\hat{u}_{l_1}^{(i)}\dots\hat{u}_{l_{j-1}}^{(i)}),\hat{C}_{U_{l_j}|U_{l_1}\dots U_{l_{j-1}}}(\hat{u}_{l_j}^{(i)}|\hat{u}_{l_1}^{(i)}\dots\hat{u}_{l_{j-1}}^{(i)}); \\ &\qquad\qquad \hat{\mathcal{F}}_{VU_{l_j};U_{l_1}\dots U_{l_{j-1}}},\hat{\theta}_{VU_{l_j};U_{l_1}\dots U_{l_{j-1}}}\Big), \end{aligned}$$

where $\hat{\mathcal{F}}_I$ and $\hat{\theta}_I$ denote the estimated family and parameter(s) of the pair copula c_I. The algorithm which we introduce sequentially constructs a D-vine while maximising the model's conditional log likelihood in each step. At the beginning of the kth step of the algorithm, the current D-vine contains *k-1* predictors. As illustrated in Figure 2, for each of the remaining variables U_j that have not been chosen yet, we fit the pair copulas needed for extension of the D-vine with order $V - U_{l_1} - \ldots - U_{l_{k-1}} - U_j$ (see the gray circles) and compute the resulting model's conditional log likelihood. Eventually, the model is updated by adding the variable corresponding to the highest cll, concluding step k. This way, the covariates are sequentially added based on their power to predict the response. If none of the remaining covariates is able to increase the model's cll in the kth step, the

algorithm stops and returns the model containing the *k-1* covariates chosen so far. This algorithm is called *an automated forward selection procedure.*
The conditional log likelihood given by Equation (2.7.32) can be easily generalized for nonparametric bivariate copulas too, as shown by Tepegjozova, Zhou, Claeskens and Czado (2022).
For comparison of nested parametric models and deciding whether a predictor significantly improves a model if added to it, we use the following statistical test (Tepegjozova (2019)):

Definition 2.7.9 *(Conditional likelihood ratio test)*
Let $\mathcal{C}_1$ be a D-vine with order $V - U_1 - \ldots - U_p$ and $\mathcal{C}_2$ a D-vine with order $V - U_1 - \ldots - U_{p-1}$. Additionally, we assume that we are given n observations on each of the considered variables, i.e. $\boldsymbol{v}, \boldsymbol{u}_j, j = 1, \ldots, p$. We define the conditional likelihood ratio test between the vine copula models $\mathcal{C}_1$ and $\mathcal{C}_2$ as the test which rejects the null hypothesis

$$H_0 : \textit{Adding } U_p \textit{ to the model } \mathcal{C}_2 \textit{ does not improve the fit}$$

at level $\alpha \in (0, 1)$, if

$$cll(\mathcal{C}_1, \boldsymbol{v}, (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_p)) - cll(\mathcal{C}_2, \boldsymbol{v}, (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_{p-1})) > \chi^2_{1-\alpha, |\hat{\boldsymbol{\theta}}_1| - |\hat{\boldsymbol{\theta}}_2|}, \tag{2.7.33}$$

where $\hat{\boldsymbol{\theta}}_1$ and $\hat{\boldsymbol{\theta}}_2$ denote the parameters in the D-vines $\mathcal{C}_1$ and $\mathcal{C}_2$ respectively, and $\chi^2_{1-\alpha, |\hat{\boldsymbol{\theta}}_1| - |\hat{\boldsymbol{\theta}}_2|}$ denotes the (1-α)-quantile of a χ^2-distribution with $|\hat{\boldsymbol{\theta}}_1| - |\hat{\boldsymbol{\theta}}_2|$ degrees of freedom.

Derivation of the log likelihood

In order to compare the fitted D-vine model with order $V - U_{l_1} - \ldots - U_{l_d}$ with the other regression models, we need to derive the log likelihood of the model, with Y as a response variable and $X_{l_1}, \ldots, X_{l_d}$ as covariates included in the model. In particular, we want to derive the conditional density $f_{Y|l_1,\ldots,l_d}$ of $Y|X_{l_1}, \ldots, X_{l_d}$. Under the simplifying assumption,

$$\begin{aligned}
f_{Y|l_1,\ldots,l_d}(y|x_{l_1}, \ldots, x_{l_d}) &= \frac{f_{Y,l_1,\ldots,l_d}(y, x_{l_1}, \ldots, x_{l_d})}{f_{l_1,\ldots,l_d}(x_{l_1}, \ldots, x_{l_d})} \\
&= \frac{f_{Yl_d|l_1,\ldots,l_{d-1}}(y, x_{l_d}|x_{l_1}, \ldots, x_{l_{d-1}}) f_{l_1,\ldots,l_{d-1}}(x_{l_1}, \ldots, x_{l_{d-1}})}{f_{l_d|l_1,\ldots,l_{d-1}}(x_{l_d}|x_{l_1}, \ldots, x_{l_{d-1}}) f_{l_1,\ldots,l_{d-1}}(x_{l_1}, \ldots, x_{l_{d-1}})} \\
&= c_{Yl_d;l_1,\ldots,l_{d-1}}(F_{Y|l_1,\ldots,l_{d-1}}(y|x_{l_1}, \ldots, x_{l_{d-1}}), F_{l_d|l_1,\ldots,l_{d-1}}(x_{l_d}|x_{l_1}, \ldots, x_{l_{d-1}})) \\
&\times f_{Y|l_1,\ldots,l_{d-1}}(y|x_{l_1}, \ldots, x_{l_{d-1}}).
\end{aligned}$$

Now we can apply the same steps to $f_{Y|l_1,\ldots,l_{d-1}}$.We repeat this until we obtain construction of all bivariate copulas that contain Y:

$$\begin{aligned}
f_{Y|l_1,\ldots,l_d}(y|x_{l_1}, \ldots, x_{l_d}) &= f_Y(y) c_{Yl_1}(F_Y(y), F_{l_1}(x_{l_1})) \\
&\prod_{j=2}^{d} c_{Yl_j;l_1\ldots l_{j-1}}\Big(F_{Y|l_1\ldots l_{j-1}}(y|x_{l_1}\ldots x_{l_{j-1}}), F_{l_j|l_1\ldots l_{j-1}}(x_{l_j}|x_{l_1}\ldots x_{l_{j-1}})\Big).
\end{aligned} \tag{2.7.34}$$

Therefore given the observed data $(\boldsymbol{y}, \boldsymbol{\mathcal{X}})$, the pseudo copula data $(\hat{\boldsymbol{v}}, \hat{\boldsymbol{\mathcal{U}}})$, the pair copula families $\hat{\boldsymbol{\mathcal{F}}}$ and the corresponding copula parameters $\hat{\boldsymbol{\theta}}$, the *likelihood* of the model is

$$\begin{aligned} L(\boldsymbol{y}|\boldsymbol{l}, \hat{\boldsymbol{\mathcal{F}}}, \hat{\boldsymbol{\theta}}) &= \prod_{i=1}^{n} \hat{f}_{Y|l_1,\ldots,l_d}(y^{(i)}|x_{l_1}^{(i)}, \ldots, x_{l_d}^{(i)}) \\ &= \prod_{i=1}^{n} \hat{f}_Y(y^{(i)}) \prod_{i=1}^{n} c_{V|U}\Big(\hat{v}^{(i)}|\hat{\boldsymbol{u}}^{(i)}; \boldsymbol{l}, \hat{\boldsymbol{\mathcal{F}}}, \hat{\boldsymbol{\theta}}\Big), \end{aligned} \tag{2.7.35}$$

where $c_{V|U}$ is the conditional copula density from Equation (2.7.32) and $\hat{f}_Y$ is the fitted density function to the response variable Y. The *log likelihood* of the model is given by

$$l(\boldsymbol{l}, \hat{\boldsymbol{\mathcal{F}}}, \hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\mathcal{U}}}) = \sum_{i=1}^{n} \ln \hat{f}_Y(y^{(i)}) + cll(\boldsymbol{l}, \hat{\boldsymbol{\mathcal{F}}}, \hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\mathcal{U}}}), \tag{2.7.36}$$

where $cll(\boldsymbol{l}, \hat{\boldsymbol{\mathcal{F}}}, \hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\mathcal{U}}})$ is defined with Equation (2.7.32). In case of nonaprametric D-vine regression model we follow the same steps, where we just discard the copula parameters $\hat{\boldsymbol{\theta}}$.
Similar as before, using this log likelihood we can compute the AIC and BIC for the model using the Equations (2.3.16) and (2.3.17), where as number of parameters we take the parameters (degrees of freedom) used to fit the D-vine, and in case of parametric estimation of marginals also the number of parameters used for their estimation.

2.8 Comparison of different regression models

After we fit different regression models to the training data sets, we want to compare them. For comparison we use the log likelihood, AIC, BIC, training and test error, as well as interval score, all of which are calculated on the original Y scale. The log likelihood, AIC, BIC and training error are calculated on the training data set with n_{tr} number of observations, whereas the test error and the interval score on the test data set which contains n_{test} number of observations.
For the purpose of defining log likelihood on original Y scale for the models which use the transformed variable $\ln(Y)$ as a response variable, we denote the transformed response variable as $\tilde{Y} = \ln(Y)$. Having the density $f_{\tilde{Y}}$ of $\tilde{Y}$, we obtain the density f_Y of $Y = g(\tilde{Y}) = \exp(\tilde{Y})$ (Lefebvre (2006)) using the formula

$$f_Y(y) = f_{\tilde{Y}}(g^{-1}(y)) \left| \frac{dg^{-1}(y)}{dy} \right|. \tag{2.8.1}$$

Without loss of generality, let us denote the estimated density of $\tilde{Y}$ by $f_{\tilde{Y}}(\cdot|\hat{\boldsymbol{\varphi}})$, where $\hat{\boldsymbol{\varphi}}$ is the vector of the estimated parameters in the respective regression model. Given n_{tr} observations $\tilde{y}_i, i = 1, \ldots, n_{tr}$ of $\tilde{Y}$ and n_{tr} observations $y_i, i = 1, \ldots, n_{tr}$ of Y, such that $\tilde{y}_i = \ln(y_i), i = 1, \ldots, n_{tr}$ we derive the likelihood on Y scale

$$L(\boldsymbol{y}, \hat{\boldsymbol{\varphi}}) = \prod_{i=1}^{n_{tr}} f_Y(\boldsymbol{y}|\hat{\boldsymbol{\varphi}}) = \prod_{i=1}^{n_{tr}} f_{\tilde{Y}}(\ln(y_i)|\hat{\boldsymbol{\varphi}}) \frac{1}{y_i} = L(\tilde{\boldsymbol{y}}, \hat{\boldsymbol{\varphi}}) \prod_{i=1}^{n_{tr}} \frac{1}{y_i}, \tag{2.8.2}$$

where $L(\tilde{\boldsymbol{y}}, \hat{\boldsymbol{\varphi}})$ is the likelihood of the respective model. From this it follows that the log likelihood on Y scale is

$$l(\boldsymbol{y}, \hat{\boldsymbol{\varphi}}) = l(\tilde{\boldsymbol{y}}, \hat{\boldsymbol{\varphi}}) - \sum_{i=1}^{n_{tr}} \ln(y_i), \tag{2.8.3}$$

where $l(\tilde{\boldsymbol{y}}, \hat{\boldsymbol{\varphi}})$ is the log likelihood of the respective model. Additionally, we give a general definition of the criteria AIC and BIC defined with Equations (2.3.16) and (2.3.17). For a given model $\mathcal{M}$, we define the *Akaike information criterion (AIC)* and *Bayesian information criterion (BIC)* (Fahrmeir, Kneib, Lang and Marx (2013)) on the original Y scale as

$$AIC^{\mathcal{M}} := -2l(\tilde{\boldsymbol{y}}, \hat{\boldsymbol{\varphi}}) + 2q \tag{2.8.4}$$

$$BIC^{\mathcal{M}} := -2l(\tilde{\boldsymbol{y}}, \hat{\boldsymbol{\varphi}}) + \ln(n_{tr})q, \tag{2.8.5}$$

where $l(\tilde{\boldsymbol{y}}, \hat{\boldsymbol{\varphi}})$ is the maximized log likelihood of the respective model $\mathcal{M}$ and q is the number of parameters in the model. The number of parameters used for AIC and BIC for the lognormal and gamma regression model is p+1, where p is the length of the vector of estimated parameters $\hat{\boldsymbol{\beta}}$; for the linear quantile regression model with conditional α quantiles is the length of the vector of the estimated parameters $\hat{\boldsymbol{\beta}}(\alpha)$; for the D-vine regression model we take the number of parameters used to estimate the marginals and the bivariate copulas in the model. The models with larger log likelihood and smaller AIC and BIC are better.
Further, we define the *training* and *test error* of a model $\mathcal{M}$ (Hastie, Tibshirani and Friedman (2009)) on the original Y scale as

$$\overline{err}_{tr}^{\mathcal{M}} := \frac{1}{n_{tr}} \sum_{i=1}^{n_{tr}} \left(y_{i,tr} - \hat{y}_{i,tr}^{\mathcal{M}} \right)^2 \tag{2.8.6}$$

$$\overline{err}_{test}^{\mathcal{M}} := \frac{1}{n_{test}} \sum_{i=1}^{n_{test}} \left(y_{i,test} - \hat{y}_{i,test}^{\mathcal{M}} \right)^2, \tag{2.8.7}$$

where n_{tr} and n_{test} are the number of observations of the training and test data sets and $y_{i,tr}$, $y_{i,test}$ are observations of the original response variable in the training and test data set respectively. The values $\hat{y}_{i,tr}^{\mathcal{M}}, i = 1, \ldots, n_{tr}$ and $\hat{y}_{i,test}^{\mathcal{M}}, i = 1, \ldots, n_{test}$ we define for each model $\mathcal{M}$ separately as:

- *Lognormal regression model*: Let $\hat{\boldsymbol{\beta}}$ be the vector of estimated parameters in the model and $\boldsymbol{x}_{i,tr}^{\top}$,$\boldsymbol{x}_{i,test}^{\top}$ be the vector of the ith observation of the predictors present in the model including the intercept, in the training and test data set respectively. Then

$$\hat{y}_{i,tr}^{\mathcal{M}} = \exp\left(\boldsymbol{x}_{i,tr}^{\top}\hat{\boldsymbol{\beta}}\right), i = 1, \ldots, n_{tr} \text{ and } \hat{y}_{i,test}^{\mathcal{M}} = \exp\left(\boldsymbol{x}_{i,test}^{\top}\hat{\boldsymbol{\beta}}\right), i = 1, \ldots, n_{test}.$$

- *Gamma regression model*: Let $\hat{\boldsymbol{\beta}}$ be the vector of estimated parameters in the model and $\boldsymbol{x}_{i,tr}^{\top}$,$\boldsymbol{x}_{i,test}^{\top}$ be the vector of the ith observation of the predictors present in the model including the intercept, in the training and test data set respectively. Then

$$\hat{y}_{i,tr}^{\mathcal{M}} = \hat{\mu}_{i,tr} = \exp\left(\boldsymbol{x}_{i,tr}^{\top}\hat{\boldsymbol{\beta}}\right), i = 1, \ldots, n_{tr} \text{ and}$$
$$\hat{y}_{i,test}^{\mathcal{M}} = \hat{\mu}_{i,test} = \exp\left(\boldsymbol{x}_{i,test}^{\top}\hat{\boldsymbol{\beta}}\right), i = 1, \ldots, n_{test}.$$

- *Linear quantile regression model with original response* Y: Let $\hat{\boldsymbol{\beta}}(0.5)$ be the vector of estimated parameters in the model, where $\alpha = 0.5$, and $\boldsymbol{x}_{i,tr}^\top$,$\boldsymbol{x}_{i,test}^\top$ be the vector of the ith observation of the predictors present in the model including the intercept, in the training and test data set respectively. Then

$$\hat{y}_{i,tr}^{\mathcal{M}} = \hat{q}_{0.5}(\boldsymbol{x}_{i,tr}) = \boldsymbol{x}_{i,tr}^\top \hat{\boldsymbol{\beta}}(0.5), i = 1, \ldots, n_{tr} \text{ and}$$
$$\hat{y}_{i,test}^{\mathcal{M}} = \hat{q}_{0.5}(\boldsymbol{x}_{i,test}) = \boldsymbol{x}_{i,test}^\top \hat{\boldsymbol{\beta}}(0.5), i = 1, \ldots, n_{test}.$$

- *Linear quantile regression model with transformed response* $\ln(Y)$: Let $\hat{\boldsymbol{\beta}}(0.5)$ be the vector of estimated parameters in the model, where $\alpha = 0.5$, and $\boldsymbol{x}_{i,tr}^\top$,$\boldsymbol{x}_{i,test}^\top$ be the vector of the ith observation of the predictors present in the model including the intercept, in the training and test data set respectively. Then

$$\hat{y}_{i,tr}^{\mathcal{M}} = \exp\left(\hat{q}_{0.5}(\boldsymbol{x}_{i,tr})\right) = \exp\left(\boldsymbol{x}_{i,tr}^\top \hat{\boldsymbol{\beta}}(0.5)\right), i = 1, \ldots, n_{tr} \text{ and}$$
$$\hat{y}_{i,test}^{\mathcal{M}} = \exp\left(\hat{q}_{0.5}(\boldsymbol{x}_{i,test})\right) = \exp\left(\boldsymbol{x}_{i,test}^\top \hat{\boldsymbol{\beta}}(0.5)\right), i = 1, \ldots, n_{test}.$$

- *D-vine quantile regression model with original response* Y: Let $\boldsymbol{x}_{i,tr}^\top$,$\boldsymbol{x}_{i,test}^\top$ be the vector of the ith observation of the predictors present in the model, in the training and test data set respectively. Then

$$\hat{y}_{i,tr}^{\mathcal{M}} = \hat{q}_{0.5}(\boldsymbol{x}_{i,tr}), i = 1, \ldots, n_{tr} \text{ and } \hat{y}_{i,test}^{\mathcal{M}} = \hat{q}_{0.5}(\boldsymbol{x}_{i,test}), i = 1, \ldots, n_{test}.$$

- *D-vine quantile regression model with transformed response* $\ln(Y)$: Let $\boldsymbol{x}_{i,tr}^\top$,$\boldsymbol{x}_{i,test}^\top$ be the vector of the ith observation of the predictors present in the model, in the training and test data set respectively. Then

$$\hat{y}_{i,tr}^{\mathcal{M}} = \exp\left(\hat{q}_{0.5}(\boldsymbol{x}_{i,tr})\right), i = 1, \ldots, n_{tr} \text{ and } \hat{y}_{i,test}^{\mathcal{M}} = \exp\left(\hat{q}_{0.5}(\boldsymbol{x}_{i,test})\right), i = 1, \ldots, n_{test}.$$

Typically, the training error is an overly optimistic estimate of the overall error, since the same data is used to fit the model and assess this error. Therefore, we favour the comparison of the test errors more. Models with smaller training and test errors are better.
Additionally, we compute the *interval score* for a $(1-\alpha)100\%$ prediction interval, introduced by Gneiting and Raftery (2007), on original scale Y on the test data set. This measure rewards the model for narrow prediction intervals and it incurs penalty, the size of which depends on α, if an observation misses the interval. Therefore, smaller interval scores are better.
First we need to define the $(1-\alpha)100\%$ central prediction interval of the original response Y for each regression model. For the gamma regression models we use the prediction interval derived by Wasef Hattab (2016). For the models that use the transformed response $\ln(Y)$, we use the exponential function to transform the prediction interval on the original scale. This transformation of the prediction intervals is allowed because the exponential function is strictly increasing.

- *Lognormal regression model*: Let $\hat{\boldsymbol{\beta}}$ be the vector of estimated parameters in the model. The prediction interval for a future observation $\ln(y_0)$ of the lognormal

model at location $\boldsymbol{x}_0$ with level $1-\alpha$ is given by (Fahrmeir, Kneib, Lang and Marx (2013))

$$\left[\boldsymbol{x}_0^\top\hat{\boldsymbol{\beta}} - t_{n-p,1-\frac{\alpha}{2}}\, s\sqrt{1+\boldsymbol{x}_0^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{x}_0}, \boldsymbol{x}_0^\top\hat{\boldsymbol{\beta}} + t_{n-p,1-\frac{\alpha}{2}}\, s\sqrt{1+\boldsymbol{x}_0^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{x}_0}\right],$$

where $t_{n-p,1-\frac{\alpha}{2}}$ is the $(1-\frac{\alpha}{2})$ quantile of the t-distribution with $n-p$ degrees of freedom, $\mathbf{X}$ is the hat matrix, s is the unbiased estimator of σ^2 defined with Equation (2.3.10), n is the number of observations in the training data set i.e. $n = n_{tr}$ and p is the length of the vector of estimated parameters $\hat{\boldsymbol{\beta}}$. Using the exponential function, the prediction interval for a future observation y_0 at location $\boldsymbol{x}_0$ with level $1-\alpha$ is given by

$$\begin{aligned}\Big[\exp\Big(\boldsymbol{x}_0^\top\hat{\boldsymbol{\beta}} - t_{n-p,1-\frac{\alpha}{2}}\, s\sqrt{1+\boldsymbol{x}_0^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{x}_0}\Big),\\ \exp\Big(\boldsymbol{x}_0^\top\hat{\boldsymbol{\beta}} + t_{n-p,1-\frac{\alpha}{2}}\, s\sqrt{1+\boldsymbol{x}_0^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{x}_0}\Big)\Big].\end{aligned} \tag{2.8.8}$$

- *Gamma regression model*: We derive the prediction interval for a new observation y_0 following Wasef Hattab (2016). For a future observation y_0 at location $\boldsymbol{x}_0$, under the model's assumptions, it holds that $Y_0 \sim Gamma(\nu, \nu/\exp(\boldsymbol{x}_0^\top\boldsymbol{\beta}))$ and that Y_0 is independent of $\boldsymbol{Y}$. The MLE estimator of $\boldsymbol{\beta}$ for a gamma regression with log link function, $\hat{\boldsymbol{\beta}}$, follows a normal distribution asymptotically, i.e. $\hat{\boldsymbol{\beta}} \overset{a}{\sim} \mathcal{N}_p(\boldsymbol{\beta}, (\boldsymbol{X}^\top\boldsymbol{X})^{-1}/\nu)$. Using that for $X \sim Gamma(\alpha,\beta)$, it holds $cX \sim Gamma(\alpha, \beta/c)$ for a constant $c>0$, and that χ^2_n random variable is $Gamma(\frac{n}{2},\frac{1}{2})$ distributed (Czado and Schmidt (2011)), we obtain that $2\nu Y_0/\exp(\boldsymbol{x}_0^\top\boldsymbol{\beta}) \sim \chi^2_{2\nu}$.
  Next, we define the **delta method** (Jiang (2010), Example 4.4), which allows for derivation of the asymptotic distribution of a nonlinear transformation of parameters whose asymptotic distribution is known. Let $a_n(\boldsymbol{\xi}_n - \boldsymbol{c}) \xrightarrow{d} \boldsymbol{V}$ as $n\to\infty$, where a_n is a sequence of positive constants such that $a_n \to \infty$ as $n\to\infty$, $\boldsymbol{\xi}_n$ is a sequence of s-dimensional random vectors, $\boldsymbol{c}$ is a constant s-dimensional vector and $\boldsymbol{V}$ is an s-dimensional random vector. If $g(\boldsymbol{x}) : \mathbb{R}^s \to \mathbb{R}$ is a continuously differentiable function it holds that

$$a_n\Big(g(\boldsymbol{\xi}_n) - g(\boldsymbol{c})\Big) \xrightarrow{d} \left[\frac{\partial g}{\partial \boldsymbol{x}}(\boldsymbol{c})\right]^\top \boldsymbol{V} \text{ as } n\to\infty, \tag{2.8.9}$$

  where $\frac{\partial g}{\partial \boldsymbol{x}}(\boldsymbol{c}) = \left(\frac{\partial g}{\partial x_1}(\boldsymbol{c}), \ldots, \frac{\partial g}{\partial x_s}(\boldsymbol{c})\right)^\top$.
  Using $\hat{\boldsymbol{\beta}} \overset{a}{\sim} \mathcal{N}_p(\boldsymbol{\beta}, (\boldsymbol{X}^\top\boldsymbol{X})^{-1}/\nu)$ and the delta method (Equation (2.8.9)), we obtain $\hat{Y_0} = \exp(\boldsymbol{x}_0^\top\hat{\boldsymbol{\beta}}) \overset{a}{\sim} \mathcal{N}(\exp(\boldsymbol{x}_0^\top\boldsymbol{\beta}), \exp(2\boldsymbol{x}_0^\top\boldsymbol{\beta})\boldsymbol{x}_0^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{x}_0/\nu)$, from which it can be derived that $\hat{Y_0}/\exp(\boldsymbol{x}_0^\top\boldsymbol{\beta}) \overset{a}{\sim} \mathcal{N}(1, \boldsymbol{x}_0^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{x}_0/\nu)$. If we take into account the uncertainty from estimating ν, and the asymptotic independence of $\hat{\nu}$ and $\hat{\boldsymbol{\beta}}$, we can estimate the distribution of $\hat{Y_0}/\exp(\boldsymbol{x}_0^\top\boldsymbol{\beta})$ by $t_{n-p}\sqrt{\boldsymbol{x}_0^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{x}_0/\hat{\nu}}+1$, where n is the number of observations in the training data set i.e. $n = n_{tr}$ and p is the length of the vector of estimated parameters $\hat{\boldsymbol{\beta}}$ in the model. We define

$$G := \frac{\chi^2_{2\nu}}{t_{n-p}\sqrt{\frac{\boldsymbol{x}_0^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{x}_0}{\hat{\nu}}}+1}$$

where the random variables in the nominator and denominator are independent. Consequently,

$$P(G_{(\alpha/2)} \leq \frac{2\nu Y_0}{\exp(\boldsymbol{x}_0^\top \boldsymbol{\beta})} \div \frac{\hat{Y}_0}{\exp(\boldsymbol{x}_0^\top \boldsymbol{\beta})} \leq G_{(1-\alpha/2)}) = P(\frac{G_{\alpha/2}\hat{Y}_0}{2\nu} \leq Y_0 \leq \frac{G_{1-\alpha/2}\hat{Y}_0}{2\nu}),$$

where Y_0 and $\hat{Y}_0$ are independent random variables and $G_{(\alpha/2)}$ and $G_{(1-\alpha/2)}$ are the $\alpha/2$ and $1-\alpha/2$ quantiles of the G distribution. Therefore, the asymptotic $(1-\alpha)100\%$ prediction interval for y_0 is given by

$$\Big[\frac{\hat{G}_{(\alpha/2)} \exp(\boldsymbol{x}_0^\top \hat{\boldsymbol{\beta}})}{2\hat{\nu}}, \frac{\hat{G}_{(1-\alpha/2)} \exp(\boldsymbol{x}_0^\top \hat{\boldsymbol{\beta}})}{2\hat{\nu}}\Big], \tag{2.8.10}$$

where $\hat{G}$ is G with $\hat{\nu}$ replacing ν. In this thesis we find $\hat{G}_{(\alpha/2)}$ and $\hat{G}_{(1-\alpha/2)}$ using 100 000 simulations of the $\chi^2_{2\hat{\nu}}$ and t_{n-p} distribution.
For the application of this method of deriving a prediction interval for a new observation for gamma model with log link function, an important discussion is the estimation of ν. In the framework of the gamma model in this thesis, we defined that $\phi = \frac{1}{\nu}$ and the estimator of ϕ using the moments is given by $\hat{\nu}^{-1} = \hat{\phi} = \frac{1}{n-p}\sum_{i=1}^{n}\left(\frac{Y_i-\hat{\mu}_i}{\hat{\mu}_i}\right)^2$. Another estimator of ϕ is using the deviance,

$$\hat{\nu}^{-1} = \hat{\phi} = \frac{D(\hat{\boldsymbol{\mu}}, \boldsymbol{y})}{n-p}. \tag{2.8.11}$$

Wasef Hattab (2016) discusses that both these estimators are apppropriate choices for estimation of ν, however in this thesis for the derivation of the prediction intervals we use the estimator given by Equation (2.8.11), which is also the choice in the case studies in the paper.

- *Linear quantile regression model with original response* Y: Let $\hat{\boldsymbol{\beta}}(\frac{\alpha}{2})$, $\hat{\boldsymbol{\beta}}(1-\frac{\alpha}{2})$ be the vector of estimated parameters in the linear quantile regression models with $\frac{\alpha}{2}$ and $1-\frac{\alpha}{2}$ conditional quantiles respectively. Then the $(1-\alpha)100\%$ prediction interval for a future observation y_0 at location $\boldsymbol{x}_0$ is given by

$$\Big[\hat{q}_{\frac{\alpha}{2}}(\boldsymbol{x}_0), \hat{q}_{1-\frac{\alpha}{2}}(\boldsymbol{x}_0)\Big], \tag{2.8.12}$$

where $\hat{q}_{\frac{\alpha}{2}}(\boldsymbol{x}_0) = \boldsymbol{x}_0^\top \hat{\boldsymbol{\beta}}(\frac{\alpha}{2})$ and $\hat{q}_{1-\frac{\alpha}{2}}(\boldsymbol{x}_0) = \boldsymbol{x}_0^\top \hat{\boldsymbol{\beta}}(1-\frac{\alpha}{2})$. When calculating prediction intervals using linear quantile regression models, one needs to be cautious about quantile crossing which leads to unreasonable prediction intervals. In this thesis however, we do not face quantile crossing in the prediction intervals calculation.

- *Linear quantile regression model with transformed response* $\ln(Y)$: Let $\hat{\boldsymbol{\beta}}(\frac{\alpha}{2})$, $\hat{\boldsymbol{\beta}}(1-\frac{\alpha}{2})$ be the vector of estimated parameters in the linear quantile regression models with $\frac{\alpha}{2}$ and $1-\frac{\alpha}{2}$ conditional quantiles respectively. Then the $(1-\alpha)100\%$ prediction interval for a future observation y_0 at location $\boldsymbol{x}_0$ is given by

$$\Big[\exp\Big(\hat{q}_{\frac{\alpha}{2}}(\boldsymbol{x}_0)\Big), \exp\Big(\hat{q}_{1-\frac{\alpha}{2}}(\boldsymbol{x}_0)\Big)\Big], \tag{2.8.13}$$

where $\hat{q}_{\frac{\alpha}{2}}(\boldsymbol{x}_0) = \boldsymbol{x}_0^\top \hat{\boldsymbol{\beta}}(\frac{\alpha}{2})$ and $\hat{q}_{1-\frac{\alpha}{2}}(\boldsymbol{x}_0) = \boldsymbol{x}_0^\top \hat{\boldsymbol{\beta}}(1-\frac{\alpha}{2})$. Again, we do not face quantile crossings, which guarantees reasonable prediction intervals.

- *D-vine quantile regression model with original response* Y: The $(1-\alpha)100\%$ prediction interval for a future observation y_0 at location $\boldsymbol{x}_0$ is given by

$$\left[\hat{q}_{\frac{\alpha}{2}}(\boldsymbol{x}_0), \hat{q}_{1-\frac{\alpha}{2}}(\boldsymbol{x}_0)\right]. \tag{2.8.14}$$

- *D-vine quantile regression model with transformed response* $\ln(Y)$: The $(1-\alpha)100\%$ prediction interval for a future observation y_0 at location $\boldsymbol{x}_0$ is given by

$$\left[\exp\left(\hat{q}_{\frac{\alpha}{2}}(\boldsymbol{x}_0)\right), \exp\left(\hat{q}_{1-\frac{\alpha}{2}}(\boldsymbol{x}_0)\right)\right]. \tag{2.8.15}$$

Finally, let $\hat{l}^{\mathcal{M}}_{i,test}$ and $\hat{u}^{\mathcal{M}}_{i,test}$ be the lower and upper limits of the prediction interval by model $\mathcal{M}$ for the ith observation in the test data set. We define the *interval score* for the $(1-\alpha)100\%$ prediction interval as (Gneiting and Raftery 2007))

$$\begin{aligned}\widehat{IS}_\alpha = \frac{1}{n_{test}} \sum_{i=1}^{n_{test}} \Big[&(\hat{u}^{\mathcal{M}}_{i,test} - \hat{l}^{\mathcal{M}}_{i,test}) + \frac{2}{\alpha}(\hat{l}^{\mathcal{M}}_{i,test} - y_{i,test}) I\{y_{i,test} < \hat{l}^{\mathcal{M}}_{i,test}\} \\ &+ \frac{2}{\alpha}(y_{i,test} - \hat{u}^{\mathcal{M}}_{i,test}) I\{y_{i,test} > \hat{u}^{\mathcal{M}}_{i,test}\}\Big],\end{aligned} \tag{2.8.16}$$

where $y_{i,test}, i = 1, \ldots, n_{test}$ are the observations of the original response in the test data set and $I\{\cdot\}$ denotes the indicator function, which has value 1 when the condition in the brackets is satisfied and value 0 otherwise.

3 Data sets

After explaining some estimation methods and discussing statistical models in Section 2, we now apply them to a real-world data sets. In particular, we are interested in the performance of the different models in modelling severity in motor insurance. For that purpose we consider a data used for the 2017 pricing game of the French Institute of Actuaries organized on November 16, 2017. The two data sets can be found in the R package `CASdatasets`.

3.1 Introduction to data sets

The data set `pg17trainpol` contains 100 000 policies for private motor insurance and `pg17trainclaim` contains 14 243 third party liability claims related to those 100.000 policies, which occurred in a span of one year. The data set of 100 000 policies contains all the relevant collected information for the insured object and the policy holder of each policy, whereas the data set of the claims contains only the individual claim amount and the corresponding policy number.
We are interested in modelling the total claim amount occurred per insurance policy, which is why we merge the two data sets, we aggregate the claim amount per policy and we consider only the policies which notified a claim during this one year. We also add an additional column with the number of claims per policy. However, knowing only the total claim amount per policy is not sufficient because it depends on the policy's coverage, also called exposure. Since this information is not provided in our data, as coverage we take the vehicle's value, which is the replacement value of the vehicle in Euros without inflation (so it is stable from a year to another). Finally, we can define our response variable *standardized claims* as

$$Y := \frac{\text{total claim amount}}{\text{number of claims} \times \text{vehicle value}}. \tag{3.1.1}$$

Before we aggregate the claim amount, we remove the observations that have a negative claim amount, which appear when the driver's liability is not engaged so there's a legal recourse. For the lognormal models we use *ln(standardized claims)* as a response variable, for the gamma regression models we use *standardized claims*, and for the linear quantile and D-vine quantile regression models we use both the transformed response *ln(standardized claims)* and the original response *standardized claims* as a response variable.
We present the continuous variables of the data in Table 1 and the discrete variables are summarized in Table 2. Before these summaries we removed the observations with a *license age* bigger than 90 (of either driver) and observations with a *vehicle weight* 0. They are unrealistic and we can't be sure why such observations are present, therefore we remove them. This data cleaning results in a data set with 10599 observations. Additionally, we introduce the additional covariate

$$\textit{difference in duration} := \text{policy duration-policy situation duration}, \tag{3.1.2}$$

as we want to investigate whether the period duration before the current policy conditions affects our response significantly.

Continuous Variable	Name	Measurement unit	Range	Description
Response				
standr_claims	standardized claims	-	[0.0008,16.6426]	Standardized claims as described with Equation (3.1.1).
ln_standr_claims	ln(standardized claims)	-	[-7.092,2.812]	Logarithm of the standardized claims variable.
Covariates				
pol_bonus	bonus malus class	-	[0.5, 1.56]	French bonus malus system. The coefficient starts at 1 and every year without a claim it decreases by 5% until it reaches a minimum of 0.5. When a claim occurs, it increases by 25%, with maximum of 3.5.
pol_duration	policy duration	Years	[1,37]	Represents how old the policy is, accounted from the beginning of the current year.
pol_sit_duration	policy situation duration	Years	[1,20]	The policy current endorsement duration and can differ from policy duration, because of change of coverage,vehicle, drivers etc.
pol_diff_duration	difference in duration	Years	[0,33]	The difference between policy duration and policy situation duration.
drv_age1	driver 1 age	Years	[19,103]	Age of driver 1 counted from the beginning of the considered year.
drv_age2	driver 2 age	Years	[19,92]	Age of driver 2 counted from the beginning of the considered year. In policies with no driver 2, the value is 0.
drv_age_lic1	driver 1 license age	Years	[1,72]	Age of the driving licence of driver 1. It is counted from the beginning of the current year.
drv_age_lic2	driver 2 license age	Years	[1,71]	Age of driving license of driver 2. In policies with no driver 2, the value is 0.
vh_age	vehicle age	Years	[1,51]	The difference between the year of release of the vehicle and the current year.
vh_cyl	engine cylinder displacement	ml	[425,5666]	Engine cylinder displacement of the vehicle. Higher values correspond to more powerful vehicle.
vh_din	motor power	Watts	[20,507]	Motor power of the vehicle.
vh_sale_begin	vehicle sale begin	Years	[1,54]	Years from the beginning of the current year to the beginning of marketing years of the vehicle.
vh_sale_end	vehicle sale end	Years	[1,46]	Years from the beginning of the current year to the end of marketing years of the vehicle.
vh_speed	vehicle speed	km/h	[88,285]	The vehicle maximum speed, as stated by the manifacturer.
vh_weight	vehicle weight	kg	[560,3200]	Weight of the vehicle.

Table 1: Description of the continuous variables in the data set.

In Table 2 we can observe the discrete covariates with their corresponding levels and the number of observations per category. As the level "AllTrips" of the covariate *policy usage* has only 20 observations and the level "Hybrid" of the covariate *type of fuel* has 15 observations, we remove them from the data set. The resulting data set contains 10 564 observations. We define an additional covariate *claim indicator*

$$\textit{claim indicator} := \begin{cases} \text{One}, & \text{number of claims} = 1 \\ \text{MoreThanOne}, & \text{number of claims} > 1 \end{cases}, \tag{3.1.3}$$

with which we want to differentiate between single and multiple claims in a policy and investigate its effect on the response.
The covariate *vehicle carmaker* has 50 levels, so in the next step we want to reduce the number of factors in this covariate. Let $\boldsymbol{x} = (\boldsymbol{x}_1^\top, \ldots, \boldsymbol{x}_{50}^\top)^\top$ be the vector of *vehicle carmaker* sorted per factor levels, and $\boldsymbol{x}_i, i = 1, \ldots, 50$ have corresponding length $n_i, i = 1, \ldots, 50$, and let $\boldsymbol{y} = (\boldsymbol{y}_1^\top, \ldots, \boldsymbol{y}_{50}^\top)^\top$ denote their corresponding values of *standardized claims*. Then for each level we define the *mean standardized claims* per *vehicle carmaker* as

$$\bar{y}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij}, \quad i = 1, \ldots, 50, \tag{3.1.4}$$

where y_{ij} is the jth observation in $\boldsymbol{y}_i$.

Discrete Covariates	Name	Number of levels	Levels and corresponding number of observations		Description
pol_coverage	policy coverage	4	Mini	190	The coverage category of the policy. "Mini" policies cover only Third Party Liability claims, whereas "Maxi" policies covers all claims, including Damage, Theft, Windshield Breaking, Assistance etc.
			Median 1	663	
			Median 2	1388	
			Maxi	8358	
pol_pay_freq	premium payment frequency	4	Monthly	3434	The price of the insurance coverage can be paid annually, biannually, quarterly or monthly.
			Quarterly	303	
			Biannual	3041	
			Yearly	3821	
pol_payd	policy per day	2	Yes	327	A dummy which indicates whether our client has subscribed a mileage-based policy or not.
			No	10272	
pol_usage	policy usage	4	AllTrips	20	It describes what usage the driver makes from his vehicle most of time. "Retired" stands for retired people (who usually drive less often) and "All Trips" is similar to Professional (includes pro tours).
			Professional	809	
			Retired	2665	
			WorkPrivate	7105	
drv_drv2	driver 2	2	Yes	3763	A dummy indicating if there is a secondary driver in the policy.
			No	6836	
drv_sex1	driver 1 gender	2	F	4414	Gender of driver 1. "F" stands for female and "M" for male.
			M	6185	
drv_sex2	driver 2 gender	3		6836	Gender of driver 2. "" denotes the observations where driver 2 is not present, "F" denotes females and "M" denotes males.
			F	2338	
			M	1425	
vh_fuel	type of fuel	3	Diesel	6681	Fuel type of the vehicle.
			Gasoline	3903	
			Hybrid	15	
vh_make	vehicle carmaker	50	RENAULT	2685	The vehicle carmaker. We present here only the three major brands Renault, Peugeot and Citroen.
			PEUGEOT	2054	
			CITROEN	1607	
vh_type	vehicle type	2	Commercial	532	The vehicle type. There are more "Commercial" types for "Professional" policy usage than for "WorkPrivate".
			Tourism	10067	
claim_ind	claim indicator	2	MoreThanOne	1028	A dummy variable which differs between only one claim and multiple claims in the policy.
			One	9571	

Table 2: Description of the discrete covariates in the data set.

Finally, based on the quantiles of $\bar{y}_i, i = 1, \ldots, 50$ we create new covariate *class carmaker* as

$$\textit{class carmaker} := \begin{cases} \text{Class1}, & \bar{y}_i \in [0.01, 0.05] \\ \text{Class2}, & \bar{y}_i = 0.06 \\ \text{Class3}, & \bar{y}_i \in [0.07, 0.09] \end{cases} . \tag{3.1.5}$$

This is summarized in Tables 3 and 4. The disadvantage of merging levels of a covariate is that we lose information from the data. Additionally, another disadvantage of this way of merging is that we include information from another column from the data set as a criterion of merging. However, this results in effective merging since now it is more likely that this covariate is significant. In our modelling, we consider only *class carmaker* and discard *vehicle carmaker*.
Additionally, we created the covariate *gender* which combines the covariates *driver 1 gender* and *driver 2 gender* the following way:

$$\textit{gender} := \begin{cases} \text{F}, & \text{only driver 1 is present and is a female} \\ \text{M}, & \text{only driver 1 is present and is a male} \\ \text{FF}, & \text{both drivers are female} \\ \text{MF}, & \text{driver 1 is a male, driver 2 is a female} \\ \text{FM}, & \text{driver 1 is a female, driver 2 is a male} \\ \text{MM}, & \text{both drivers are male} \end{cases} \tag{3.1.6}$$

and its summary can be seen in Table 4. For modelling we discard the covariates *driver 1 gender* and *driver 2 gender* and use only *gender*.

i	Vehicle carmaker level $\boldsymbol{x}_i$	n_i	$\bar{y}_i$	i	Vehicle carmaker level $\boldsymbol{x}_i$	n_i	$\bar{y}_i$
1	EBRO	1	0.01	**26**	SUBARU	7	0.05
2	IVECO	1	0.01	**27**	VOLKSWAGEN	653	0.05
3	PORSCHE	6	0.01	**28**	BMW	189	0.06
4	BUICK	1	0.02	**29**	CITROEN	1601	0.06
5	DAIMLER	1	0.02	**30**	DAIHATSU	8	0.06
6	JAGUAR	9	0.02	**31**	DODGE	5	0.06
7	LEXUS	1	0.02	**32**	FORD	449	0.06
8	PONTIAC	1	0.02	**33**	HYUNDAI	77	0.06
9	AUDI	202	0.03	**34**	LANCIA	17	0.06
10	CHRYSLER	29	0.03	**35**	MINI	46	0.06
11	LADA VAZ	3	0.03	**36**	NISSAN	210	0.06
12	LAND ROVER	31	0.03	**37**	OPEL	378	0.06
13	MITSUBISHI	35	0.03	**38**	PEUGEOT	2052	0.06
14	SAAB	13	0.03	**39**	TOYOTA	351	0.06
15	SSANGYONG	5	0.03	**40**	DAEWOO	10	0.07
16	VOLVO	67	0.03	**41**	FIAT	262	0.07
17	JEEP	23	0.04	**42**	HONDA	69	0.07
18	MERCEDES BENZ	375	0.04	**43**	KIA	55	0.07
19	MG	3	0.04	**44**	MAZDA	62	0.07
20	SMART	15	0.04	**45**	RENAULT	2681	0.07
21	ALFA ROMEO	49	0.05	**46**	ROVER	19	0.07
22	ISUZU	2	0.05	**47**	ARO	1	0.08
23	SANTANA	8	0.05	**48**	SUZUKI	99	0.08
24	SEAT	128	0.05	**49**	CHEVROLET	31	0.09
25	SKODA	48	0.05	**50**	DACIA	175	0.09

Table 3: Levels of *vehicle carmaker* sorted by the values of *mean standardized claims*. The horizontal lines indicate the different levels of *class carmaker*.

Discrete Covariates	Name	Number of levels	Levels and corresponding number of observations		Description
class_make	class carmaker	3	Class1	1717	Vehicle carmaker merged based on the values of mean standardized claims.
			Class2	5383	
			Class3	3464	
gender	gender	6	F	2909	Driver 1 gender and driver 2 gender combined in a covariate.
			M	267	
			FF	1225	
			MF	3905	
			FM	2065	
			MM	193	

Table 4: Description of the discrete covariates *class carmaker* and *gender*.

Now we want to separate the data set into *training* and *test data set*. We will fit our models on the training data set and then evaluate and compare them on the test data set. We split them randomly, such that the training data set would contain 9564 observations, whereas the test data set 1000 observations. An important realisation is that out of 10564 observations, 7950 are in the best bonus malus class 0.5. Intuitively, we would expect that the "good" drivers which are in the best bonus malus class show different behaviour than the other drivers. For that reason we additionally separate the training and test data set to data sets where the value of bonus malus class is 0.5 and data sets where the bonus malus class is different than 0.5.
For D-vine quantile regression we consider the some of the continuous covariates as ordinal (for eg. *policy situation duration*), because they have a small number of unique values (more details in Sections 4 and 5). To make the data consistent, we remove one observation of the Bad Driver test data set, which had a value of *policy situation duration* which was

not present in the Bad Driver training data set. Finally, we summarize our final data sets in Table 5.

Data set	Name	Number of obsevations	Number of variables	Description
Good Driver Data				
data1	training data set	7214	26	The training data set of observations where the bonus malus class=0.5.
test_data1	test data set	736	26	The test data set of observations where the bonus malus class=0.5.
Bad Driver Data				
data2	training data set	2350	27	The training data set of observations where the bonus malus class≠0.5.
test_data2	test data set	263	27	The test data set of observations where the bonus malus class≠0.5.

Table 5: Final data sets we use in this thesis. In the number of variables we include the response variables *standardized claims* and *ln(standardized claims)*.The number of variables for Bad Driver Data is bigger because there we can consider *bonus malus class* as a covariate.

In Section 3.2 and 3.3 we present exploratory data analysis for Good and Bad Driver Data respectively, which we perform on the training data sets. We only present the covariates that show influence on the response variable.

3.2 Exploratory data analysis for Lognormal and Gamma regression on Good Driver Data

For lognormal and gamma regression we need to analyse the relationship between the response *ln(standardized claims)* and the covariates on the Good Driver training data set. In gamma regression, our response variable is *standardized claims*, but since we use the log link function, we want the relationship between *ln(standardized claims)* and the covariates to be linear. Therefore, the exploratory data analysis for gamma and lognormal regression coincides. We begin with the main effects.

Main effects

Continuous covariates

Scatterplots between the continuous covariates and the response *ln(standardized claims)* are presented and the relationship between the variables is analysed. Since in some of the plots the relationship is nonlinear, which can be seen by adding a smoother line to the plot, we transform the covariates with appropriate functions. In that case, we fit a linear model using only the transformed covariate with *ln(standardized claims)* as a response variable, and add its fitted values to the scatterplot, in order to compare them with the smoother line from the raw data. If the smoother line and the fitted linear model line are almost identical, we know that our transformation is adequate. The influential continuous covariates on *ln(standardized claims)* in the Good Driver training data set with their respective transformations are shown in Table 6.
Additonally, we present their scatterplots against the response *ln(standardized claims)* in Figure 3. All scatterplots are manually trimmed at the ends for better visualisation. The number of observations omitted per covariate scatterplot is shown in Table 6. For *driver*

2 age the number of ommited observations is so high because the value of *driver 2 age* is 0 in the observations where driver 2 is not present. However, as they are zeros, they do not influence the response and the relationship between *driver 2 age* and *ln(standardized claims)*, therefore we remove them from the plot.

Continuous variables	Transformations	Number of ommited observations in the scatterplots in Figure 3
pol_sit_duration	-	20
drv_age1	poly(drv_age1,2)	19
drv_age2	poly(drv_age2,2)	4.509
drv_age_lic1	-	25
drv_age_lic2	poly(drv_age_lic2,2)	18
vh_cyl	log(vh_cyl)	18
vh_din	poly(vh_din,5)	18
vh_speed	-	14
vh_weight	poly(vh_weight,3)	25

Table 6: Influential continuous covariates on *ln(standardized claims)* in the Good Driver training data set, where $\text{poly}(\cdot, n)$ is the orthogonal polynomial function of nth degree and $\log(\cdot)$ is the natural logarithm function. For some covariates no transformation was needed.

An important caution before we start modelling is to avoid multicollinearity, which is a phenomenon in which one covariate can be linearly predicted from other covariate(s) and is usually indicated by a high correlation between these two covariates, as it leads to unstability of the models to small changes of the data. For that purpose we investigate Kendall's taus (defined by Definition 2.5.5) between these continuous covariates in Table 7. We use Kendall's tau as a correlation measure, because it does not depend on the marginal distributions of the covariates, therefore is a more robust measure of correlation. We can spot that some of the Kendall's taus are high, which may lead to multicollinearity in our models. We solve this by not includuing the covariates `drv_age_lic1` and `drv_age_lic2`, since the models with `drv_age1` and `drv_age2` showed the most parsimonious fit. We include both covariates `vh_din` and `vh_speed` to our models in the beginning, however the reduced models, which showed a more parsimonious fit, do not include both covariates.

	pol_sit_ duration	drv_ age1	drv_ age2	drv_age _lic1	drv_age _lic2	vh_ cyl	vh_ din	vh_ speed	vh_ weight
pol_sit_ duration	1.00								
drv_age1	0.11	1.00							
drv_age2	-0.05	0.01	1.00						
drv_age _lic1	0.09	0.80	0.02	1.00					
drv_age _lic2	-0.05	-0.01	0.96	0.01	1.00				
vh_cyl	-0.02	-0.07	0.03	-0.03	0.03	1.00			
vh_din	-0.09	-0.06	0.06	-0.02	0.06	0.57	1.00		
vh_speed	-0.06	-0.03	0.06	-0.01	0.06	0.42	0.72	1.00	
vh_weight	-0.09	-0.08	0.05	-0.03	0.05	0.59	0.61	0.41	1.00

Table 7: Kendall's tau correlations for the influential continuous covariates in Good Driver training data set. The high Kendall's taus are marked with gray.

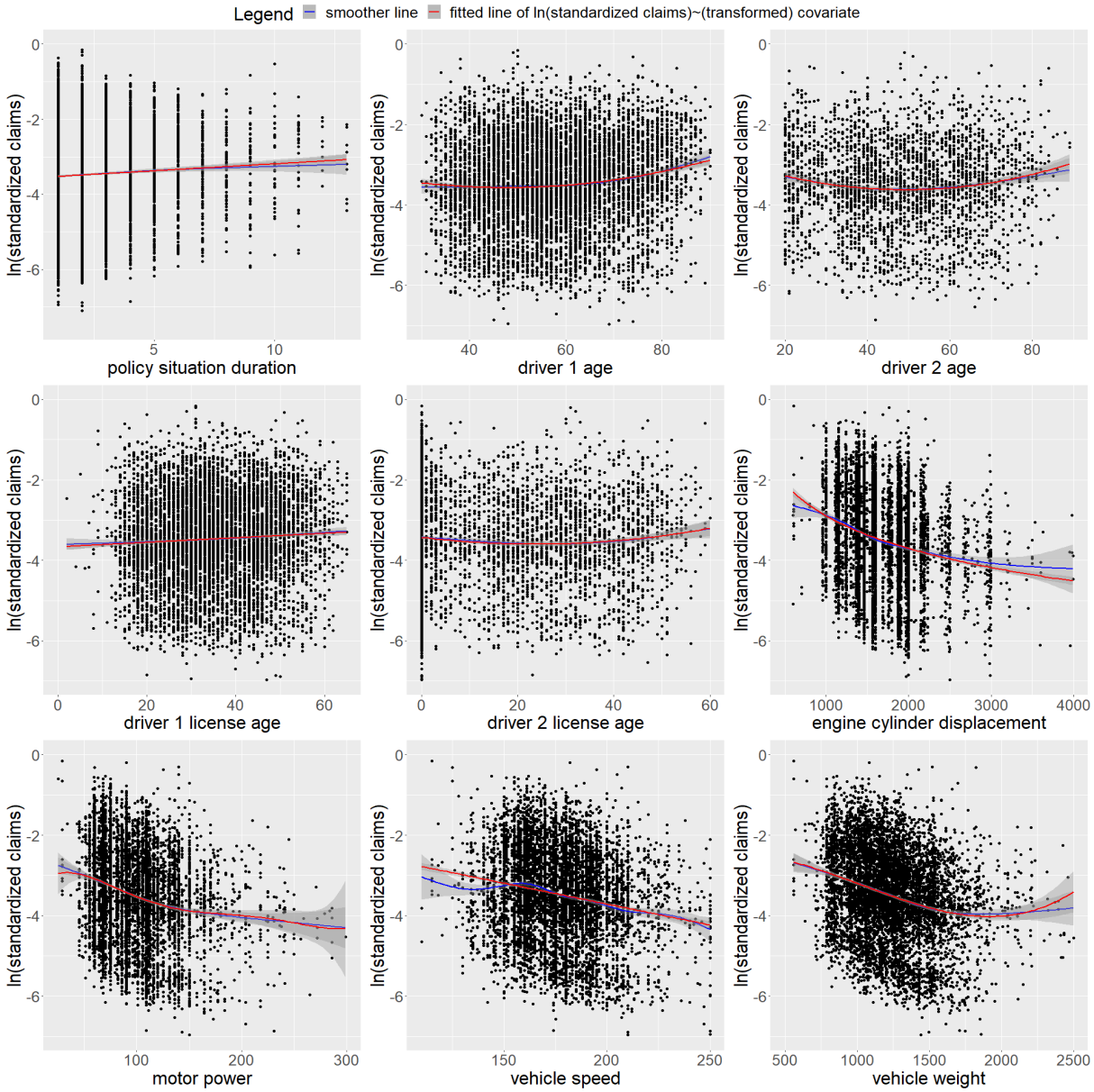

Figure 3: Trimmed scatterplots of the influential continuous variables against the response variable *ln(standardized claims)* in Good Driver training data set.

Discrete covariates

The influential covariates on the response *ln(standardized claims)* are shown in Table 8. Their respective boxplots against the response *ln(standardized claims)*, as well as the density plots of *ln(standardized claims)* for different levels of the covariates are presented in Figure 4. As it can be seen in Figure 4, the influence of the levels "Maxi", "Median1" and "Median2" on the covariate *policy coverage* on the response is similar (the medians of *ln(standardized claims)* on these levels are similar), which is why we transform it into a new covariate *policy coverage* where we merge these levels into one level "NoMini". Similarly, for the covariate *class carmaker* we merge "Class2" and "Class3" into one level "Class2&3". From this point on, when we mention the covariates *policy coverage* and *class carmaker* of the Good Driver training data set, we refer to these transformed covariates. Respectively, we will use these transformed covariates in our models.

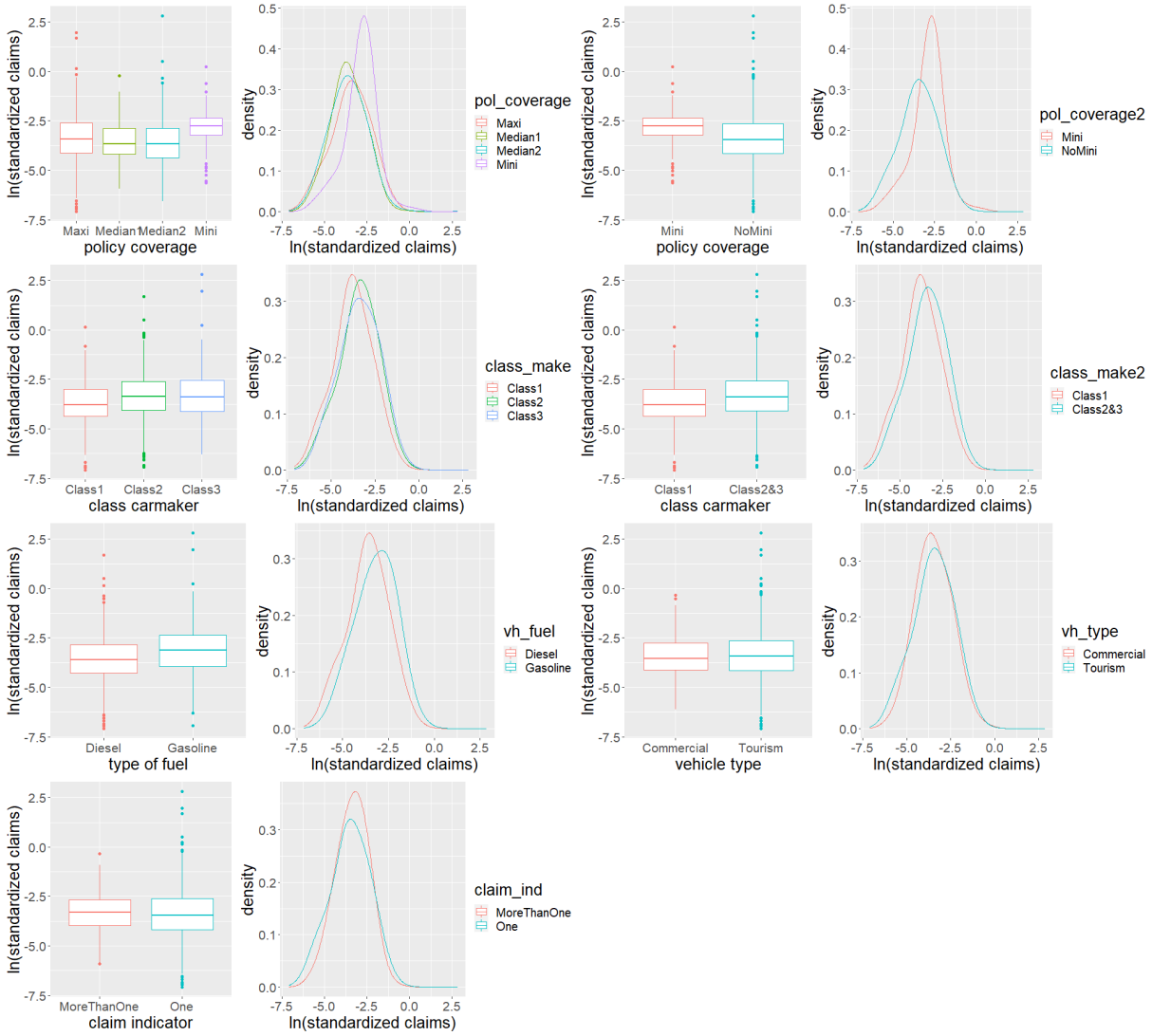

Figure 4: Boxplots and density plots of the response variable *ln(standardized claims)* for different levels of the discrete influential covariates from Table 8.

Discrete covariates	Transformed covariate	Name of the transformed covariate	Levels and corresponding number of observations of the (transformed) covariate		Description of the transformed covariate
pol_coverage	pol_coverage2	policy coverage	Mini	109	We merge the levels "Median1", "Median2" and "Maxi" in one level "No Mini".
			NoMini	7105	
class_make	class_make2	class carmaker	Class1	1131	We merge the levels "Class2" and "Class3" in one level "Class2&3".
			Class2&3	6083	
vh_fuel	-	-	Diesel	4523	-
			Gasoline	2691	
vh_type	-	-	Commercial	400	-
			Tourism	6814	
claim_ind	-	-	MoreThanOne	685	-
			One	6529	

Table 8: Influential discrete covariates on *ln(standardized claims)*. For *class carmaker* and *policy coverage* we introduce new versions which are created by merging some of their levels in one, as described in the Description column.

Interactions

We only analyse interaction terms between the covariates which are included as main effects. For the continuous covariates in polynomial form, we can include an interaction term with them with any degree of polynomial which is smaller or equal than the degree of the main effect. All the interactions that we consider in our models are summarized in Table 9.

Continuous vs Continuous	drv_age1:poly(drv_age2, 2) drv_age1:vh_speed drv_age1:poly(vh_weight,3) drv_age2:poly(vh_weight,2) log(vh_cyl):vh_speed poly(vh_din,3):vh_speed poly(vh_din,5):poly(vh_weight,3) vh_speed:poly(vh_weight,3)
Continuous vs Categorical	drv_age1:pol_coverage2 drv_age1:claim_ind poly(drv_age2, 2):pol_coverage2
Categorical vs Categorical	pol_coverage2:claim_ind

Table 9: Interactions for Good Driver Data.

The surface plots of the interactions of continuous vs continuous variables are presented on original scale, without their orthogonal polynomial transformations. We can spot an interaction term due to the nonlinearity of the surfaces. The resulting plots are presented in Figure 5.

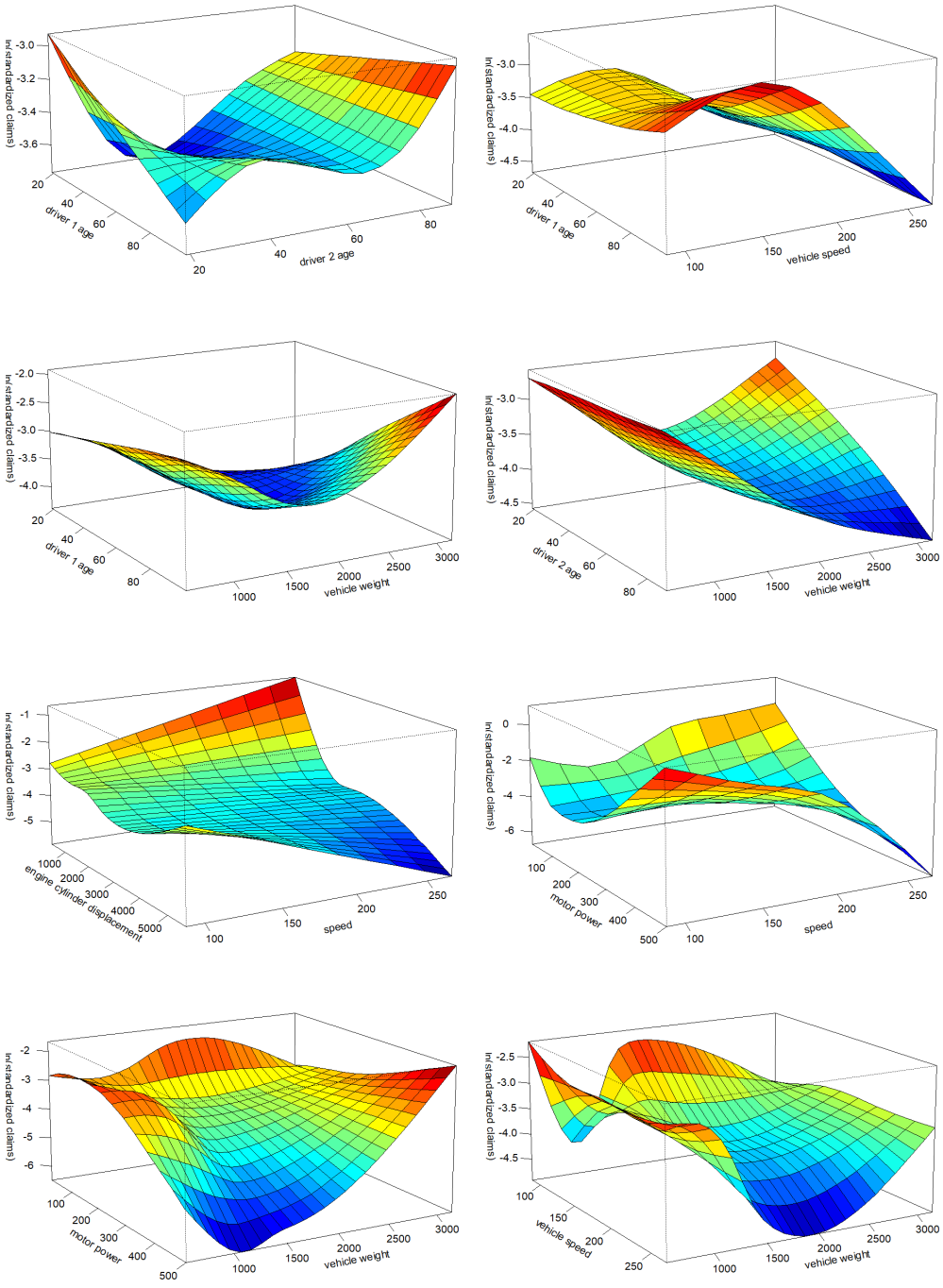

Figure 5: Plots of interaction terms between continuous covariates on original scale in Good Driver training data set, which are summarized in Table 9.

For interactions between continuous and categorical covariates, we tried to transform the continuous covariate such that the different levels of interaction look linear, but for some cases none of the transformations helped (for eg. the interaction of *driver 1 age* and *policy coverage*). Therefore, we leave them on original scale. The interaction plot of *driver 2 age* and *policy coverage*, which includes orthogonal polynomial of second degree of *driver 2 age*, is plotted such that we show how different levels of *policy coverage* affect the polynomial fit of *driver 2 age* and the response *ln(standardized claims)*. The interaction plots of continuous vs categorical and categorical vs categorical variables are presented in Figure 6, whereas in Table 10 we can see number of observations per level of the interaction between *policy coverage* and *claim indicator*. There are only 3 observations for the levels Mini and MoreThanOne, which needs to be taken into account when interpreting the boxplots and the estimated parameters in our models.

		claim indicator	
		One	MoreThanOne
policy coverage	Mini	106	3
	NoMini	6423	682

Table 10: Number of observations per level of interaction for *policy coverage* and *claim indicator*.

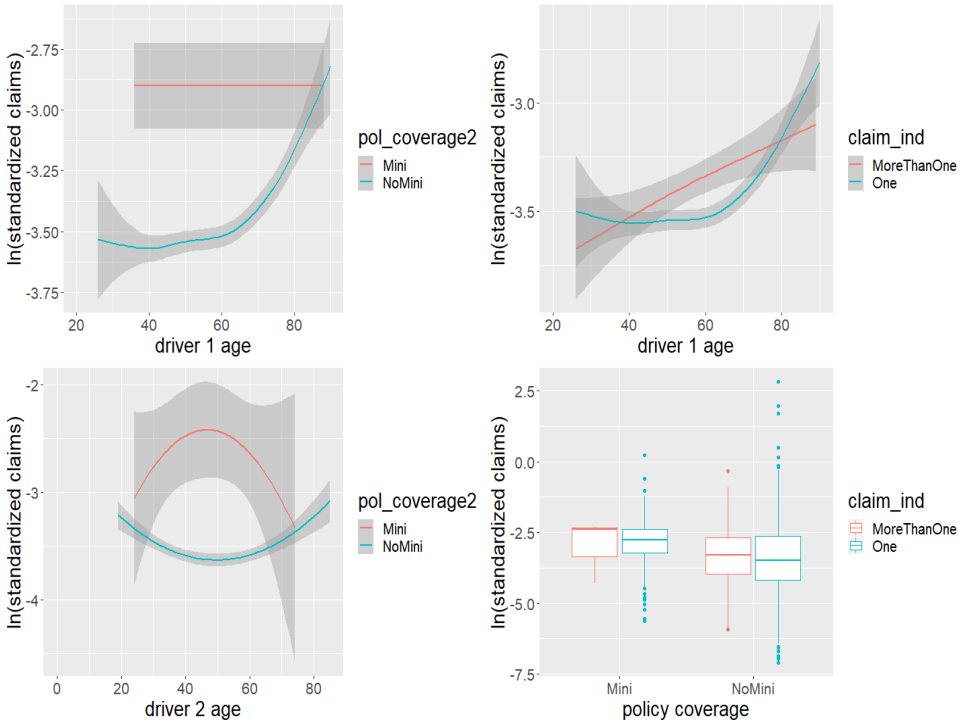

Figure 6: Plots of interaction terms for continuous vs categorical and categorical vs categorical covariates in Good Driver training data set, which are summarized in Table 9.

3.3 Exploratory data analysis for Lognormal and Gamma regression on Bad Driver Data

Similarly as for Good Driver Data, we are interested in the relationship between the response *ln(standardized claims)* and the covariates on Bad Driver training data set. In particular, we are interested that this relationship in linear. We begin with the main effects and then follow with interactions.

Main effects

Continuous covariates

We look at the scatterplots of the continuous covariates against the response *ln(standardized claims)*. Some nonlinearity can be spotted, which is why we transform certain covariates using polynomials, as these transformations showed a most parsimonious fit. To the scatterplots we add a smoother line as a reference for comparison with our fitted linear model to the (transformed) covariates and the response *ln(standardized claims)*. The influential covariates and their respective transformations can be seen in Table 11.

Continuous variables	Transformations	Number of ommited observations in the scatterplots in Figure 7
pol_bonus	-	10
pol_duration	-	2
pol_sit_duration	-	3
drv_age1	poly(drv_age1,4)	9
drv_age2	poly(drv_age2,3)	1698
drv_age_lic1	poly(drv_age_lic1,3)	4
drv_age_lic2	poly(drv_age_lic2,4)	7
vh_cyl	poly(vh_cyl,2)	6
vh_din	poly(vh_din, 3)	4
vh_speed	poly(vh_speed, 2)	3
vh_weight	poly(vh_weight,5)	9

Table 11: Influential continuous covariates on *ln(standardized claims)* in Bad Driver training data set, where poly(·, n) is the orthogonal polynomial function of nth degree. For some covariates no transformation was needed.

We present their scatterplots against the response *ln(standardized claims)* in Figure 7. All scatterplots are manually trimmed at the ends for better visualisation and the number of observations omitted per covariate scatterplot is shown in Table 6. For driver 2 age the number of ommited observations is so high because the value of driver 2 age is 0 in the observations where driver 2 is not present. However, as they are zeros, they do not influence the relationship between driver 2 age and *ln(standardized claims)*, therefore we remove them from the plot.

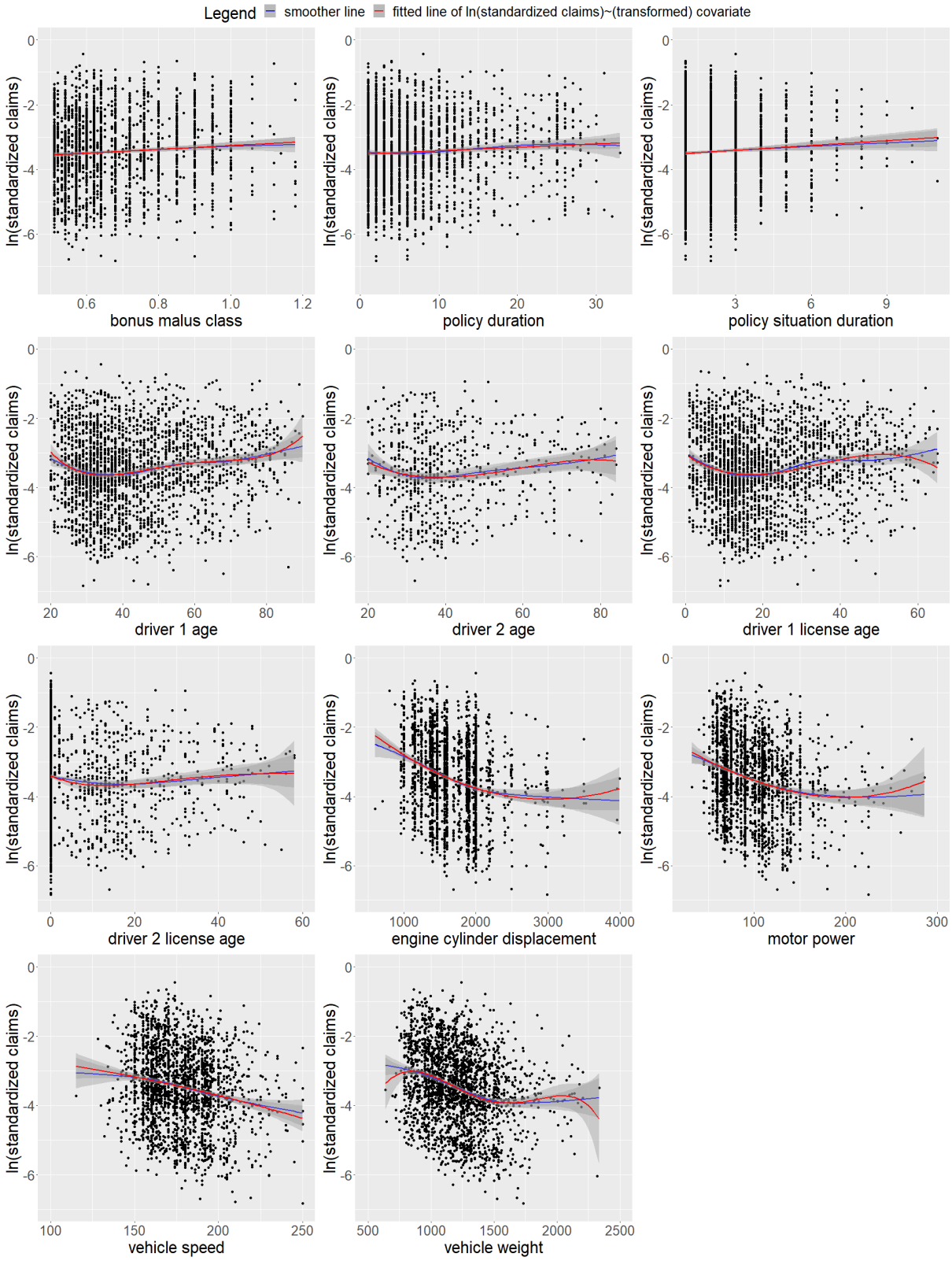

Figure 7: Trimmed scatterplots of the influential continuous variables against the response variable *ln(standardized claims)* in Bad Driver training data set.

In this training data set, multicollinearity can be spotted too. The Kendall't taus of the continuous covariates are presented in Table 12. The highest Kendall's tau is between *driver 2 age* and *driver 2 license age*, which is why we avoid using both covariates at the same time in a model. Additionally, *driver 1 age* and *driver 1 license age*, as well as *motor power* and *vehicle speed* have also high Kendall's tau values. Therefore, we do not include the covariates `drv_age1` and `drv_age_lic2` for the lognormal regression model, whereas for gamma regression we omit the covariates `drv_age_lic1` and `drv_age_lic2`. Additionally, the reduced models of both regression methods, which showed most parsimonious fits, contain `vh_din` but do not include `vh_speed`.

	pol_ bonus	pol_ duration	pol_sit_ duration	drv_ age1	drv_ age2	drv_age _lic1	drv_age _lic2	vh_ cyl	vh_ din	vh_ speed	vh_ weight
pol_ bonus	1.00										
pol_ duration	-0.20	1.00									
pol_sit_ duration	-0.07	0.27	1.00								
drv_ age1	-0.20	0.31	0.16	1.00							
drv_ age2	-0.15	0.03	-0.04	0.09	1.00						
drv_age _lic1	-0.24	0.32	0.15	0.82	0.11	1.00					
drv_age _lic2	-0.15	0.03	-0.04	0.08	0.97	0.11	1.00				
vh_cyl	-0.07	-0.01	-0.03	0.01	0.09	0.04	0.09	1.00			
vh_din	-0.08	-0.01	-0.08	0.03	0.10	0.07	0.10	0.57	1.00		
vh_speed	-0.05	-0.03	-0.07	-0.01	0.07	0.01	0.07	0.45	0.74	1.00	
vh_weight	-0.08	0.01	-0.06	0.06	0.12	0.09	0.13	0.59	0.62	0.43	1.00

Table 12: Kendall's tau correlations for the influential continuous covariates in Bad Driver training data set. The high Kendall's taus are marked with gray.

Discrete covariates

The influential discrete covariates on the response *ln(standardized claims)* are shown in Table 13. Their respective boxplots against the response *ln(standardized claims)*, as well as the density plots of *ln(standardized claims)* for different levels of the covariates are presented in Figure 8. As for Good Driver training data set, the influence of the levels "Maxi", "Median1" and "Median2" of the covariate *policy coverage* on the response is similar (the medians of *ln(standardized claims)* on these levels are similar), which is why we transform it into a new covariate *policy coverage* where we merge these levels into one level "NoMini". Similarly, for the covariate *class carmaker* we merge "Class2" and "Class3" into one level "Class2&3", for *policy usage* we merge "Professional" and "WorkPrivate" to "NoRetired", and for *gender* we merge "F" and "FF" to "FemalesOnly" and "FM", "M", "MF", "MM" to "AllOther". From this point on, when we mention the covariates *policy coverage*, *class carmaker*, *policy usage* and *gender* of the Bad Driver training data set, we refer to these transformed covariates. Respectively, we will use these transformed covariates in our models.

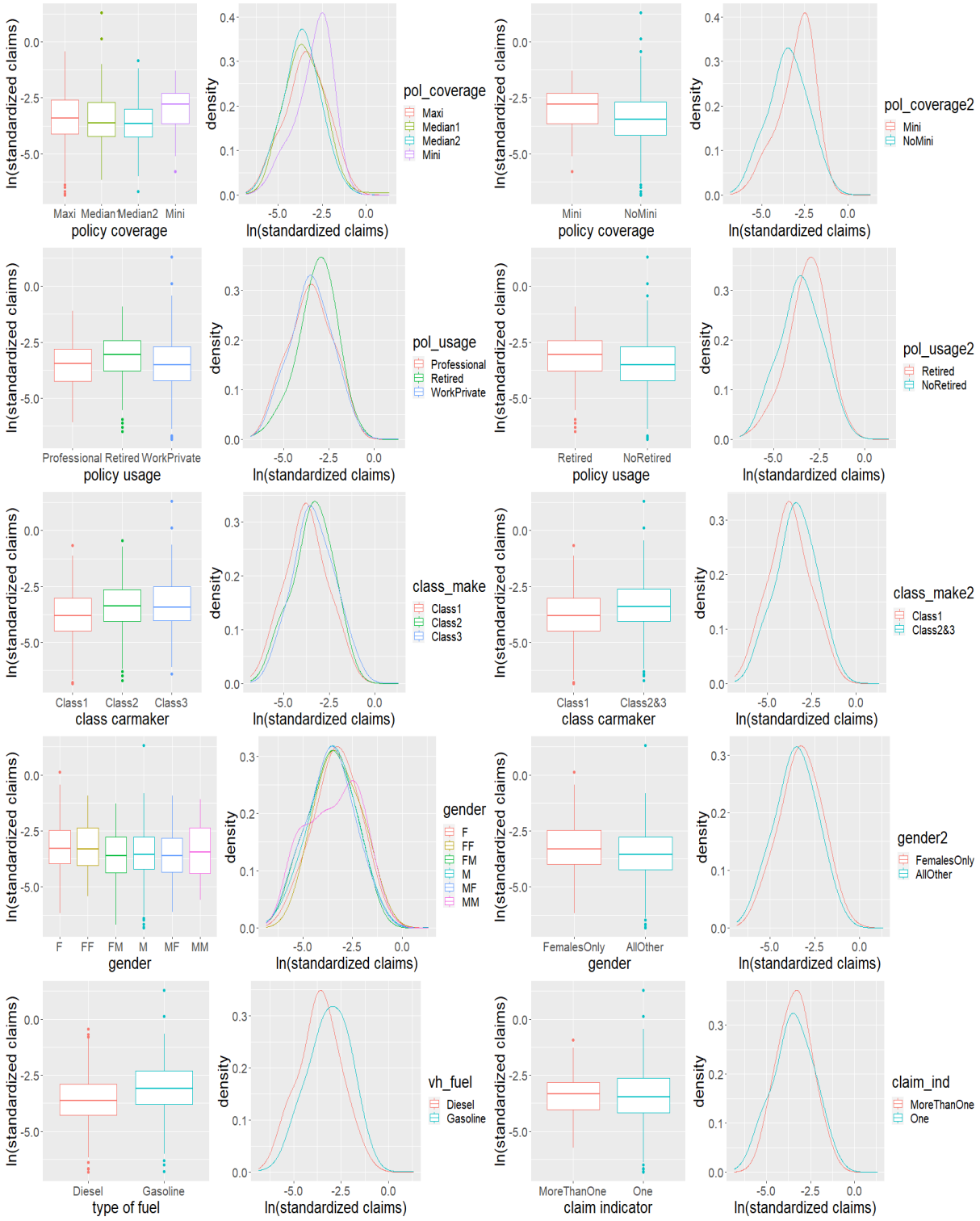

Figure 8: Boxplots and density plots of the response variable *ln(standardized claims)* for different levels of the discrete influential covariates from Table 13.

Discrete covariates	Transformed covariate	Name of the transformed covariate	Levels and corresponding number of observations of the (transformed) covariate		Description of the transformed covariate
pol_coverage	pol_coverage2	policy coverage	Mini	65	We merge the levels "Median1", "Median2" and "Maxi" in one level "NoMini".
			NoMini	2285	
pol_usage	pol_usage2	policy usage	Retired	274	We merge the levels "Professional" and "WorkPrivate" in one level "NoRetired".
			NoRetired	2076	
class_make	class_make2	class carmaker	Class1	408	We merge the levels "Class2" and "Class3" in one level "Class2&3".
			Class2&3	1942	
gender	gender2	gender	FemalesOnly	829	We merge "F" and "FF" to "FemalesOnly" and all other levels into "AllOther"
			AllOther	1521	
vh_fuel	-	-	Diesel	1522	-
			Gasoline	828	
claim_ind	-	-	One	2106	-
			MoreThanOne	244	

Table 13: Influential discrete covariates on ln(standardized claims). For *policy coverage*, *policy usage*, *class carmaker* and *gender* we introduce new versions, which are created by merging some of their levels in one, as described in the Description column.

Interactions

We only analyse interaction terms between the covariates which are included as main effects. For continuous covariates in polynomial form, we can include an interaction term with them with any degree of polynomial which is smaller or equal than the degree of the main effect. Additionally, in the models for Bad Driver training data set we don't consider any categorical vs categorical variable interaction since there seems to be no interaction between the influential categorical covariates. All the interactions that we consider in our models are summarized in Table 14.

Continuous vs Continuous	pol_bonus:poly(drv_age1,4) pol_bonus:drv_age_lic1 pol_bonus:vh_cyl pol_bonus:poly(vh_din,3) pol_bonus:poly(vh_weight,2) poly(drv_age1,4):vh_cyl poly(drv_age1,4):poly(vh_weight,2) drv_age_lic1:vh_cyl drv_age_lic1:poly(vh_din,3) drv_age_lic1:poly(vh_weight,5) vh_cyl:vh_din vh_cyl:poly(vh_weight,2) vh_din:poly(vh_weight,2)
Continuous vs Categorical	pol_bonus:pol_coverage2 pol_bonus:class_make2 drv_age_lic1:pol_coverage2 vh_din:vh_fuel

Table 14: Interactions for Bad Driver Data.

The surface plots of the interactions of continuous vs continuous variables are presented on original scale, without their orthogonal polynomial transformations. We can spot an interaction term due to the nonlinearity of the surfaces. The resulting plots are presented in Figure 9 and Figure 10.

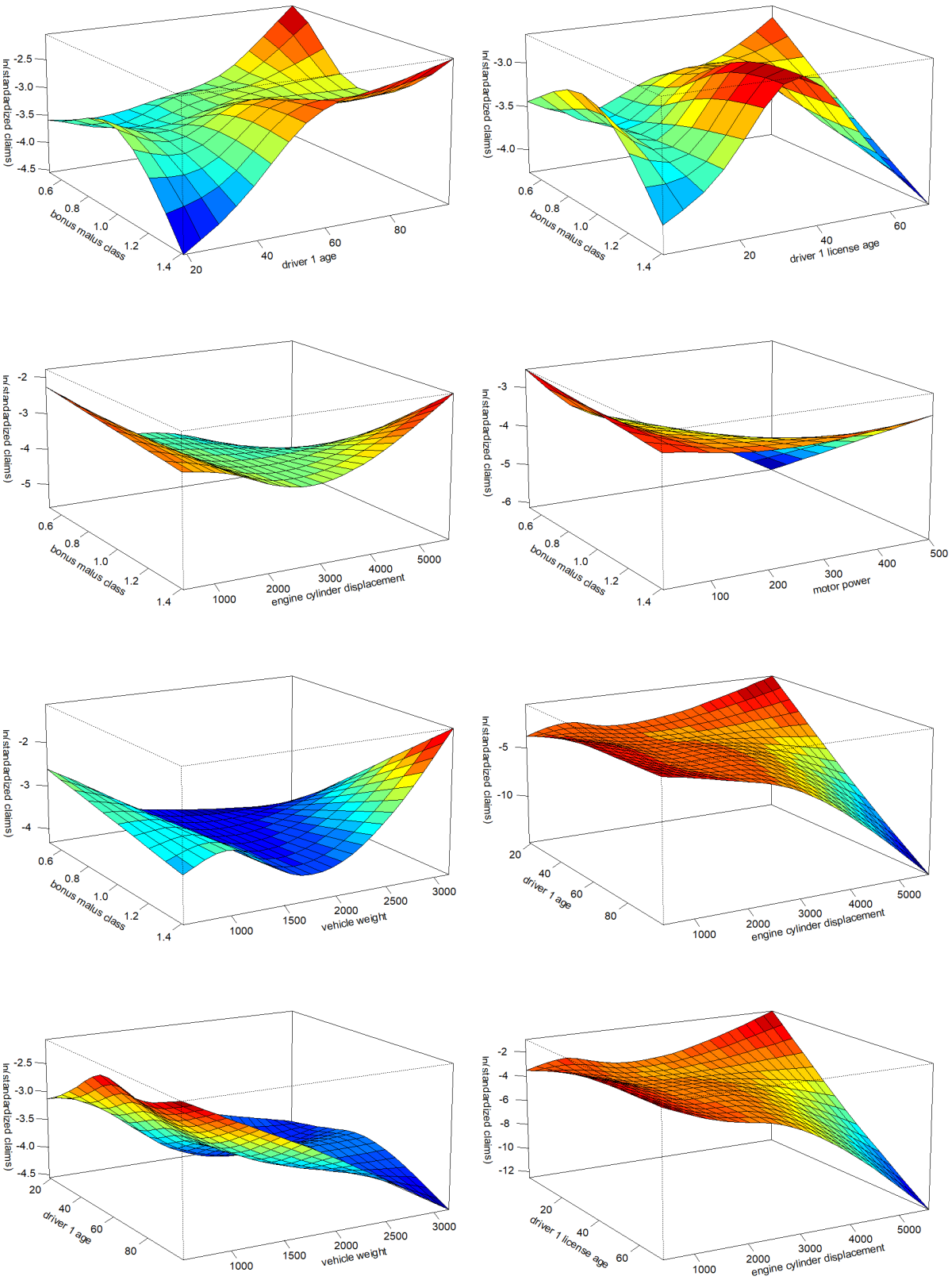

Figure 9: Plots of interaction terms between continuous covariates on original scale in Bad Driver training data set, which are summarized in Table 14.

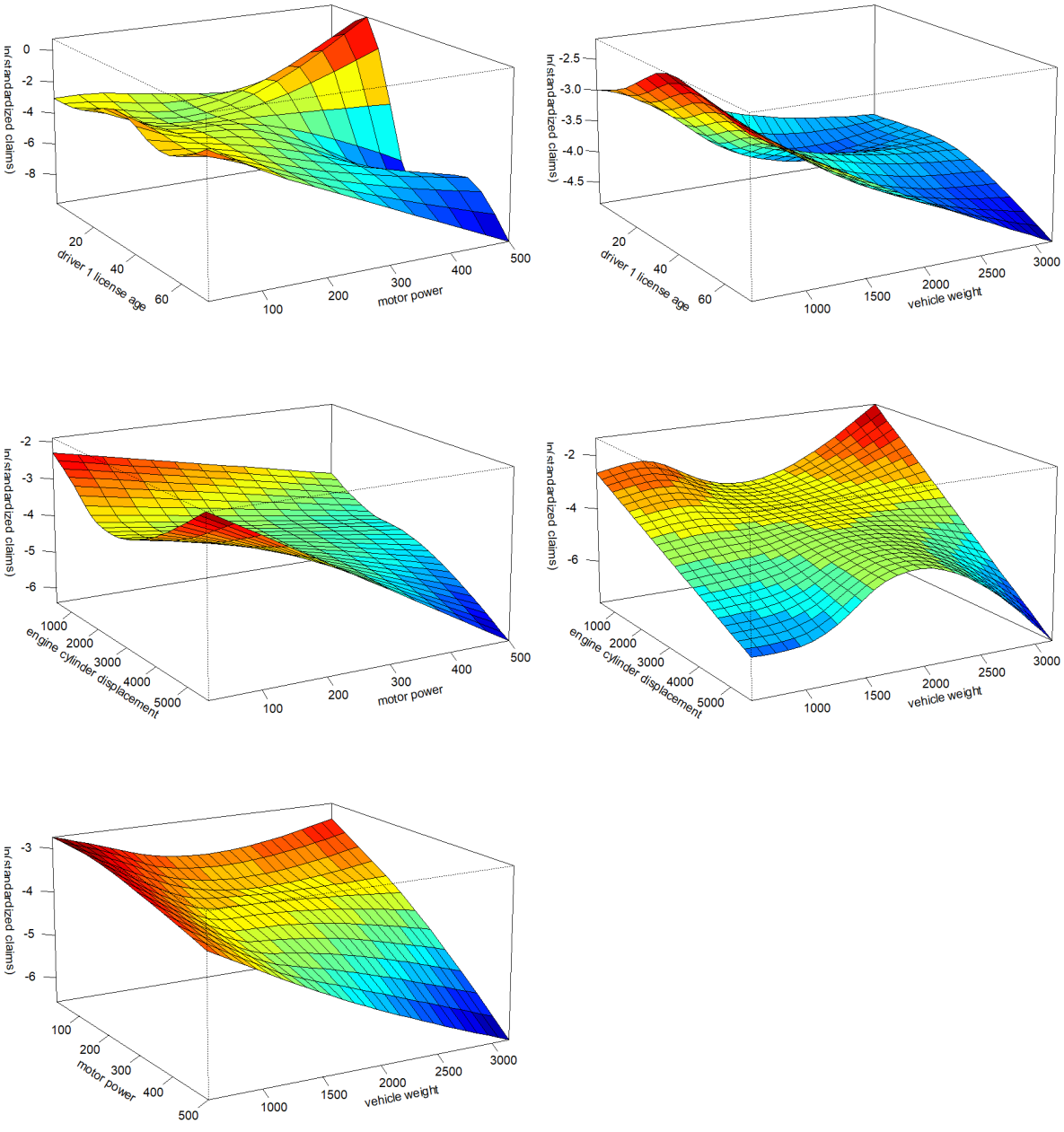

Figure 10: Plots of interaction terms between continuous covariates on original scale in Bad Driver training data set, which are summarized in Table 14.

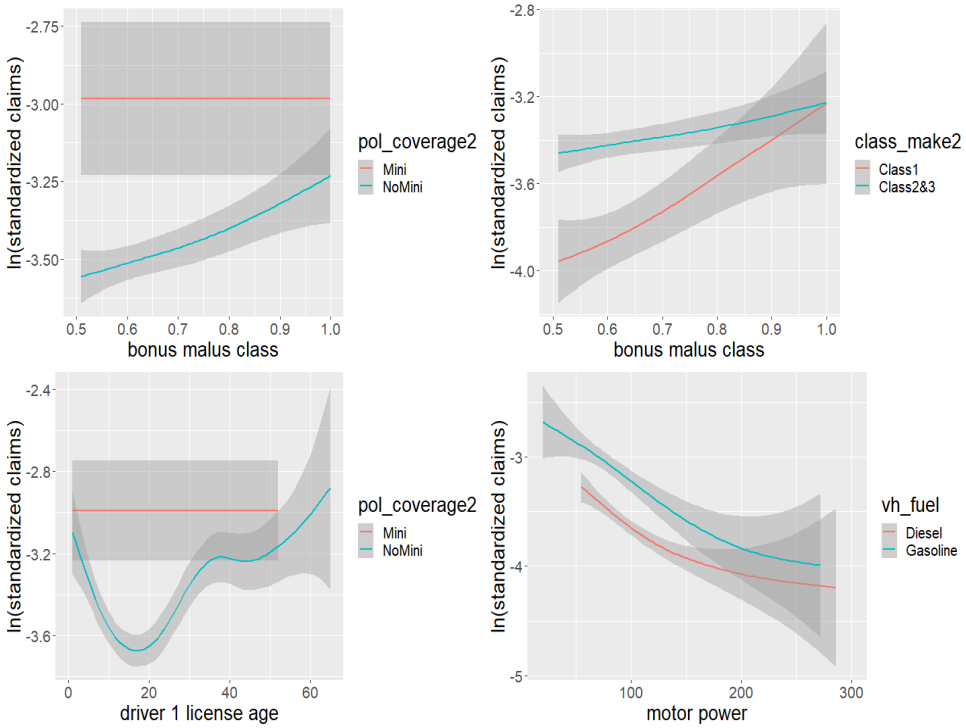

Figure 11: Plots of interaction terms for continuous vs categorical covariates in Bad Driver training data set, which are summarized in Table 14.

The interaction plot of a continuous vs categorical covariate,is plotted such that we show how different levels of the categorical covariate affect the smoother line of the continuous covariate with respect to the response *ln(standardized claims)*. We tried to transform the continuous covariate such that the different levels of interaction look linear, but for some cases none of the transformations helped (for eg. the interaction of *driver 1 license age* and *policy coverage*). Therefore, we leave them on original scale. The interaction plots of continuous vs categorical variables are presented in Figure 11.

4 Modelling on Good Driver Data

In this section we present the lognormal, gamma, linear quantile and D-vine quantile regression results for the Good Driver Data. After the exploratory data analysis in Section 3.2, we can proceed with fitting of lognormal and gamma regression with log link function. We fit D-vine regression using both the original response *standardized claims* and the transformed response *ln(standardized claims)*, with nonparametric marginals and nonparametric bivariate copulas. Consequently, we fit two linear quantile regression models using the same covariates which were selected in the D-vine regression models. At the end of the section we compare all four model approaches.

4.1 Lognormal regression model

Due to the high correlation of `drv_age_lic1` and `drv_age_lic2` with `drv_age1` and `drv_age2` respectively, we don't include `drv_age_lic1` and `drv_age_lic2` in our lognormal models. This decision was based on fitting different models by removing one of the covariates who is highly correlated to another one. Therefore, the main effects lognormal model which includes the covariates presented in Section 3.2 is:

```
LogLM1_main<-lm(ln_standr_claims~pol_sit_duration+poly(drv_age1,2)+
              poly(drv_age2,2)+log(vh_cyl)+poly(vh_din,5)+vh_speed+
              poly(vh_weight,3)+pol_coverage2+
              vh_fuel+claim_ind+class_make2+vh_type, data=data1)
```

$R^2 = 9.31\%$, $R^2_{adj} = 9.06\%$	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.08	0.88	0.09	0.93
pol_sit_duration	0.00	0.01	0.51	0.61
poly(drv_age1, 2)1	5.71	1.23	4.66	0.00
poly(drv_age1, 2)2	3.80	1.10	3.47	0.00
poly(drv_age2, 2)1	-0.77	1.09	-0.70	0.48
poly(drv_age2, 2)2	-1.00	1.19	-0.84	0.40
log(vh_cyl)	-0.37	0.10	-3.50	0.00
poly(vh_din, 5)1	-5.90	3.87	-1.53	0.13
poly(vh_din, 5)2	0.14	1.42	0.10	0.92
poly(vh_din, 5)3	-0.82	1.22	-0.67	0.50
poly(vh_din, 5)4	-0.57	1.16	-0.49	0.62
poly(vh_din, 5)5	3.58	1.10	3.25	0.00
vh_speed	-0.00	0.00	-1.17	0.24
poly(vh_weight, 3)1	-7.03	2.39	-2.94	0.00
poly(vh_weight, 3)2	2.77	1.41	1.96	0.05
poly(vh_weight, 3)3	1.82	1.29	1.41	0.16
pol_coverage2NoMini	-0.54	0.11	-5.08	0.00
vh_fuelGasoline	0.17	0.04	4.55	0.00
claim_indOne	-0.16	0.04	-3.68	0.00
class_make2Class2&3	0.10	0.04	2.75	0.01
vh_typeTourism	0.01	0.06	0.19	0.85

Table 15: Maximum likelihood estimates, estimated standard errors, t-values and corresponding p-values, as well as R^2 and R^2_{adj} for the model `LogLM1_main`.

As already discussed, in the case when a high degree term of a polynomial shows significance but lower degree terms are not significant, we keep all lower degree terms of the polynomial in the model too, which is the case of the covariate `vh_din`. The advantage of using orthogonal polynomials in our models is that the significance of the parameters of a polynomial can be interpreted as significance of each term isolated from the influence of all lower terms of the polynomial. The overall fit of this model is not very good, since the R^2_{adj} value is less than 10%. Additionally, we can see that some of the main effects are not significant, which is why we reduce the covariates of the model using the `step` function with BIC criterion. The backward model however does not include the covariate `class_make2`, which shows significance in `LogLM1_main`, therefore we manually add it to the model. The resulting model is:

```
step(LogLM1_main,direction="backward",data=data1,k=ln(nrow(data1)))
LogLM1_main_red<-lm(ln_standr_claims ~ poly(drv_age1, 2)+
                    log(vh_cyl)+vh_speed+poly(vh_weight,3)+
                    pol_coverage2+vh_fuel+claim_ind+class_make2,
                    data = data1)
```

$R^2 = 9.11\%$, $R^2_{adj} = 8.97\%$	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.72	0.68	1.06	0.29
poly(drv_age1, 2)1	5.50	1.10	5.00	0.00
poly(drv_age1, 2)2	3.91	1.09	3.60	0.00
log(vh_cyl)	-0.40	0.09	-4.25	0.00
vh_speed	-0.00	0.00	-4.69	0.00
poly(vh_weight, 3)1	-9.71	1.76	-5.52	0.00
poly(vh_weight, 3)2	2.94	1.25	2.36	0.02
poly(vh_weight, 3)3	2.87	1.09	2.63	0.01
pol_coverage2No Mini	-0.54	0.11	-5.11	0.00
vh_fuelGasoline	0.15	0.03	4.27	0.00
claim_indOne	-0.16	0.04	-3.69	0.00
class_make2Class2&3	0.10	0.04	2.70	0.01

Table 16: Maximum likelihood estimates, estimated standard errors, t-values and corresponding p-values, as well as R^2 and R^2_{adj} for the model `LogLM1_main_red`.

R^2_{adj} is almost the same in model `LogLM1_main_red` compared to model `LogLM1_main`, but all of the covariates in the model `LogLM1_main_red` are significant. Therefore, `LogLM1_main_red` shows much more parsimonious fit. Now we try to improve the model `LogLM1_main_red` by allowing interaction effects to enter the model:

```
LogLM1_inter<-lm(ln_standr_claims~poly(drv_age1, 2)+log(vh_cyl) +
               vh_speed + poly(vh_weight, 3)+pol_coverage2+vh_fuel+
               claim_ind+class_make2+drv_age1:vh_speed+
               drv_age1:poly(vh_weight,3)+log(vh_cyl):vh_speed+
               vh_speed:poly(vh_weight,3)+drv_age1:pol_coverage2+
               drv_age1:claim_ind+pol_coverage2:claim_ind, data=data1)
```

$R^2 = 9.27\%$, $R^2_{adj} = 9.00\%$	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2.40	4.05	0.59	0.55
poly(drv_age1, 2)1	-21.28	15.01	-1.42	0.16
poly(drv_age1, 2)2	3.70	1.11	3.32	0.00
log(vh_cyl)	-0.66	0.54	-1.21	0.23
vh_speed	-0.02	0.02	-0.74	0.46
poly(vh_weight, 3)1	21.07	16.08	1.31	0.19
poly(vh_weight, 3)2	-25.36	10.68	-2.37	0.02
poly(vh_weight, 3)3	-1.36	9.41	-0.14	0.89
pol_coverage2No Mini	-1.07	0.74	-1.46	0.15
vh_fuelGasoline	0.13	0.04	3.68	0.00
claim_indOne	-0.06	0.66	-0.09	0.93
class_make2Class2&3	0.10	0.04	2.54	0.01
vh_speed:drv_age1	0.00	0.00	1.38	0.17
poly(vh_weight, 3)1:drv_age1	-0.09	0.10	-0.89	0.37
poly(vh_weight, 3)2:drv_age1	0.17	0.10	1.63	0.10
poly(vh_weight, 3)3:drv_age1	0.04	0.09	0.39	0.70
log(vh_cyl):vh_speed	0.00	0.00	0.43	0.67
vh_speed:poly(vh_weight, 3)1	-0.15	0.09	-1.72	0.08
vh_speed:poly(vh_weight, 3)2	0.12	0.05	2.21	0.03
vh_speed:poly(vh_weight, 3)3	0.00	0.05	0.10	0.92
pol_coverage2No Mini:drv_age1	0.01	0.01	1.25	0.21
claim_indOne:drv_age1	-0.00	0.00	-0.22	0.83
pol_coverage2No Mini:claim_indOne	-0.06	0.64	-0.09	0.93

Table 17: Maximum likelihood estimates, estimated standard errors, t-values and corresponding p-values, as well as R^2 and R^2_{adj} for the model `LogLM1_inter`.

The fit of the model `LogLM1_inter` is not much better than `LogLM1_main_red`, because the R^2_{adj} is almost the same and a lot of the covariates are insignificant. Using again the backward selection on the model `LogLM1_inter` with the help of the `step` function and BIC criterion

```
step(LogLM1_inter,direction="backward",data=data1,
    scope=list(upper= LogLM1_inter, lower= LogLM1_main_red),
       k=ln(nrow(data1)))
```

we get the main effects model `LogLM1_main_red`. Therefore, none of the interaction effects were kept by using the backward selection. In order to statistically test the importance of the interactions we need to perform an F-test for the models `LogLM1_main_red` and `LogLM1_inter`. This is done using the `anova` function in R. The resulting p-value is not smaller than 0.05, which tells us that none of the interaction effects can be considered as statistically significant at 0.05 level:

```
anova(LogLM1_main_red,LogLM1_inter)
```

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	7202	8432				
2	7191	8418	11	14.36	1.12	0.34

Table 18: F-test for the interaction effects in model `LogLM1_inter`.

However, if we manually try to remove the insignificant interactions on 0.1 level and leave the significant ones in the model, we end up with the reduced interaction effects model which adds only one interaction effect:

```
LogLM1_inter_red<-lm(ln_standr_claims~poly(drv_age1, 2) + log(vh_cyl) +
              vh_speed + poly(vh_weight, 3) + pol_coverage2 + vh_fuel +
              claim_ind+vh_speed:poly(vh_weight,2), data=data1)
```

$R^2 = 9.18\%$, $R^2_{adj} = 9.03\%$	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.78	0.68	1.15	0.25
poly(drv_age1, 2)1	5.57	1.10	5.07	0.00
poly(drv_age1, 2)2	3.83	1.09	3.52	0.00
log(vh_cyl)	-0.44	0.09	-4.66	0.00
vh_speed	-0.00	0.00	-2.37	0.02
poly(vh_weight, 2)1	17.53	8.94	1.96	0.05
poly(vh_weight, 2)2	-16.31	8.05	-2.03	0.04
pol_coverage2No Mini	-0.56	0.11	-5.25	0.00
vh_fuelGasoline	0.13	0.03	3.70	0.00
claim_indOne	-0.16	0.04	-3.65	0.00
class_make2Class2&3	0.10	0.04	2.56	0.01
vh_speed:poly(vh_weight, 2)1	-0.16	0.05	-3.13	0.00
vh_speed:poly(vh_weight, 2)2	0.12	0.05	2.53	0.01

Table 19: Maximum likelihood estimates, estimated standard errors, t-values and corresponding p-values, as well as R^2 and R^2_{adj} for the model `LogLM1_inter_red`.

We can remove all these interactions, because performing an F-test for the removed covariates using the ANOVA table results in p-value of 0.75, which means that there is no statistical evidence that either one of them is significant. If we compare the model `LogLM1_inter_red` to `LogLM1_main_red`, we can spot that the R^2_{adj} of `LogLM1_inter_red` is slightly bigger and the interaction term is significant. We can test the significance of this term using the F-test again, which confirms that the interaction parameters are significant:

```
anova(LogLM1_main_red,LogLM1_inter_red)
```

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	7202	8432				
2	7201	8426	1	6.43	5.49	0.02

Table 20: F-test for the interaction term in model `LogLM1_inter_red`.

Finally, in our later model and methods comparison, for Good Driver Data we will consider the lognormal models `LogLM1_main_red` and `LogLM1_inter_red` as best lognormal models. We can see that the model `LogLM1_inter_red` has slightly better R^2_{adj} value, however since they do not differentiate much, we would like to know the performace of both models. For model diagnostics however, we focus only on `LogLM1_inter_red`.

The plot of the internally studentized residuals of `LogLM1_inter_red` is given in Figure 12. If the model assumptions are satisfied, these residuals are approximately $\mathcal{N}(0,1)$-distributed, therefore almost all of the observations should lie on the interval [-3,3]. As observed, they randomly fluctuate around zero and only 8 observations are outside of the interval [-3,3]. Therefore, there is no statistical evidence against the model assumptions. The blue line at 0 is the smoother line of the residuals, whereas the red line around the 0 is the vertical line $y = 0$. The red dashed lines denote the values $y = -3$ and $y = 3$.

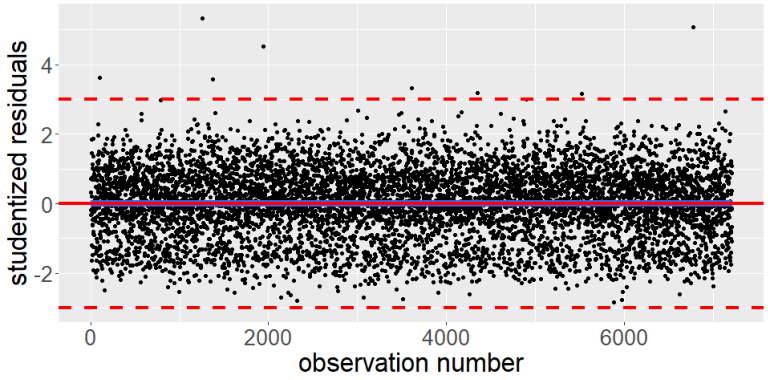

Figure 12: Internally studentized residuals of the model `LogLM1_inter_red`.

Additionally, in Figure 13 we observe the Q-Q plot of the internally studentized residuals, where the values follow the 45° line. We can spot that the distribution shows right skewness and heavier right tail than the normal distribution, but only for few observations.

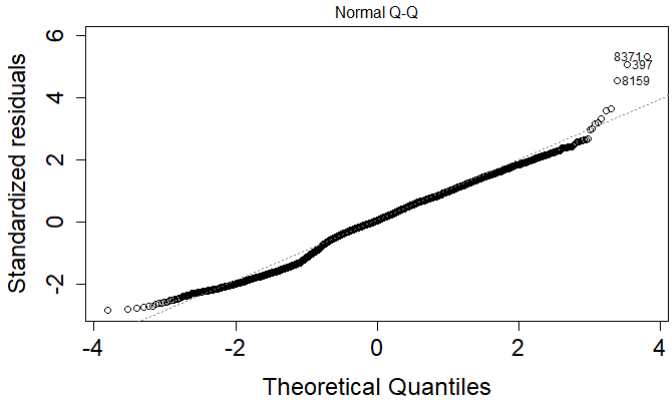

Figure 13: Q-Q plot of the internally studentized residuals of the model `LogLM1_inter_red`.

The model has 408 high leverage points, but since these show only x-outliers we don't consider them as influential observations. What is important that none of the Cook's Distance values is bigger than 1, so there are no influential points that need to be studied or removed. Finally, we can conclude that there is no evidence against the lognormal regression assumptions.

4.2 Gamma regression model

Similarly as for the lognormal regression models, we don't consider the covariates `drv_age_lic1` and `drv_age_lic2` due to their high correlation with `drv_age1` and `drv_age2` respectively. This decision was based on fitting different models by removing one of the covariates who is highly correlated to another one. The main effects gamma model which includes the covariates presented in Section 3.2 is:

```
GamReg1_main<-glm(standr_claims ~ pol_sit_duration+poly(drv_age1,2)+
                poly(drv_age2,2)+log(vh_cyl)+poly(vh_din,5)+
                poly(vh_speed,5)+poly(vh_weight,3)+pol_coverage2+
                vh_fuel+claim_ind+class_make2+vh_type, data=data1,
                family=Gamma(link="log"))
```

$\phi = 6.07$	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.80	2.00	0.40	0.69
pol_sit_duration	0.02	0.02	1.14	0.26
poly(drv_age1, 2)1	5.66	2.79	2.03	0.04
poly(drv_age1, 2)2	0.28	2.49	0.11	0.91
poly(drv_age2, 2)1	2.81	2.49	1.13	0.26
poly(drv_age2, 2)2	-4.63	2.72	-1.70	0.09
log(vh_cyl)	-0.51	0.24	-2.15	0.03
poly(vh_din, 5)1	-6.38	8.81	-0.72	0.47
poly(vh_din, 5)2	-1.28	3.24	-0.40	0.69
poly(vh_din, 5)3	0.02	2.79	0.01	0.99
poly(vh_din, 5)4	0.75	2.64	0.29	0.77
poly(vh_din, 5)5	1.97	2.50	0.79	0.43
vh_speed	0.00	0.00	0.15	0.88
poly(vh_weight, 3)1	-10.76	5.45	-1.97	0.05
poly(vh_weight, 3)2	6.11	3.22	1.90	0.06
poly(vh_weight, 3)3	-1.06	2.95	-0.36	0.72
pol_coverage2NoMini	-0.28	0.24	-1.16	0.25
vh_fuelGasoline	0.10	0.08	1.18	0.24
claim_indOne	0.09	0.10	0.90	0.37
class_make2Class2&3	0.11	0.09	1.27	0.21
vh_typeTourism	0.03	0.14	0.20	0.84

Table 21: Maximum likelihood estimates , estimated standard errors, Wald ratios and corresponding p-values for the model `GamReg1_main`.

The estimate for `vh_speed` is shown as 0.00 due to the rounding to two decimals. The model's residual deviance is 8313.4 on 7193 degrees of freedom. If we try to remove the insignificant covariates using the `step` function with BIC criterion, we obtain again a model in which a lot of covariates are not significant. Therefore, we present a reduced main effects model where we manually removed the insignificant covariates from the model `GamReg1_main`:

```
GamReg1_main_red<-glm(standr_claims ~ drv_age1+poly(drv_age2,2)+
                  log(vh_cyl)+poly(vh_weight,2), data=data1,
                  family=Gamma(link="log"))
```

$\phi = 6.25$	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2.04	1.43	1.43	0.15
drv_age1	0.01	0.00	2.51	0.01
poly(drv_age2, 2)1	2.63	2.51	1.05	0.29
poly(drv_age2, 2)2	-4.95	2.75	-1.80	0.07
log(vh_cyl)	-0.71	0.19	-3.72	0.00
poly(vh_weight, 2)1	-14.42	3.92	-3.68	0.00
poly(vh_weight, 2)2	7.51	2.61	2.87	0.00

Table 22: Maximum likelihood estimates , estimated standard errors, Wald ratios and corresponding p-values for the model `GamReg1_main_red`.

The model `GamReg1_main_red` has residual deviance 8373.2 on 7207 degrees of freedom and the significance of the covariates is much better. If we perform partial deviance test for the removed covariates from `GamReg1_main`, we obtain the p-value 0.77, which indicates that the null hypothesis cannot be rejected i.e. we can proceed with the reduced main effects model `GamReg1_main_red`. Finally, performing a residual deviance test on the model `GamReg1_main_red` results in p-value 1, therefore the model shows no lack of fit.
In the next step we want to improve the fit using interaction effects. We include the interaction `drv_age1:poly(vh_weight,3)` as `drv_age1:poly(vh_weight,2)` because the polynomial of `vh_weight` in the main effects is of second degree.

```
GamReg1_inter<-glm(standr_claims ~ drv_age1+poly(drv_age2,2)+
              log(vh_cyl)+poly(vh_weight,2)+drv_age1:poly(drv_age2,2)+
              drv_age1:poly(vh_weight,2)+drv_age2:poly(vh_weight,2),
              data=data1, family=Gamma(link="log"))
```

$\phi = 5.30$	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.98	1.32	1.50	0.13
drv_age1	0.01	0.00	2.82	0.00
poly(drv_age2, 2)1	-10.32	15.66	-0.66	0.51
poly(drv_age2, 2)2	10.26	11.76	0.87	0.38
log(vh_cyl)	-0.71	0.18	-4.02	0.00
poly(vh_weight, 2)1	-6.51	11.48	-0.57	0.57
poly(vh_weight, 2)2	-7.50	12.03	-0.62	0.53
drv_age1:poly(drv_age2, 2)1	0.25	0.25	0.99	0.32
drv_age1:poly(drv_age2, 2)2	-0.28	0.18	-1.52	0.13
drv_age1:poly(vh_weight, 2)1	-0.12	0.19	-0.61	0.54
drv_age1:poly(vh_weight, 2)2	0.23	0.20	1.18	0.24
poly(vh_weight, 2)1:drv_age2	-0.05	0.10	-0.52	0.61
poly(vh_weight, 2)2:drv_age2	0.05	0.09	0.55	0.58

Table 23: Maximum likelihood estimates , estimated standard errors, Wald ratios and corresponding p-values for the model `GamReg1_inter`.

The summary output of the model is given by Table 23. The residual deviance is 8338.0 on 7201 degrees of freedom. Many of covariates are insignificant, so we would like to reduce this model as well. If we perform backward stepwise selection of the model using BIC criterion with `GamReg1_main_red` as a lower limit, we get the model `GamReg1_main_red`. If

we perform partial deviance test for the interaction terms in `GamReg1_inter`, we obtain a p-value of 0.36, hence the interactions do not show statistical significance at a 0.05 level. By removing the interactions `drv_age1:poly(vh_weight,2)` and `drv_age2:poly(vh_weight, 2)` from the model `GamReg1_inter` we obtain the reduced model

```
GamReg1_inter_red<-glm(standr_claims ~ drv_age1+poly(drv_age2,2)+
                    log(vh_cyl)+poly(vh_weight,2)+drv_age1:poly(drv_age2, 2),
                    data=data1, family=Gamma(link="log"))
```

$\phi = 5.49$	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2.02	1.34	1.51	0.13
drv_age1	0.01	0.00	2.92	0.00
poly(drv_age2, 2)1	-11.92	15.85	-0.75	0.45
poly(drv_age2, 2)2	9.66	11.92	0.81	0.42
log(vh_cyl)	-0.72	0.18	-4.00	0.00
poly(vh_weight, 2)1	-14.06	3.67	-3.83	0.00
poly(vh_weight, 2)2	7.32	2.45	2.99	0.00
drv_age1:poly(drv_age2, 2)1	0.27	0.25	1.07	0.29
drv_age1:poly(drv_age2, 2)2	-0.28	0.19	-1.49	0.14

Table 24: Maximum likelihood estimates , estimated standard errors, Wald ratios and corresponding p-values for the model `GamReg1_inter_red`.

whose residual deviance is 8352.2 on 7205 degrees of freedom. The interaction effect in this model shows significance on 0.15 level, which is not particularly strong. Partial deviance test for the removed terms from the model `GamReg1_inter` results in p-value of 0.61, therefore we can proceed with the reduced model `GamReg1_inter_red`. Performing a residual deviance test for this interaction with the help of the model `GamReg1_main_red`, we obtain a p-value of 0.15, which indicates that we cannot reject the null hypothesis on 0.05 level, however the p-value is also not particularly large so we would like to investigate the performance of this model as well on the test data set. The residual deviance test for `GamReg1_inter_red` results in a p-value of 1, therefore the model shows no lack of fit. Finally, as gamma regression models for Good Driver Data we focus on `GamReg1_main_red` and `GamReg1_inter_red`.

4.3 D-vine quantile regression model

We fit two D-vine regression models using the response variable on original scale *standardized claims* and the transformed response variable *ln(standardized claims)*. For Good Driver Data we have fully nonparametric approach i.e. we estimate the marginals using kernel smoothing estimator which in implemented in the R package `kde1d` (Nagler and Vatter (2022)) and then fit a D-vine regression model using nonparametric bivariate copulas. This nonparametric estimator is implemented in the R package `vinereg` (Nagler (2022)), which we use to fit these D-vine regression models, however before we fit the models we look at the histograms of the marginals of the copula data, which we obtain using the R package `kde1d` (Nagler and Vatter (2022)).

In our D-vine models we consider all the covariates, except `drv_age2` and `drv_age_lic2`, since they contain a lot of zeros for the observations where driver 2 is not present. First we present the empirical normalized contour plots of our continuous variables in Figure 14, which are obtained by transforming the original data to copula data scale using marginal empirical distributions.

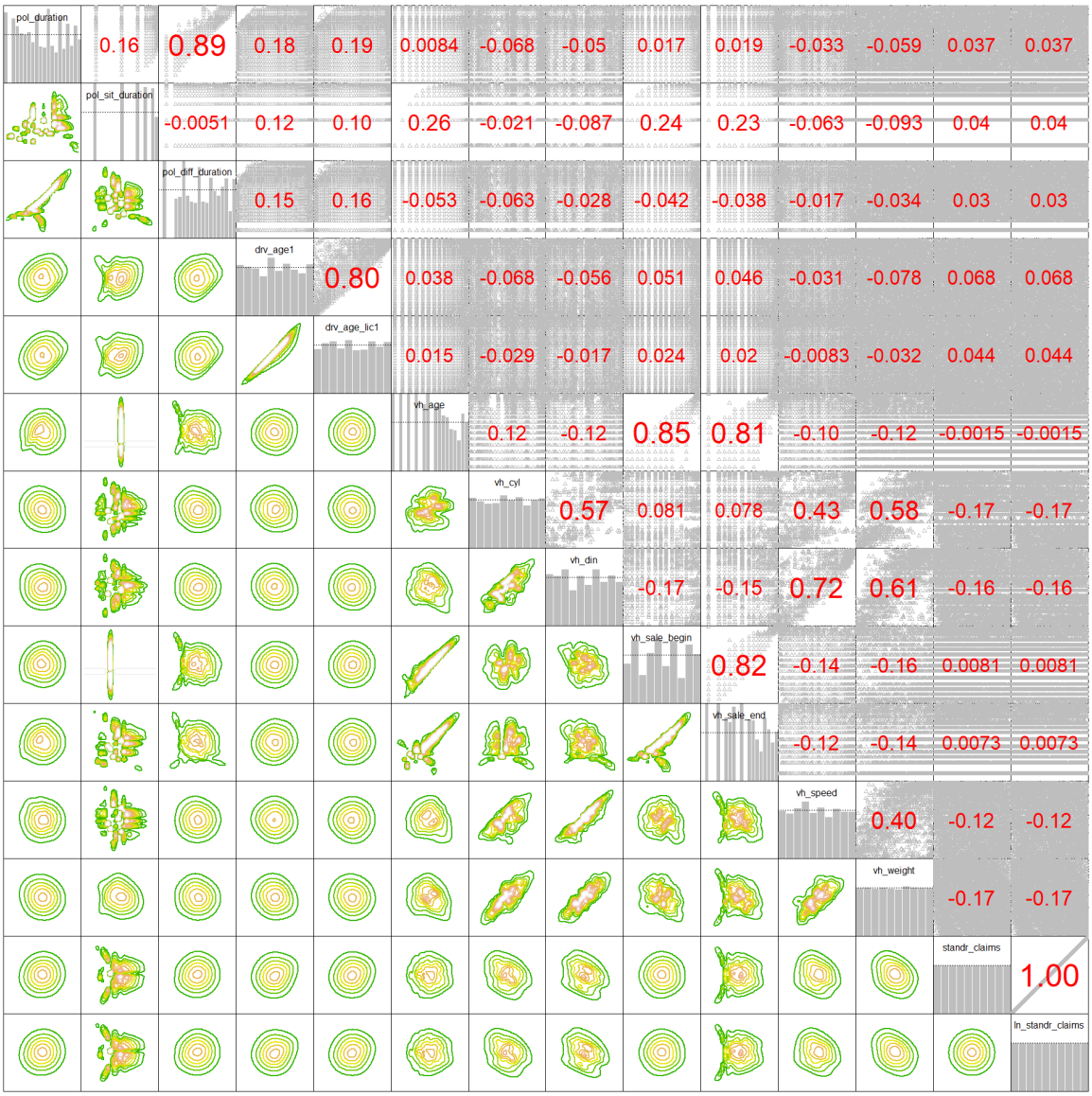

Figure 14: Good Driver training data set. *Lower:* empirical normalized contour plots for the pair copulas, *diagonal:* histogram of the margins, *upper:* pairs plots of copula data and their Kendall's taus.

We can see that the empirical copula data histograms do not look uniform for `pol_duration`, `pol_sit_duration`, `vh_age`, `vh_sale_begin` and `vh_sale_end`. Additionally, the pair copulas where one of the margins is `pol_sit_duration` look strange, which is another indication that we may need to transform this covariate to ordinal. Eventually we expect that some

of the pair copulas that have high Kendall's tau will be included in our models. Once we have our D-vine regression models, we can compare the fitted normalized contour plots with the empirical ones, However, before we transform any of the marginals to ordinal variables and fit the models, we estimate the marginals nonparametrically, using the R function `kde1d`, which can be found in the R package with the same name (Nagler and Vatter (2022)). The estimated densities of the marginals and their respective histograms are presented in Figure 15.

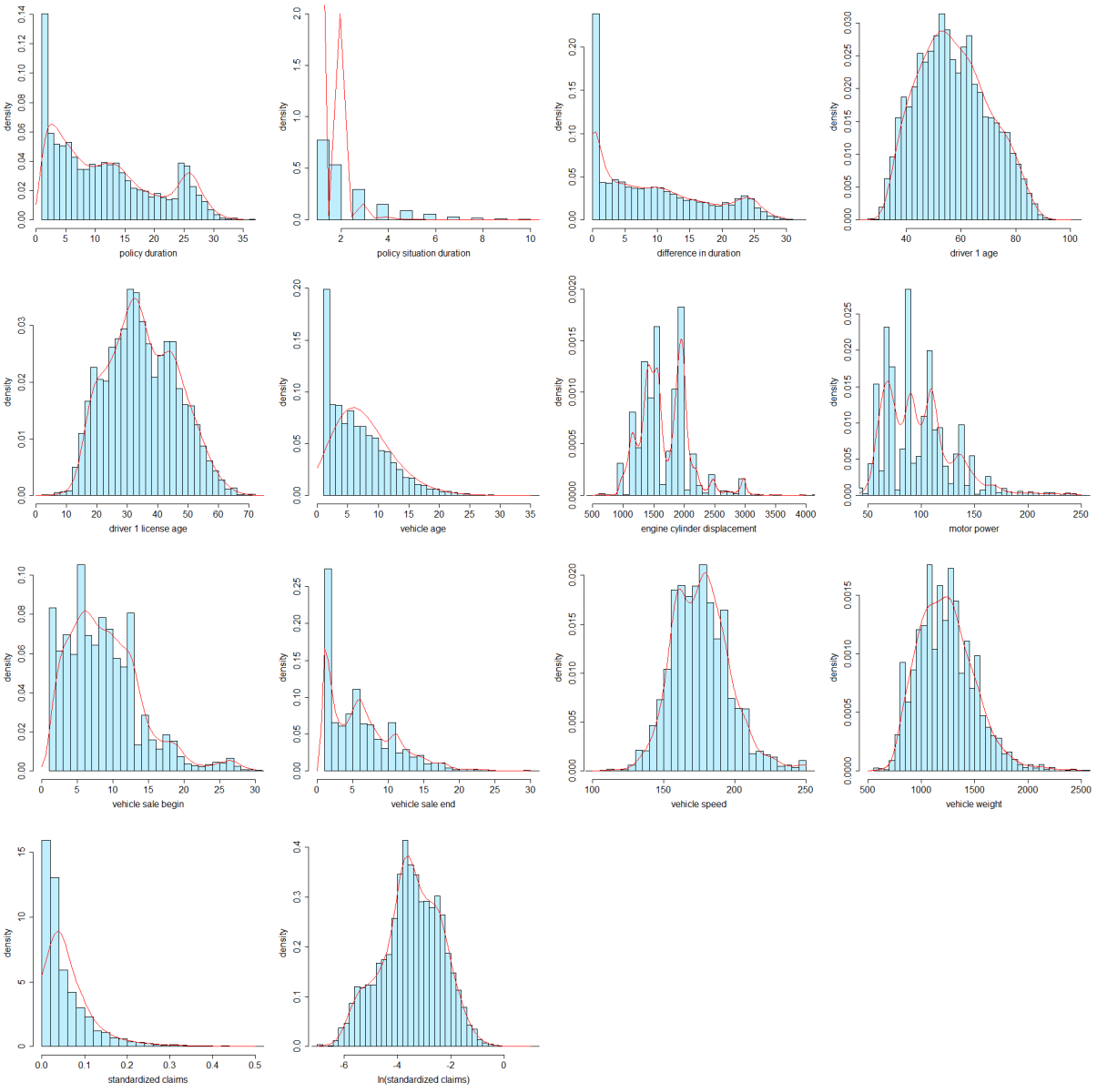

Figure 15: Histograms of the continuous marginals in Good Driver training data set. The red line denotes their `kde1d` estimators.

Using these `kde1d` estimators, we transform the data to copula data and we present the histograms of the marginals in Figure 16. The histograms of *policy situation duration* and *vehicle sale end* look particularly nonuniform, which is why we transform these covariates to ordinal: `sitdur_ordinal` and `saleend_ordinal`. The have respectively 20 and 33 levels.

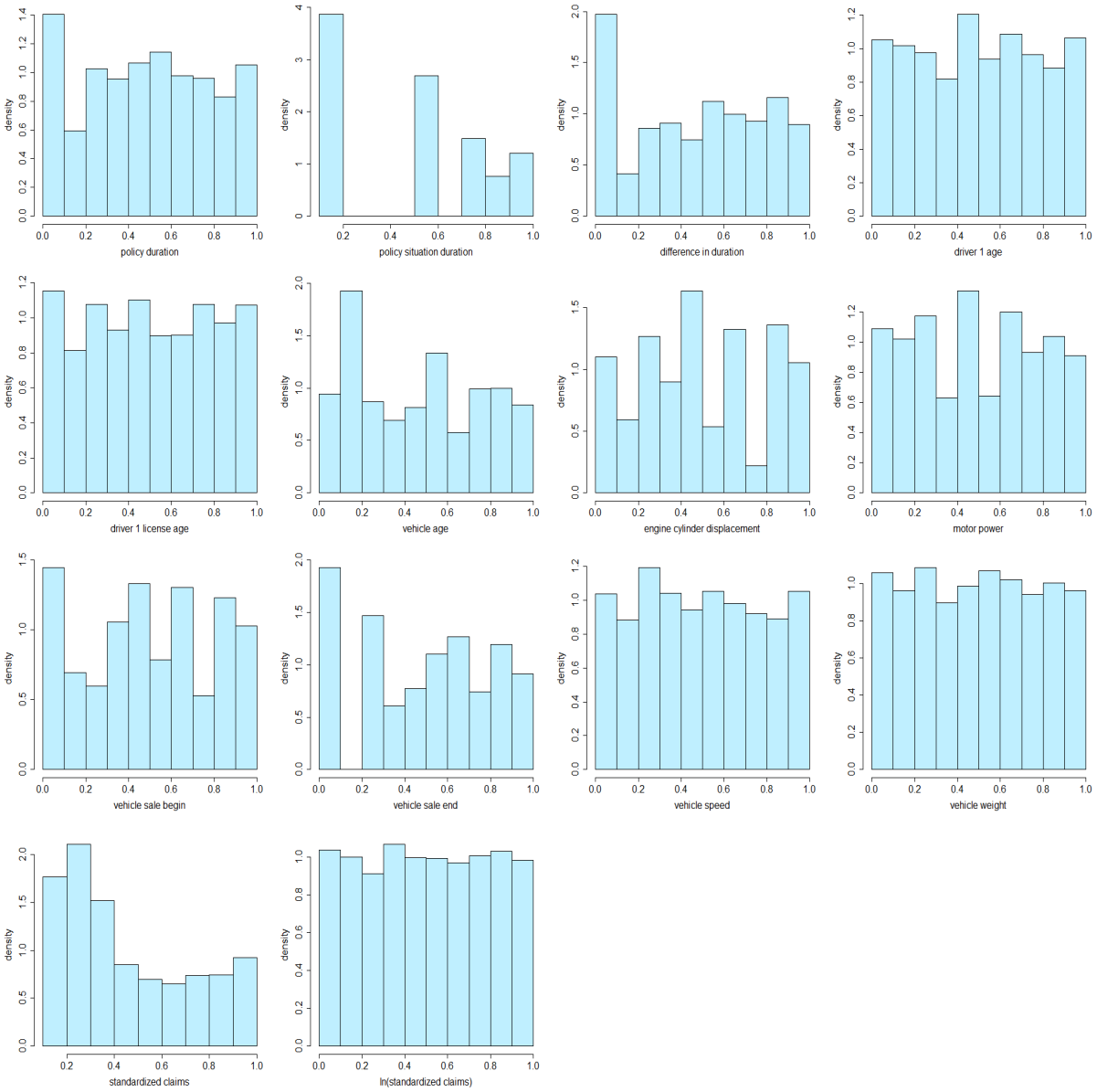

Figure 16: Histograms of the continuous marginals of the copula scale data in Good Driver training data set.

The histograms of the copula data marginals *difference in duration*, *vehicle age*, *engine cylinder displacement*, *motor power*, *vehicle sale begin* and *standardized claims* also do not look so uniform, however, we continue to work with them in our models.
An important preliminary step is transforming all of the discrete covariates in the Good Driver training data set to ordinal covariates. Their levels were sorted in increasing order based on their influence on the response. After we analysed our data and did the

necessary transformations of some of the covariates to ordinal type, we can proceed with modelling.
We present the model `DVReg1_nonpar_orig` which models the response variable on original scale *standardized claims*, and the model `DVReg1_nonpar_ln` which predicts the transformed response variable *ln(standardized claims)*. The advantage of using the transformed response as our response variable is that it guarantees positive value of the variable *standardized claims*, which we can obtain using the exponential function. The order of the variables in these models is presented in Table 25 and the fitted normalized contour plots in Figure 17 and Figure 18, respectively.

Order	**DVReg1_nonpar_orig**	**DVReg1_nonpar_ln**
1	standr_claims	ln_standr_claims
2	vh_cyl	vh_weight
3	pol_payd	vh_cyl
4	pol_usage	drv_age1
5	vh_fuel	vh_sale_begin
6	claim_ind	claim_ind
7	pol_coverage	vh_din
8	class_make2	pol_duration
9	saleend_ordinal	vh_fuel
10	vh_din	pol_coverage
11	drv_age1	saleend_ordinal
12	drv_drv2	
13	gender	
14	vh_age	
15	vh_speed	
16	drv_age_lic1	
17	pol_diff_duration	
18	pol_pay_freq	
19	sitdur_ordinal	
20	vh_sale_begin	
21	pol_duration	

Table 25: Order of variables in the D-vine regression models `DVReg1_nonpar_orig` and `DVReg1_nonpar_ln`.

From Table 25 we can see that the model `DVReg1_nonpar_orig` contains much more covariates than the model with transformed response, `DVReg1_nonpar_ln`. Additionally, in the model `DVReg1_nonpar_orig` almost all of the ordinal covariates are present (except for `vh_type`) and they are influential, since they are included early in the model. On the other hand, in `DVReg1_nonpar_ln` the highly influential covariates are of continuous type and most of the convariates in the model are continuous. All of the covariates that are present in the model `DVReg1_nonpar_ln` are also present in `DVReg1_nonpar_orig`, except for the most influential covariate, `vh_weight`.
In Figure 17 we can see that few of the fitted pair copulas in `DVReg1_nonpar_orig` indicate high Kendall's tau values. There are pair copulas that look particularly nonparametric, for example the pair copula with marginals (`drv_drv2`, `gender`). We are interested in comparing the fitted pair copulas in the first tree of the D-vine regression model `DVReg1_nonpar_orig` with continuous marginals, which are presented in the last row of Figure 17, with their respective empirical normalized contour plots in Figure 14. The normalized contour plots of the pair copulas with marginals (`saleend_ordinal`, `vh_din`) and (`sitdur_ordinal`, `vh_sale_begin`) look particularly different from their respective empirical normalized contour plots. In these cases, the nonparametric fitted copulas do not

catch all of the patterns in the data. Additionally, the normalized contour plots of the pair copulas with marginals (`standr_claims`, `vh_cyl`) and (`vh_age`, `vh_speed`) look different for higher contour levels than their empirical plots. All other fitted pair copulas in the first tree with continuous marginals look similar to their respective empirical pair copulas in Figure 14.

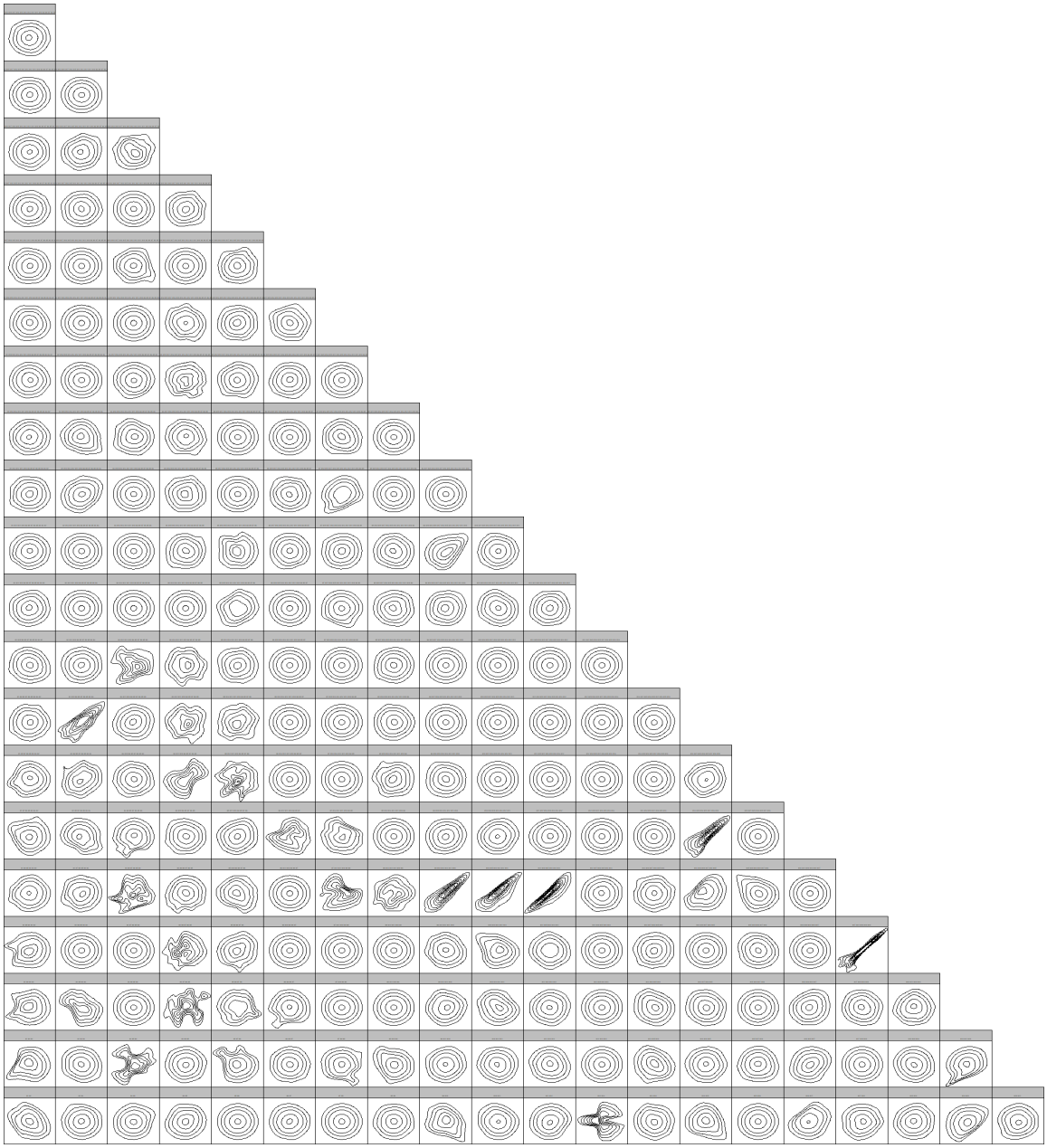

Figure 17: Normalized fitted contour plots for the nonparametric pair copulas of the D-vine regression model `DVReg1_nonpar_orig`, where the variables of the model are presented by X1=standr_claims, X2=vh_cyl, X3=pol_payd, X4=pol_usage, X5=vh_fuel, X6=claim_ind, X7=pol_coverage, X8=class_make2, X9=saleend_ordinal, X10=vh_din, X11=drv_age1, X12=drv_drv2, X13=gender, X14=vh_age, X15=vh_speed, X16=drv_age_lic1, X17=pol_diff_duration, X18=pol_pay_freq, X19=sitdur_ordinal, X20=vh_sale_begin and X21=pol_duration as their order in Table 25.

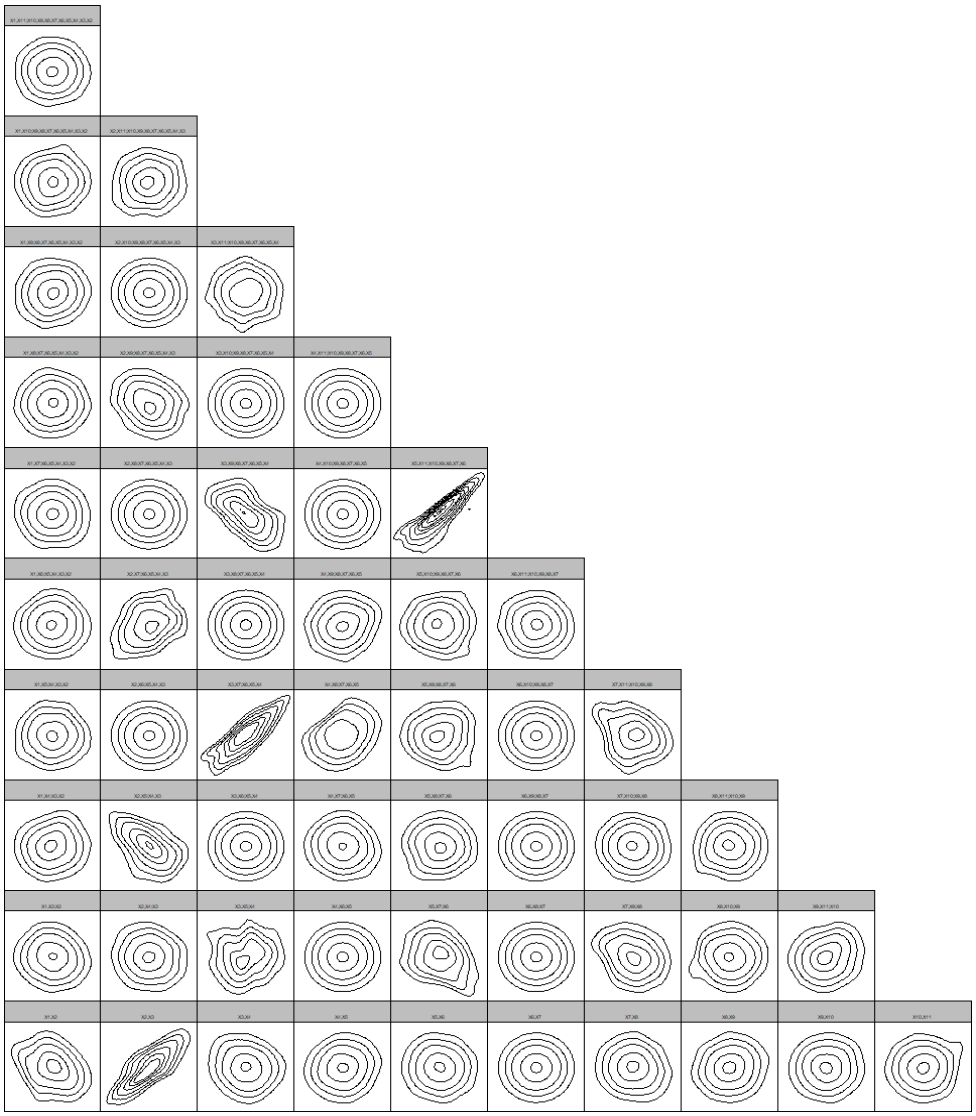

Figure 18: Normalized fitted contour plots for the nonparametric pair copulas of the D-vine regression model `DVReg1_nonpar_ln`, where the variables of the model are presented by X1=ln_standr_claims, X2=vh_weight, X3=vh_cyl, X4=drv_age1, X5=vh_sale_begin, X6=claim_ind, X7=vh_din, X8=pol_duration, X9=vh_fuel, X10=pol_coverage and X11=saleend_ordinal, as their order in Table 25.

Compared to `DVReg1_nonpar_orig`, the model `DVReg1_nonpar_ln` is less complex. In Figure 14 we can see that the empirical normalized contour plots of the pair copulas of *standardized claims* and the covariates look identical with the empirical normalized contour plots of the pair copulas of *ln(standardized claims)* and the covariates. Surprisingly however, we obtain much different D-vine regression model `DVReg1_nonpar_ln` using the transformed response variable than the model with the response on original scale `DVReg1_nonpar_orig`. Again, we are interested in the fitted normalized contour plots of the pair copulas in the first tree of the model `DVReg1_nonpar_ln` with continuous marginals, which are visualised in the last row of Figure 18, as we want to compare them with their respective empirical normalized contour plots in Figure 14. In this model, all of the normalized contour plots of the fitted nonaprametric pair copulas look similar to the normalized contour plots of

their respective empirical pair copulas. Additionally, we can see that early in the model two highly dependent variables were included in the model, `vh_weight` and `vh_cyl`. In the next section we introduce the linear quantile regression models, which contain the same variables as the D-vine regression models, and we compare all four regression approaches on the Good Driver Data.

4.4 Comparison of the models

Before we proceed with comparison of the regression models, we fit linear quantile regression models `LQReg1_orig` and `LQReg1_ln` using the same variables which are present in the models `DVReg1_nonpar_orig` and `DVReg1_nonpar_ln` respectively, such that instead of the ordinal covariates `sitdur_ordinal` and `saleend_ordinal` we use the continuous covariates `pol_sit_duration` and `vh_sale_end`. As a result, the model `LQReg1_orig` uses the response on the original scale *standardized claims*, whereas the model `LQReg1_ln` uses the response on ln scale, *ln(standardized claims)*. However, in the model `LQReg1_orig`, we discard the covariates `gender` and `pol_duration`, which are present in `DVReg1_nonpar_orig`, in order to avoid singular design matrix. The decision was based on fitting a linear regression model using the variables from the model `DVReg1_nonpar_orig`, which in its summary presented `NA` outputs for these variables.
The main obstacle in the comparison of all resulting models is that lognormal and gamma regression models predict the conditional mean of the response, whereas quantile regression predicts the conditional meadian of the response for a quantile level 0.5. We compare the models based on log likelihood, AIC ,BIC, training error, test error and interval score values on the scale of *standardized claims*. The transformation of these measures on original scale was studied in Section 2.8. The models with larger log likelihood and smaller AIC, BIC, training error, test error and interval score values are considered to be better. The log likelihood, AIC, BIC and the training error are calculated on the same data set we use to fit the models, which is the Good Driver training data set. The test error and the interval score are calculated on a new data set, which is the test data set of the Good Driver Data. To make these measures comparable for all models, we calculate them on the *standardized claims* scale, using the exponential function for the models with *ln(standardized claims)* as a response variable. Models with smaller training and test error are considered to be better, since that indicates more parsimonious fit of the model. The training errors by themselves are not as informative, because smaller training error of a model also indicates overfitting. The test errors are of special interest, because they indicate the performance of the model on a new data set.
Additionally, we calculate the interval score values for 95% prediction interval and for 90% prediction interval. The results are presented in Table 26, where the best values per performance measure are identified in gray.

	df	LogLik	AIC	BIC	Training error	Test error	Interval score $\alpha = 0.05$	Interval score $\alpha = 0.1$
LogLM1_main_red	13	14203.96	-28381.91	**-28292.42**	0.054444	0.004635	0.00834	0.01302
LogLM1_inter_red	14	14206.71	**-28385.41**	-28289.04	0.054448	0.004640	0.00826	0.01294
GamReg1_main_red	8	13556.03	-27096.05	-27040.98	0.053512	0.004288	0.00804	0.01300
GamReg1_inter_red	10	13566.68	-27113.37	-27044.53	**0.053445**	0.004294	0.00805	0.01296
DVReg1_nonpar_orig	2486	13609.09	-22246.16	-5133.01	0.053883	**0.004270**	0.00890	0.01312
DVReg1_nonpar_ln	941	**14793.27**	-27702.92	-21219.72	0.054206	0.004468	0.00811	0.01245
LQReg1_orig	24	10575.79	-21103.58	-20938.37	0.054400	0.004582	**0.00744**	0.01217
LQReg1_ln	13	13907.66	-27789.31	-27699.82	0.054361	0.004534	0.00752	**0.01215**

Table 26: Comparison of different regression models on Good Driver Data based on log likelihood, AIC, BIC, training error, test error and interval score values on the original *standardized claims* scale. The best values per column are identified in gray.

For the nonparametric D-vine regression models, as a penalty for complexity in computing AIC and BIC criterion we take the degrees of freedom of all the pair copulas in the model. From Table 26, we can see that the complexity of the D-vine regression models is much bigger than the other regression models, based on the degrees of freedom. However, the model `DVReg1_nonpar_orig` has significantly more degrees of freedom than `DVReg1_nonpar_ln`, while all perfomance measure values of `DVReg1_nonpar_orig` are worse except for the training and test errors.
Based on the log likelihood, the D-vine regression model with transformed response variable *ln(standardized claims)* shows best performance. Based on the AIC and BIC criterion however, the lognormal models show the best performance. According to training and test error values, the gamma regression models `GamReg1_main_red` and `GamReg1_inter_red` and the D-vine regression model `DVReg1_nonpar_orig` have the lowest errors on the training and test data set of the Good Driver Data. The performance of the models on the test data set is particularly important, because that way we can measure the predictive accuracy of the model on a new data set. For that purpose we can look at the interval score values, where the linear quantile regression models have the lowest values.
What we can also notice from the performance measures is that the D-vine regression model and the linear quantile regression model with the transformed response have better values for most performance measures than the models with the original response variable. The linear quantile regression models have better interval score values than their respective D-vine regression models, however their log likelihood, training and test error values are worse than the D-vine regression models. In particular, even though the number of degrees of freedom of `DVReg1_nonpar_orig` is much larger than the number of degrees of freedom of `LQReg1_nonpar_ln`, the AIC value of `DVReg1_nonpar_orig` is better than the AIC value of `LQReg1_nonpar_orig`.
The gamma regression models have worse log likelihood, AIC and BIC values than lognormal regression models, but their training and test error values are better. Additionally, the interaction terms in both gamma and lognormal regression models do not improve the main effects models significantly.
Finally, we take the models `LogLM1_inter_red`, `GamReg1_inter_red`, `DVReg1_nonpar_ln` and `LQReg1_ln` as most parsimonious per regression method. First, we present the covariates of these four models in Table 28.

LogLM1_inter_red	GamReg1_inter_red	DVReg1_nonpar_ln
poly(drv_age1,2)	drv_age1	vh_weight
log(vh_cyl)	poly(drv_age2,2)	vh_cyl
vh_speed	log(vh_cyl)	drv_age1
poly(vh_weight, 2)	poly(vh_weight, 2)	vh_sale_begin
pol_coverage2	drv_age1:poly(drv_age2, 2)	claim_ind
vh_fuel		vh_din
claim_ind		pol_duration
class_make2		vh_fuel
vh_speed:poly(vh_weight,2)		pol_coverage
		saleend_ordinal

Table 27: Covariates included in each of the models `LogLM1_inter_red`, `GamReg1_inter_red` and `DVReg1_nonpar_ln`. The model `LQReg1_ln` has the same covariates as `DVReg1_nonpar_ln`, but instead of `saleend_ordinal` we use `vh_sale_end`.

Only three covariates are present in all four models, `drv_age1`, `vh_cyl` and `vh_weight`.

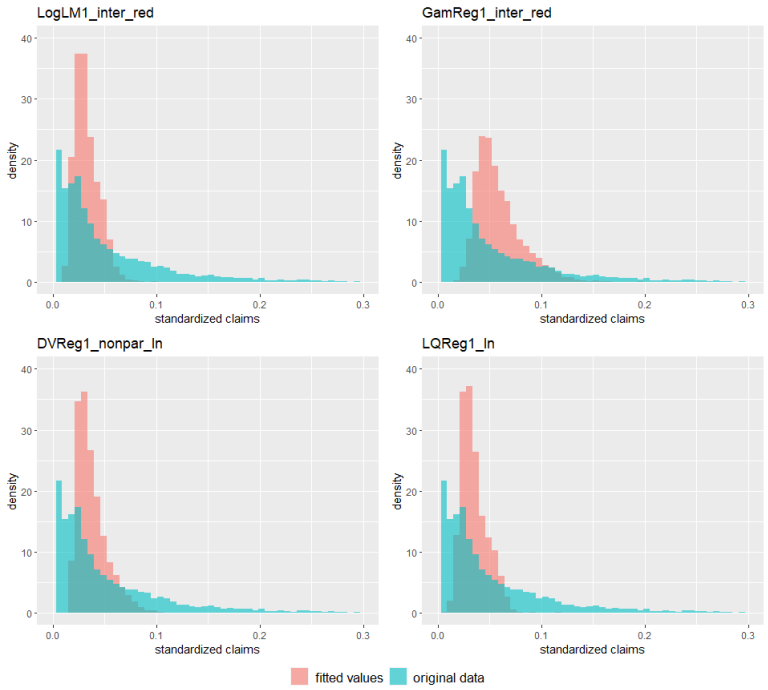

Figure 19: Histograms of the fitted values of *standardized claims* of the models `LogLM1_inter_red`, `GamReg1_inter_red`, `DVReg1_nonpar_ln` and `LQReg1_ln` on the Good Driver training data set.

In Figure 19 we present the histograms of fitted values on original scale of the models `LogLM1_inter_red`, `GamReg1_inter_red`, `DVReg1_nonpar_ln` and `LQReg1_ln` on the Good Driver training data set and the histogram of the original data of the variable *standardized claims*. By fitted values of the models `DVReg1_nonpar_ln` and `LQReg1_ln`, we mean the 0.5 conditional quantiles of the response on the original *standardized claims* scale. All of the histograms indicate that the variables included in the models can not predict the response *standardized claims* well. The histograms of `LogLM1_inter_red`, `DVReg1_nonpar_ln` and `LQReg1_ln` are similar and these models predict well the small values of *standardized*

claims, however, they underestimate the tail. The model `GamReg1_inter_red` predicts better the larger values of the response, but it overestimates the small values. The range of the fitted values of the model `GamReg1_inter_red` is larger, which is not the case with the other three models who have much smaller range of the fitted values.
We are more interested in the histograms of the predicted values of *standardized claims* on the Good Driver test data set, which are presented in Figure 20. The histograms look similar to the histograms on the training data set. In this case however, the D-vine regression model `DVReg1_nonpar_ln` predicts the tail better than the models `LogLM1_inter_red` and `LQReg1_ln`. The model `GamReg1_inter_red` predicts well the big values of *standardized claims*, but overestimates the small values.

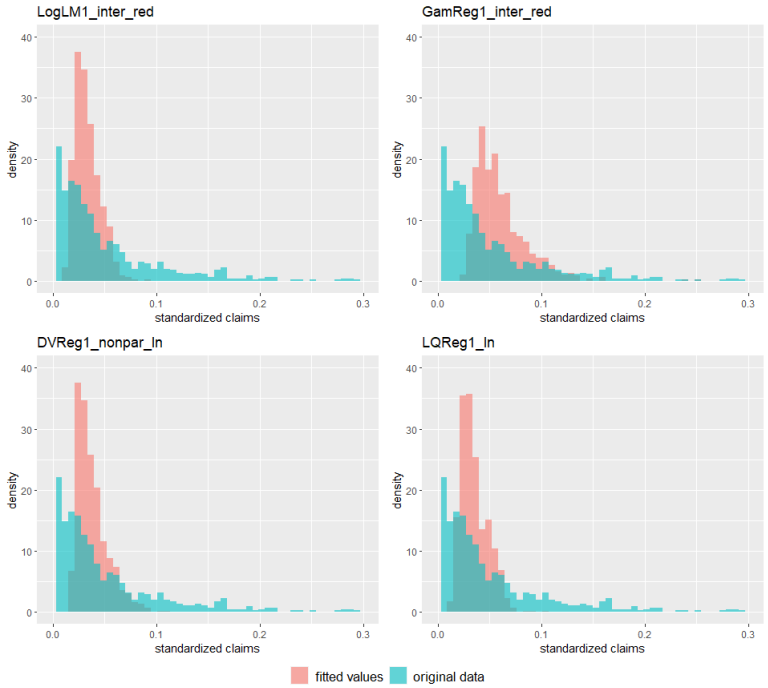

Figure 20: Histograms of the predicted values of *standardized claims* of the models `LogLM1_inter_red`, `GamReg1_inter_red`, `DVReg1_nonpar_ln` and `LQReg1_ln` on the Good Driver test data set.

Additionally, we present the 90% prediction intervals of the models in Figure 21. We can see that most of the original values of *standardized claims* belong to the prediction intervals of the models, except the big values of *standardized claims* at the right end of the plots. The model `LQReg1_inter_red` has most narrow prediction interval, although, this model has lowest interval score for the 90% prediction interval. Overall, the prediction intervals of the four regression models look similarly.

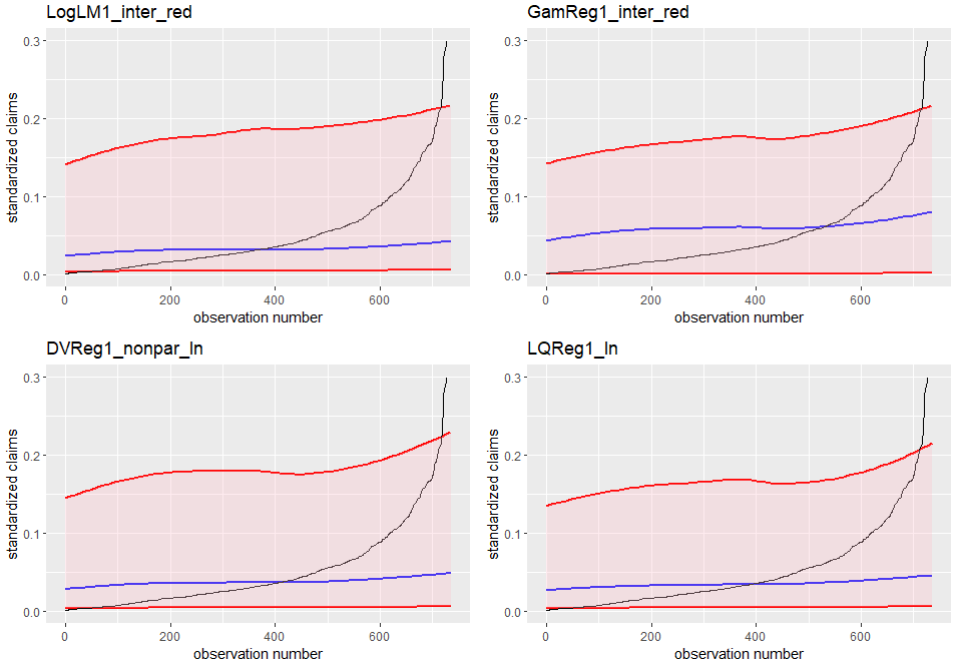

Figure 21: 90% prediction intervals of *standardized claims* on the Good Driver test data set for the models `LogLM1_inter_red`, `GamReg1_inter_red`, `DVReg1_nonpar_ln` and `LQReg1_ln`. The black line denotes the original values of *standardized claims*, the blue line denotes the smoothed line of the predicted values and the red area denotes the prediction interval, where the red lines are the smoothed limits of the prediction interval. The observations are sorted in increasing order based on the original values of *standardized claims*.

In Figure 22 we present the plots of the predicted values of the models on original scale against the values of *standardized claims*, which we plot using the Good Driver test data set. Again, as predicted values of the quantile regression models we use the 0.5 conditional quantiles. If the fit of the model is good, we expect the values to be randomly scattered around the 45° line. As shown in Figure 22, that is hardly the case and again the models `LogLM1_main_red`, `DVReg1_nonpar_ln` and `LQReg1_ln` underestimate the big values of the response variable, whereas the model `GamReg1_inter_red` better estimates the bigger values of the response, but underestimates the small values.

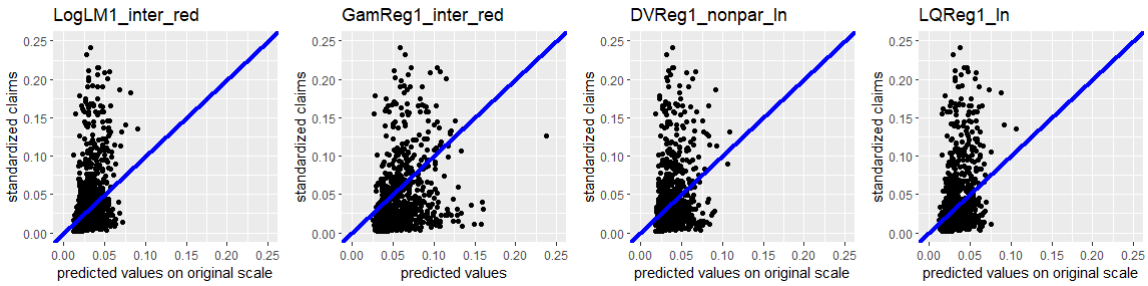

Figure 22: Plots of the fitted values on original scale against the original values of *standardized claims* on the Good Driver test data set for the models `LogLM1_inter_red`, `GamReg1_inter_red`, `DVReg1_nonpar_ln` and `LQReg1_ln`. The blue line denotes the 45° line.

Finally, since the models `DVReg1_nonpar_ln` and `LQReg1_ln` have similar performance measures values and show similar behaviour, we would like to investigate the marginal effects of the continuous covariates on the predicted quantiles of the models. The marginal effect of a continuous covariate is presented by a plot of the continuous covariate x_i against the fitted conditional quantiles $\hat{q}_\alpha^i, i = 1, \ldots, n_{tr}$, where all other covariates are set to their observed value. The marginal effect plots of the models for three different quantile levels 0.1, 0.5 and 0.9. are presented in Figure 23.

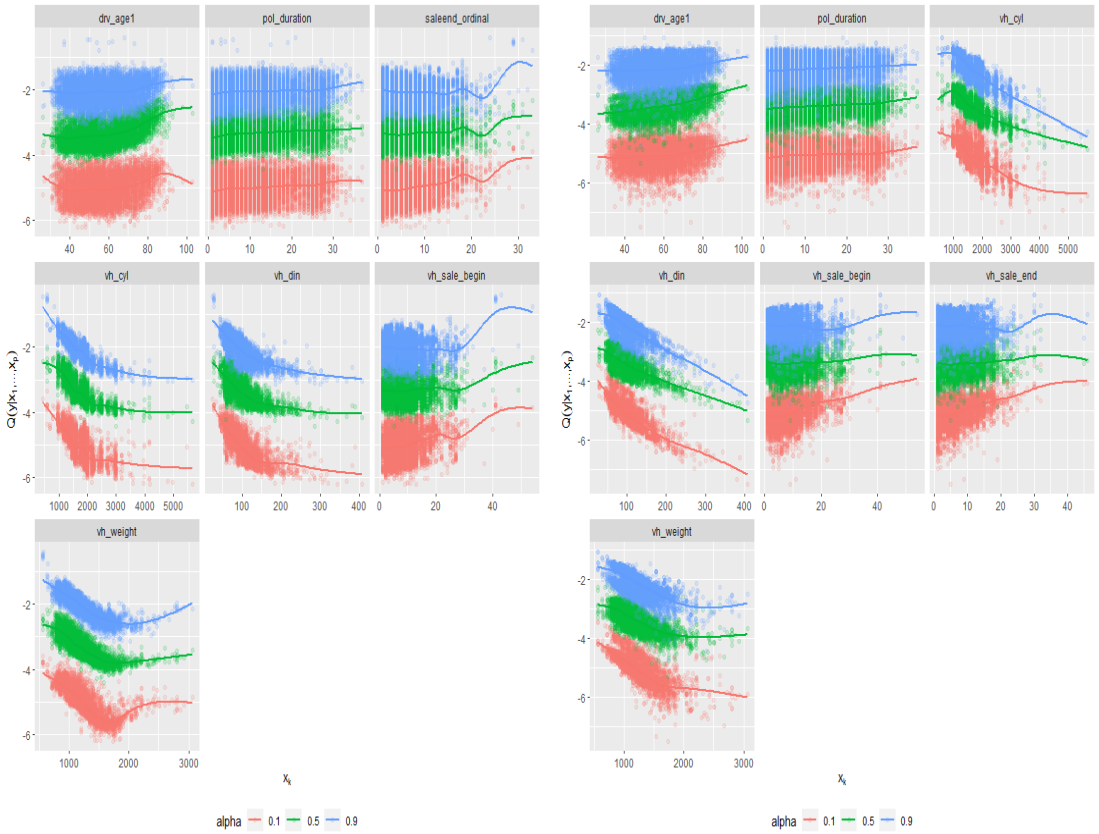

Figure 23: Marginal effect plots for the continuous covariates on the predicted quantiles on Good Driver training data set for three different quantile levels. *Left*:`DVReg1_nonpar_ln`, *right*:`LQReg1_ln`.

The marginal effect plots for the two quantile regression models look similar where there is sufficient data. For the both models we can spot nonlinearity in the plots for the covariates `vh_sale_end` (respectively `saleend_ordinal`), `vh_cyl`, `vh_sale_begin` and `vh_weight`. However, the marginal effect plots of the model `DVReg1_nonpar_ln` indicate stronger nonlinearity than the marginal effect plots of the model `DVReg1_nonpar_ln`.
Considering the advantages of using D-vine quantile regression over lognormal, gamma and linear quantile regression, some of which are model flexibility, avoiding issues like multicollinearity, need for transformations and interactions, we can say that on the Good Driver Data set, this D-vine quantile regression model seems to be beneficial. One of the disadvantages of the model is its complexity, which is large due its nonparametric marginal and pair copula esimates. For Bad Driver Data Set however, we will also fit a

parametric D-vine regression model, therefore we will obtain a model whose complexity is more easily comparable with lognormal, gamma and linear regression models.

5 Modelling on Bad Driver Data

After fitting and comparing different regression models on Good Driver Data, we repeat this procedure on Bad Driver Data, where the observations do not belong to the best bonus malus class 0.5. Again, in the lognormal and gamma regression models we allow for interactions, and we fit nonparametric D-vine regression models using both the response variable *standardized claims* and the transformed response *ln(standardized claims)* as a dependent variable. In this case however, the lognormal models on Bad Driver Data have higher R^2_{adj} than the lognormal models fitted on Good Driver Data, therefore we also fit parametric D-vine regression models using finite mixture of skew-normal and normal distributions as marginal estimates and parametric pair copulas. Additinally, we fit two linear quantile regression models, one predicting the original response and one predicting the transformed response, which contain the same variables as the best D-vine regression models with these respective response variables. Finally, we compare the fitted models using the log likelihood, AIC, BIC, train error, test error and interval score as performance measures.

$R^2 = 12.15\%$, $R^2_{adj} = 11.12\%$	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-3.34	0.23	-14.52	0.00
pol_bonus	0.23	0.18	1.31	0.19
pol_duration	0.00	0.00	0.34	0.73
pol_sit_duration	0.02	0.02	1.04	0.30
poly(drv_age2, 3)1	-1.09	1.18	-0.92	0.36
poly(drv_age2, 3)2	-0.11	1.18	-0.09	0.93
poly(drv_age2, 3)3	0.33	1.09	0.30	0.76
poly(drv_age_lic1, 3)1	5.12	1.65	3.10	0.00
poly(drv_age_lic1, 3)2	1.38	1.30	1.06	0.29
poly(drv_age_lic1, 3)3	-3.92	1.14	-3.43	0.00
poly(vh_cyl, 2)1	-6.61	2.14	-3.08	0.00
poly(vh_cyl, 2)2	3.44	1.93	1.78	0.08
poly(vh_din, 3)1	-1.02	4.04	-0.25	0.80
poly(vh_din, 3)2	-2.68	1.99	-1.35	0.18
poly(vh_din, 3)3	-2.82	1.65	-1.71	0.09
poly(vh_speed, 2)1	-0.53	2.97	-0.18	0.86
poly(vh_speed, 2)2	-1.60	1.70	-0.94	0.35
poly(vh_weight, 5)1	-4.89	2.39	-2.04	0.04
poly(vh_weight, 5)2	1.26	1.46	0.86	0.39
poly(vh_weight, 5)3	0.99	1.24	0.79	0.43
poly(vh_weight, 5)4	-3.38	1.19	-2.84	0.00
poly(vh_weight, 5)5	1.74	1.11	1.56	0.12
pol_coverage2NoMini	-0.42	0.14	-3.05	0.00
pol_usage2NoRetired	0.02	0.10	0.17	0.86
vh_fuelGasoline	0.17	0.07	2.48	0.01
claim_indOne	-0.13	0.07	-1.77	0.08
class_make2Class2&3	0.14	0.06	2.25	0.02
gender2AllOther	0.03	0.05	0.51	0.61

Table 28: Maximum likelihood estimates, estimated standard errors, t-values and corresponding p-values, as well as R^2 and R^2_{adj} for model `LogLM2_main`.

5.1 Lognormal regression model

In Section 3.3 we showed that the Kendall's taus of `drv_age1` and `drv_age_lic1`, and of `drv_age2` and `drv_age_lic2` are close to 1, so in order to avoid multicollinearity, we include only `drv_age2` and `drv_age_lic1` in our models, since the combination of these covariates in the model showed most parsimonious fit. The main effects lognormal model which includes the covariates presented in Section 3.3 is:

```
LogLM2_main<-lm(ln_standr_claims~pol_bonus+pol_duration+pol_sit_duration+
                poly(drv_age2,3)+poly(drv_age_lic1,3)+poly(vh_cyl,2)+
                poly(vh_din,3)+poly(vh_speed,2)+poly(vh_weight,5)+
                pol_coverage2+pol_usage2+vh_fuel+claim_ind+
                class_make2+gender2, data=data2)
```

and its summary is presented in Table 28. We can notice that the R^2_{adj} is small, only 11.12%, and some of the covariates are insignificant, especially the higher terms of some of the polynomials. Since backward elimination of covariates using the BIC criterion eliminates a lot of significant covariates, we remove the covariates from the model manually. Namely, we remove `pol_duration`, `pol_sit_duration`, `poly(drv_age2,3)`, `poly(vh_speed,2)`, `pol_usage2`, `gender2` and we lower the degree of polynomial of `vh_din`. The resulting model is:

```
LogLM2_main_red<-lm(ln_standr_claims~pol_bonus+poly(drv_age_lic1,3)+
              poly(vh_cyl,2)+poly(vh_din,2)+poly(vh_weight,5)+
              pol_coverage2+vh_fuel+claim_ind+class_make2, data=data2)
```

$R^2 = 11.94\%$, $R^2_{adj} = 11.30\%$	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-3.28	0.21	-15.70	0.00
pol_bonus	0.25	0.18	1.43	0.15
poly(drv_age_lic1, 3)1	5.28	1.13	4.67	0.00
poly(drv_age_lic1, 3)2	1.24	1.15	1.08	0.28
poly(drv_age_lic1, 3)3	-3.80	1.12	-3.39	0.00
poly(vh_cyl, 2)1	-6.06	2.10	-2.89	0.00
poly(vh_cyl, 2)2	3.75	1.89	1.99	0.05
poly(vh_din, 2)1	-1.62	2.13	-0.76	0.45
poly(vh_din, 2)2	-3.82	1.71	-2.23	0.03
poly(vh_weight, 5)1	-5.72	2.06	-2.77	0.01
poly(vh_weight, 5)2	2.06	1.36	1.52	0.13
poly(vh_weight, 5)3	0.36	1.12	0.32	0.75
poly(vh_weight, 5)4	-3.63	1.11	-3.26	0.00
poly(vh_weight, 5)5	2.02	1.09	1.86	0.06
pol_coverage2NoMini	-0.42	0.14	-3.04	0.00
vh_fuelGasoline	0.16	0.07	2.45	0.01
claim_indOne	-0.13	0.07	-1.79	0.07
class_make2Class2&3	0.14	0.06	2.28	0.02

Table 29: Maximum likelihood estimates, estimated standard errors, t-values and corresponding p-values, as well as R^2 and R^2_{adj} for model `LogLM2_main_red`.

and its output is presented in Table 29. In the model `LogLM2_main_red` all of the covariates are significant and on 0.1 level the R^2_{adj} is slightly improved, except for the covariate `pol_bonus`, whose p-value is 0.15 so it is debatable whether we should keep it in the model, however, we decide to keep it because this variable appears in a lot of interaction terms presented in Section 3.3. If we perform an F-test for the removed covariates, we obtain a p-value of 0.85, which indicates that none of the removed terms is statistically significant. The R^2_{adj} of the reduced model `LogLM2_main_red` is still small, so in the next step we will try to improve it by allowing for interaction effects in the model. To the model `LogLM2_main_red` we add all of the possible interactions between the covariates, which defines the model `LogLM2_inter`, whose output is presented in Table 30.

$R^2 = 12.75\%$, $R^2_{adj} = 11.24\%$	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-3.04	1.01	-3.01	0.00
pol_bonus	1.32	1.71	0.77	0.44
poly(drv_age_lic1, 3)1	2.76	13.94	0.20	0.84
poly(drv_age_lic1, 3)2	1.62	1.27	1.28	0.20
poly(drv_age_lic1, 3)3	-4.24	1.23	-3.43	0.00
poly(vh_cyl, 2)1	5.65	14.40	0.39	0.69
poly(vh_cyl, 2)2	2.84	4.58	0.62	0.54
poly(vh_din, 2)1	-15.97	18.53	-0.86	0.39
poly(vh_din, 2)2	20.22	14.53	1.39	0.16
poly(vh_weight, 5)1	-1.13	16.64	-0.07	0.95
poly(vh_weight, 5)2	-6.86	11.05	-0.62	0.53
poly(vh_weight, 5)3	0.19	4.71	0.04	0.97
poly(vh_weight, 5)4	-10.17	3.20	-3.18	0.00
poly(vh_weight, 5)5	1.16	2.78	0.42	0.68
pol_coverage2NoMini	-0.84	0.81	-1.04	0.30
vh_fuelGasoline	0.22	0.20	1.12	0.26
claim_indOne	-0.13	0.07	-1.84	0.07
class_make2Class2&3	0.57	0.28	2.03	0.04
pol_bonus:drv_age_lic1	0.01	0.01	0.47	0.64
pol_bonus:vh_cyl	-0.00	0.00	-0.87	0.38
pol_bonus:poly(vh_din, 2)1	27.80	15.65	1.78	0.08
pol_bonus:poly(vh_din, 2)2	-20.66	18.21	-1.13	0.26
pol_bonus:poly(vh_weight, 2)1	-15.39	13.61	-1.13	0.26
pol_bonus:poly(vh_weight, 2)2	10.28	9.99	1.03	0.30
drv_age_lic1:vh_cyl	-0.00	0.00	-0.50	0.62
poly(vh_din, 2)1:drv_age_lic1	0.02	0.17	0.11	0.91
poly(vh_din, 2)2:drv_age_lic1	-0.27	0.20	-1.34	0.18
poly(vh_weight, 5)1:drv_age_lic1	-0.07	0.17	-0.43	0.67
poly(vh_weight, 5)2:drv_age_lic1	0.26	0.18	1.45	0.15
poly(vh_weight, 5)3:drv_age_lic1	0.08	0.20	0.38	0.71
poly(vh_weight, 5)4:drv_age_lic1	0.27	0.15	1.78	0.07
poly(vh_weight, 5)5:drv_age_lic1	0.04	0.13	0.35	0.73
vh_cyl:vh_din	-0.00	0.00	-0.27	0.79
vh_cyl:poly(vh_weight, 2)1	0.01	0.01	0.72	0.47
vh_cyl:poly(vh_weight, 2)2	0.00	0.00	0.02	0.99
poly(vh_weight, 2)1:vh_din	-0.00	0.08	-0.04	0.97
poly(vh_weight, 2)2:vh_din	-0.05	0.06	-0.82	0.41
pol_bonus:pol_coverage2NoMini	0.39	0.93	0.42	0.68
pol_bonus:class_make2Class2&3	-0.63	0.40	-1.57	0.12
pol_coverage2NoMini:drv_age_lic1	0.01	0.01	0.49	0.63
vh_fuelGasoline:vh_din	-0.00	0.00	-0.32	0.75

Table 30: Maximum likelihood estimates, estimated standard errors, t-values and corresponding p-values, as well as R^2 and R^2_{adj} for model `LogLM2_inter`.

Compared to `LogLM2_main_red`, the model `LogLM2_inter` has smaller R^2_{adj} and rarely which of the parameters shows significance on a 0.1 level. If we perform an backward elimination of covariates on the model `LogLM2_inter` using the `step` function and BIC criterion, with the model `LogLM2_main_red` as a lower boundary model,

```
step(LogLM2_inter,direction="backward",data=data2,
     scope=list(upper= LogLM2_inter, lower= LogLM2_main_red),
     k=ln(nrow(data2)))
```

we obtain the main effect model `LogLM2_main_red`, which means that using this elimination criterion, all of the interaction terms were eliminated. If we perform an F-test for the interaction terms in the model `LogLM2_inter`, we obtain a p-value of 0.55, meaining that none of the interaction terms are significant. This result may be due to the same few covariates being present in the interaction terms in the model, which is why also the main effect terms lost their significance. That is why in the next step we try to reduce the model `LogLM2_inter` by manually removing the insignificant terms, in order to obtain a reduced model which also contains significant interaction terms. The resulting model is:

```
LogLM2_inter_red<-lm(ln_standr_claims ~ pol_bonus+poly(drv_age_lic1,3)+
       poly(vh_cyl,2)+poly(vh_din,2)+poly(vh_weight,5)+pol_coverage2+
       vh_fuel+claim_ind+class_make2+drv_age_lic1:poly(vh_weight,4)+
       pol_bonus:class_make2,data=data2)
```

$R^2 = 12.37\%$, $R^2_{adj} = 11.55\%$	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-3.61	0.30	-12.22	0.00
pol_bonus	0.76	0.36	2.14	0.03
poly(drv_age_lic1, 3)1	5.21	1.14	4.57	0.00
poly(drv_age_lic1, 3)2	1.48	1.16	1.28	0.20
poly(drv_age_lic1, 3)3	-3.90	1.14	-3.43	0.00
poly(vh_cyl, 2)1	-6.11	2.10	-2.91	0.00
poly(vh_cyl, 2)2	3.75	1.92	1.95	0.05
poly(vh_din, 2)1	-1.55	2.14	-0.73	0.47
poly(vh_din, 2)2	-3.52	1.71	-2.06	0.04
poly(vh_weight, 5)1	-4.07	3.01	-1.35	0.18
poly(vh_weight, 5)2	-1.26	3.19	-0.39	0.69
poly(vh_weight, 5)3	-1.91	3.00	-0.64	0.52
poly(vh_weight, 5)4	-8.86	2.40	-3.69	0.00
poly(vh_weight, 5)5	2.70	1.34	2.01	0.04
pol_coverage2NoMini	-0.43	0.14	-3.14	0.00
vh_fuelGasoline	0.17	0.07	2.58	0.01
claim_indOne	-0.13	0.07	-1.80	0.07
class_make2Class2&3	0.56	0.27	2.08	0.04
drv_age_lic1:poly(vh_weight, 4)1	-0.05	0.09	-0.58	0.56
drv_age_lic1:poly(vh_weight, 4)2	0.19	0.13	1.44	0.15
drv_age_lic1:poly(vh_weight, 4)3	0.11	0.14	0.83	0.41
drv_age_lic1:poly(vh_weight, 4)4	0.27	0.11	2.48	0.01
pol_bonus:class_make2Class2&3	-0.62	0.39	-1.61	0.11

Table 31: Maximum likelihood estimates, estimated standard errors, t-values and corresponding p-values, as well as R^2 and R^2_{adj} for model `LogLM2_inter_red`.

and its output is presented in Table 31. In this model `pol_bonus` is significant variable and we also keep the interaction `pol_bonus:class_make2`, even though it is significant on a 0.11 level. This interaction terms seems to increase the significance of the main effect term `pol_bonus`. Additionally, the value of R^2_{adj} has increased to 11.55%. If we perform an F-test for the removed interaction terms from `LogLM2_inter`,

```
anova(LogLM2_inter_red, LogLM2_inter)
```

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	2327	2606				
2	2309	2595	18	11.18	0.55	0.93

Table 32: F-test for the interaction effects removed from the model `LogLM2_inter`.

we obtain a p-value of 0.93, meaning that we can indeed remove them. Additinally, performing an F-test for the interaction terms in the model `LogLM2_inter_red` results in p-value of 0.04, which means that on a 0.05 level at least one of the interaction parameters in the model is significant. Now we proceed with model diagnostics for the model `LogLM2_inter_red`, since this model has highest R^2_{adj} while maintaining the significance of the terms in the model. In the comparison of the different modelling methods on Bad Driver Data however, we consider both `LogLM2_main_red` and `LogLM2_inter_red`.
First, we plot the internally studentized residuals of `LogLM2_inter_red` against the observation number in Figure 24. We expect most of the residuals to be randomly distributed around the zero within the interval [-3,3]. As observed in Figure 24, the residuals show random fluctuation around the 0, with only 3 observations being outside of the interval [-3,3]. The blue line on the plot at 0 denotes the smoother line of the residuals, the red line around 0 denotes the line $y = 0$ and the red dashed lines denote $y = -3$ and $y = 3$.

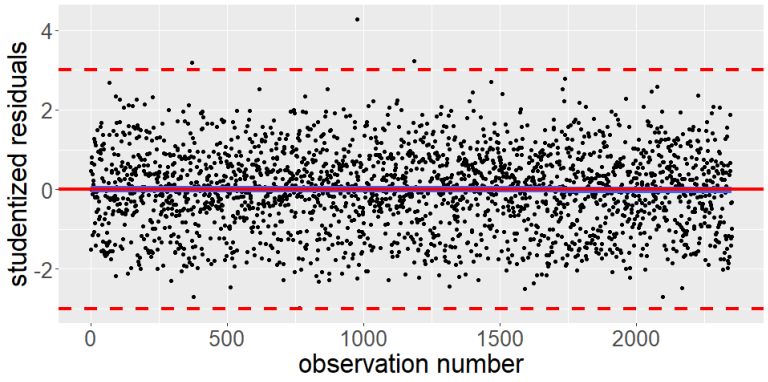

Figure 24: Internally studentized residuals of the model `LogLM2_inter_red`.

Second, in Figure 25 we observe the Q-Q plot of the internally studentized residuals, where the values follow the 45° line. This means that we do not have an indication that model the assumptions are not satisfied. We can spot that the left tail shows non-normal properties, but only for few observations.

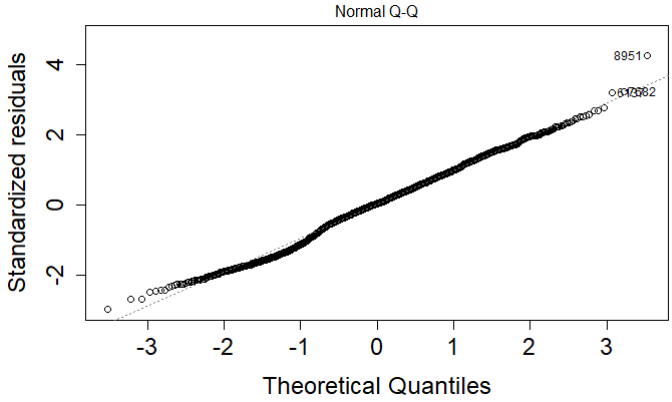

Figure 25: Q-Q plot of the internally studentized residuals of the model `LogLM2_inter_red`.

Third, the model has 168 high leverage points, but since these indicate only x-outliers we don't consider them as influential observations. None of the Cook's Distance values is bigger than 1, therefore there are no observations that need to be further studied or removed. Finally, we can conclude that given these results, we did not find any evedence against the model assumptions.

5.2 Gamma regression model

In the gamma regression models, we don't include `drv_age_lic1` and `drv_age_lic2` to avoid multicollinearity, because elimination of these two covariates showed most parsimonious fit. The main effects gamma model which includes the covariates presented in Section 3.3 is:

```
GamReg2_main<-glm(standr_claims ~ pol_bonus+pol_duration+
              pol_sit_duration+poly(drv_age1,4)+poly(drv_age2,3)+
              poly(vh_cyl,2)+poly(vh_din, 3)+poly(vh_speed, 2)+
              poly(vh_weight,5)+pol_coverage2+pol_usage2+vh_fuel+
              claim_ind+class_make2+gender2, data=data2,
              family=Gamma(link="log"))
```

whose output is presented in Table 33. We can observe that many of the main effects are not significant. The model's residual deviance is 2529.1 on 2321 degrees of freedom. If we use the `step` function to remove insignificant covariates, we end up with a model with only two covariates, which is less than the significant covariates in the model `GamReg2_main`. Therefore, removing insignificant covariates manually results in the reduced model:

```
GamReg2_main_red<-glm(standr_claims ~ pol_bonus+poly(drv_age1,4)+
                  vh_cyl+poly(vh_din, 3)+vh_weight+vh_fuel+
                  class_make2, data=data2,family=Gamma(link="log"))
```

$\phi = 1.96$	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-3.24	0.31	-10.31	0.00
pol_bonus	0.44	0.23	1.91	0.06
pol_duration	-0.00	0.01	-0.01	0.99
pol_sit_duration	0.01	0.02	0.24	0.81
poly(drv_age1, 4)1	3.32	2.42	1.37	0.17
poly(drv_age1, 4)2	-0.96	1.84	-0.52	0.60
poly(drv_age1, 4)3	-0.83	1.48	-0.56	0.58
poly(drv_age1, 4)4	3.45	1.52	2.28	0.02
poly(drv_age2, 3)1	-1.62	1.55	-1.04	0.30
poly(drv_age2, 3)2	-0.96	1.55	-0.62	0.54
poly(drv_age2, 3)3	0.71	1.45	0.49	0.63
poly(vh_cyl, 2)1	-5.50	2.82	-1.95	0.05
poly(vh_cyl, 2)2	2.47	2.55	0.97	0.33
poly(vh_din, 3)1	-1.43	5.32	-0.27	0.79
poly(vh_din, 3)2	-2.71	2.63	-1.03	0.30
poly(vh_din, 3)3	-2.63	2.18	-1.21	0.23
poly(vh_speed, 2)1	-1.10	3.92	-0.28	0.78
poly(vh_speed, 2)2	-1.16	2.25	-0.52	0.61
poly(vh_weight, 5)1	-5.91	3.16	-1.87	0.06
poly(vh_weight, 5)2	0.88	1.93	0.46	0.65
poly(vh_weight, 5)3	0.41	1.64	0.25	0.80
poly(vh_weight, 5)4	-1.38	1.57	-0.88	0.38
poly(vh_weight, 5)5	0.22	1.47	0.15	0.88
pol_coverage2NoMini	-0.13	0.18	-0.70	0.48
pol_usage2NoRetired	-0.07	0.16	-0.46	0.64
vh_fuelGasoline	0.16	0.09	1.83	0.07
claim_indOne	0.06	0.10	0.59	0.55
class_make2Class2&3	0.11	0.08	1.32	0.19
gender2AllOther	0.02	0.07	0.28	0.78

Table 33: Maximum likelihood estimates , estimated standard errors, Wald ratios and corresponding p-values for `GamReg2_main`.

$\phi = 1.94$	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-2.63	0.34	-7.81	0.00
pol_bonus	0.47	0.22	2.10	0.04
poly(drv_age1, 4)1	3.56	1.46	2.44	0.01
poly(drv_age1, 4)2	-0.30	1.47	-0.20	0.84
poly(drv_age1, 4)3	-0.91	1.45	-0.63	0.53
poly(drv_age1, 4)4	3.18	1.42	2.24	0.02
vh_cyl	-0.00	0.00	-1.64	0.10
poly(vh_din, 3)1	-4.08	2.58	-1.58	0.11
poly(vh_din, 3)2	-0.22	1.46	-0.15	0.88
poly(vh_din, 3)3	-2.49	1.43	-1.75	0.08
vh_weight	-0.00	0.00	-1.83	0.07
vh_fuelGasoline	0.22	0.08	2.79	0.01
class_make2Class2&3	0.13	0.08	1.56	0.12

Table 34: Maximum likelihood estimates , estimated standard errors, Wald ratios and corresponding p-values for `GamReg2_main_red`.

which has a residual deviance 2541.8 on 2337 degrees of freedom. In this model, the significance of the covariates is much better. We keep the covariate `class_make2` in the model because it will allow us more interactions later. Performing a partial deviance test

for the removed main effects terms from the model `GamReg2_main`, we obtain a p-value 0.98, which means that we can proceed with the reduced model. The residual deviance test on the model `GamReg2_main_red` results in p-value 1, therefore the model shows no lack of fit. Next we want to improve this model by allowing for interaction terms. For that purpose we fit the model `GamReg2_inter`. In this model however, the main effect term of `vh_weight` is takes as a polynomial of second degree, since a lot of interaction terms with `vh_weight` have this transformed form. The output of the model is presented in Table 35.

$\phi = 1.68$	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-3.84	1.15	-3.34	0.00
pol_bonus	1.95	1.66	1.17	0.24
poly(drv_age1, 4)1	-2.92	12.51	-0.23	0.82
poly(drv_age1, 4)2	17.69	12.48	1.42	0.16
poly(drv_age1, 4)3	-1.28	12.49	-0.10	0.92
poly(drv_age1, 4)4	10.41	11.13	0.93	0.35
vh_cyl	0.00	0.00	0.12	0.90
poly(vh_din, 3)1	-29.64	20.97	-1.41	0.16
poly(vh_din, 3)2	9.64	36.45	0.26	0.79
poly(vh_din, 3)3	-1.75	22.05	-0.08	0.94
poly(vh_weight, 2)1	12.24	17.41	0.70	0.48
poly(vh_weight, 2)2	-14.62	11.30	-1.29	0.20
vh_fuelGasoline	0.14	0.24	0.58	0.56
class_make2Class2&3	0.53	0.34	1.53	0.13
pol_bonus:poly(drv_age1, 4)1	7.31	11.48	0.64	0.52
pol_bonus:poly(drv_age1, 4)2	-21.07	12.50	-1.69	0.09
pol_bonus:poly(drv_age1, 4)3	-10.58	12.64	-0.84	0.40
pol_bonus:poly(drv_age1, 4)4	6.56	11.03	0.59	0.55
pol_bonus:vh_cyl	-0.00	0.00	-0.70	0.48
pol_bonus:poly(vh_din, 3)1	39.92	27.30	1.46	0.14
pol_bonus:poly(vh_din, 3)2	-18.83	65.66	-0.29	0.77
pol_bonus:poly(vh_din, 3)3	-3.15	40.89	-0.08	0.94
pol_bonus:poly(vh_weight, 2)1	-41.32	17.02	-2.43	0.02
pol_bonus:poly(vh_weight, 2)2	11.33	12.75	0.89	0.37
poly(drv_age1, 4)1:vh_cyl	0.00	0.01	0.06	0.95
poly(drv_age1, 4)2:vh_cyl	-0.00	0.01	-0.36	0.72
poly(drv_age1, 4)3:vh_cyl	0.00	0.01	0.58	0.56
poly(drv_age1, 4)4:vh_cyl	-0.01	0.01	-1.19	0.23
poly(drv_age1, 4)1:poly(vh_weight, 2)1	-50.13	134.18	-0.37	0.71
poly(drv_age1, 4)2:poly(vh_weight, 2)1	67.84	150.92	0.45	0.65
poly(drv_age1, 4)3:poly(vh_weight, 2)1	-108.87	127.79	-0.85	0.39
poly(drv_age1, 4)4:poly(vh_weight, 2)1	123.05	135.15	0.91	0.36
poly(drv_age1, 4)1:poly(vh_weight, 2)2	193.88	99.61	1.95	0.05
poly(drv_age1, 4)2:poly(vh_weight, 2)2	-276.67	131.66	-2.10	0.04
poly(drv_age1, 4)3:poly(vh_weight, 2)2	67.62	104.34	0.65	0.52
poly(drv_age1, 4)4:poly(vh_weight, 2)2	59.61	115.31	0.52	0.61
vh_cyl:vh_din	0.00	0.00	0.03	0.98
vh_cyl:poly(vh_weight, 2)1	0.01	0.01	1.16	0.25
vh_cyl:poly(vh_weight, 2)2	0.00	0.00	0.64	0.52
poly(vh_weight, 2)1:vh_din	-0.06	0.10	-0.60	0.55
poly(vh_weight, 2)2:vh_din	-0.02	0.08	-0.20	0.84
pol_bonus:class_make2Class2&3	-0.60	0.50	-1.22	0.22
vh_fuelGasoline:vh_din	0.00	0.00	0.06	0.95

Table 35: Maximum likelihood estimates , estimated standard errors, Wald ratios and corresponding p-values for `GamReg2_inter`.

The model `GamReg2_inter` has residual deviance of 2478.9 on 2307 degrees of freedom and contains a lot of insignificant terms. If we perform a backward elimination using the BIC criterion, with the main effects model `GamReg2_main_red` as a lower limit, all of the interaction terms are removed. Therefore, we manually remove step by step the

insignificant terms, starting with the ones who have the biggest p-value. This way we obtain the reduced interaction model `GamReg2_inter_red`,

```
GamReg2_inter_red<-glm(standr_claims ~ pol_bonus+poly(drv_age1,4)+
                  vh_cyl+poly(vh_din,3)+poly(vh_weight,2)+vh_fuel+
                  class_make2+pol_bonus:poly(drv_age1,2)+
                  pol_bonus:vh_din+pol_bonus:vh_weight+
                  poly(drv_age1,2):poly(vh_weight,2)+vh_cyl:vh_weight,
                  data=data2,family=Gamma(link="log"))
```

$\phi = 1.70$	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-2.86	0.28	-10.09	0.00
pol_bonus	1.68	0.93	1.80	0.07
poly(drv_age1, 4)1	-4.56	7.59	-0.60	0.55
poly(drv_age1, 4)2	17.73	8.11	2.19	0.03
poly(drv_age1, 4)3	-3.35	1.64	-2.05	0.04
poly(drv_age1, 4)4	5.23	1.51	3.47	0.00
vh_cyl	-0.00	0.00	-2.22	0.03
poly(vh_din, 3)1	-24.19	10.25	-2.36	0.02
poly(vh_din, 3)2	-2.65	2.04	-1.30	0.19
poly(vh_din, 3)3	-2.24	1.42	-1.58	0.11
poly(vh_weight, 2)1	4.88	12.02	0.41	0.68
poly(vh_weight, 2)2	-2.32	2.20	-1.05	0.29
vh_fuelGasoline	0.17	0.08	2.07	0.04
class_make2Class2&3	0.12	0.08	1.57	0.12
pol_bonus:poly(drv_age1, 2)1	9.81	10.84	0.90	0.37
pol_bonus:poly(drv_age1, 2)2	-24.09	11.24	-2.14	0.03
pol_bonus:vh_din	0.02	0.01	2.01	0.05
pol_bonus:vh_weight	-0.00	0.00	-2.43	0.02
poly(vh_weight, 2)1:poly(drv_age1, 2)1	-43.04	71.34	-0.60	0.55
poly(vh_weight, 2)2:poly(drv_age1, 2)1	143.22	71.38	2.01	0.04
poly(vh_weight, 2)1:poly(drv_age1, 2)2	2.21	83.88	0.03	0.98
poly(vh_weight, 2)2:poly(drv_age1, 2)2	-296.70	95.66	-3.10	0.00
vh_cyl:vh_weight	0.00	0.00	1.72	0.09

Table 36: Maximum likelihood estimates , estimated standard errors, Wald ratios and corresponding p-values for `GamReg2_inter_red`.

which has residual deviance of 2496.6 on 2327 degrees of freedom. Although the main effect term `class_make2` has p-value of 0.12, we keep it in the model because it improves the significance of other effects. If we perform a partial deviance test for the removed terms from model `GamReg2_inter` we obtain a p-value of 0.96, therefore we can proceed with the reduced model `GamReg2_inter_red`. Additionally, performing a partial deviance test for the interaction terms in the model `GamReg2_inter_red` which were added to `GamReg2_main_red` gives p-value of 0.003, so we prefer the model `GamReg2_inter_red` over `GamReg2_main_red`. The residual deviance test for `GamReg2_inter_red` results in p-value 1, so the model does not show a lack of fit. In the comparison of the models, we focus on `GamReg2_main_red` and `GamReg2_inter_red` as best gamma regression models on the Good Driver training data set.

5.3 D-vine quantile regression model

For Bad Driver Data, apart from nonparametric estimations of the marginals, we also estimate the marginals parametrically using finite mixture of distributions. Depending on our marginal estimation approach, we fit the pair copulas of the models parametrically or nonparametrically, which results in fully parametric and nonparametric D-vine quantile regression models. We rely on the all covariates presented in Section 3, including `pol_bonus`, where we omit the covariates `drv_age2` and `drv_age_lic2` because they contain a lot of zeros for the observations where driver 2 is not present in the policy. First we present the empirical normalized contour plots of the continuous variables in Figure 26, which are obtained by transforming the original data to copula data scale using marginal empirical distributions.

Figure 26: Bad Driver training data set. *Lower:* empirical normalized contour plots for the pair copulas, *diagonal:* histogram of the margins, *upper:* pairs plots of copula data and their Kendall's taus.

The empirical copula data histograms of `pol_duration`, `pol_sit_duration`, `pol_diff_duration` and `vh_sale_end` do not look uniform. Some of the pair copulas where one of the marginals is `pol_duration`, `pol_sit_duration` or `pol_diff_duration` look strange, which indicate that we may need to transform these covariates to ordinal covariates. We can spot high Kendall's taus up to 0.86 and we expect that some of the pair copulas with high Kendall's tau will be included in our models. The pair copulas of our response variables, `standr_claims` and `ln_standr_claims`, and the continuous covariates, are identical with same Kendall's tau values respectively. Therefore, assuming that our data in on the copula data scale, we would expect to obtain the same order of covariates in the D-vine regression models, regardless whether our response is `standr_claims` or `ln_standr_claims`. We will use the empirical contour plots in Figure 26 to compare them with the normalized contour plots of the fitted pair copulas in our D-vine regression models with respective marginals. In the next two sections we present the nonparametric and parametric D-vine regression models fitted on the Bad Driver training data set, whose order of variables in given in Table 37. Since the function `vinereg` (Nagler (2022)) in R allows only for ordinal covariates, we transform the discrete covariates to ordinal before we fit the models and their levels are sorted based on their influence on the response.

Order	**DVReg2_nonpar_orig**	**DVReg2_nonpar_ln**	**DVReg2_par_orig**	**DVReg2_par_ln**
1	standr_claims	ln_standr_claims	standr_claims	ln_standr_claims
2	vh_cyl	vh_weight	vh_cyl	vh_cyl
3	pol_usage2	drv_age_lic1	drv_age1	drv_age1
4	gender2	vh_cyl	vh_weight	vh_weight
5	pol_coverage2	vh_speed	pol_bonus	pol_bonus
6	claim_ind	vh_sale_end	vh_age	vh_age
7	drv_age1	claim_ind	vh_sale_end	vh_sale_end
8	vh_sale_end	pol_coverage2		
9	vh_fuel	vh_fuel		
10	vh_weight			
11	pol_payd			

Table 37: Order of variables in the D-vine regression models `DVReg2_nonpar_orig`, `DVReg2_nonpar_ln`, `DVReg2_par_orig` and `DVReg2_par_ln`.

Nonparametric D-vine quantile regression

Before we fit nonparametric D-vine regression models, we want to investigate the nonparametric estimation of the continuous marginals, in particular, to see whether some of them need to be transformed to ordinal covariates (in case the histograms of the pseudo data marginals do not look uniform), and to transform the discrete covariates to ordinal. First, we estimate the continuous variables nonparametrically using the kernel smoothing estimator implemented in the R function `kde1d`, in the R package `kde1d` (Nagler and Vatter (2022)). Although this function is implemented in the R function `vinereg` (Nagler (2022)), which we use to fit a D-vine regression models, in the preliminary step we want to see how the histograms of the continuous marginals of our copula data look like. For that purpose, first in Figure 27 we present the histograms of the continuous marginals and their respective nonparametrically estimated densities.

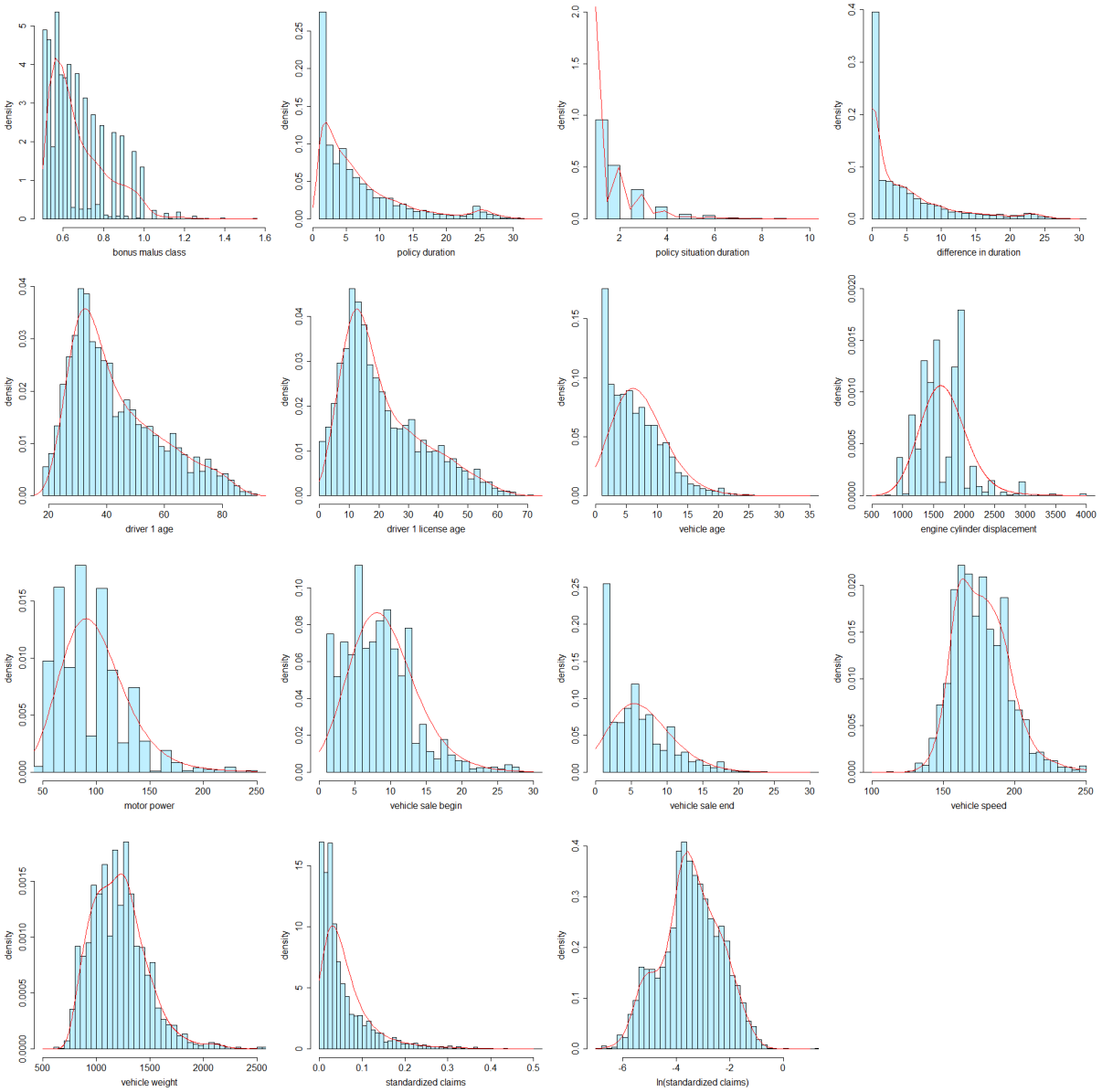

Figure 27: Histograms of the continuous marginals in Bad Driver training data set. The red line denotes their `kde1d` estimators.

The `kde1d` estimators can't fully capture the multimodality of *engine cylinder displacement*, *motor power*, *vehicle weight* and it underestimates the modes of *bonus malus class*, *policy duration*, *difference in duration*, *vehicle age*, *standardized claims*. In Figure 28 we present the histograms of marginals on copula data scale, which were obtained using the nonparametric estimators from Figure 27. The histograms of *policy duration*, *policy situation duration* and *difference in duration* look particularly nonuniform, which is why we transform them to ordinal covariates `dur_ordinal`, `sitdur_ordinal` and `diffdur_ordinal`. They have 33, 12 and 32 levels respectively. Additionally, the histograms of *vehicle age*, *engine cylinder displacement*, *motor power*, *vehicle sale begin*, *vehicle sale end* and *standardized claims* do not look so uniform, however, we continue to work with them in our nonparametric D-vine regression models.

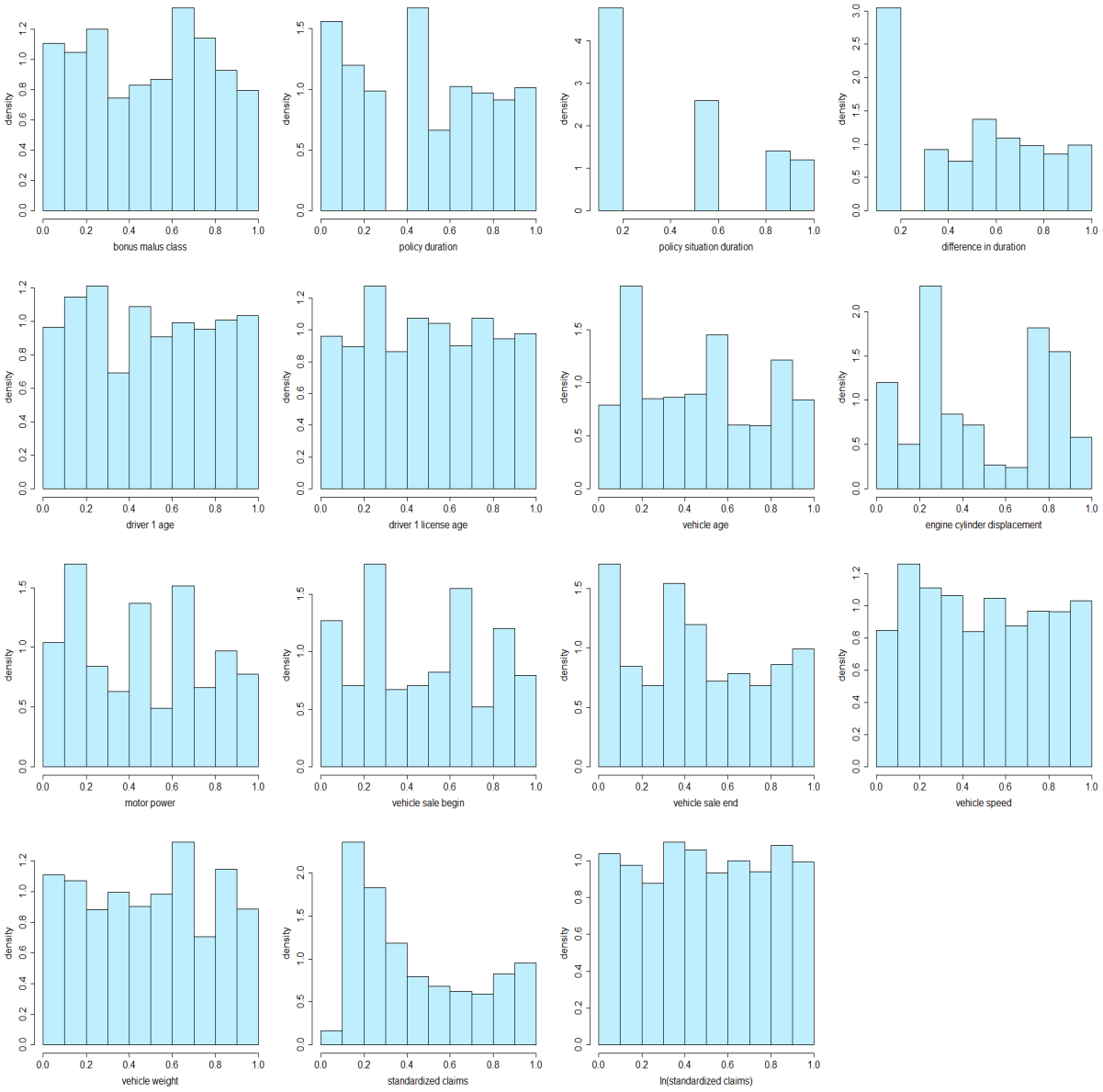

Figure 28: Histograms of the continuous marginals of the copula scale data in Bad Driver training data set, which were transformed on copula scale using their respective estimators from Figure 27.

Finally, as a last preliminary step we transform the discrete covariates in the Bad Driver Data to ordinal, and their levels are sorted based on their influence on the response variable. Currently it is possible to consider ordinal covariates for D-vine regression with the function `vinereg` in R only if we estimate the marginals nonparametrically using the `kde1d` estimator (which is implemented in the `vinereg` function in R (Nagler (2022)). This is not be possible if the continuous marginals are estimated differently, which is our case for the parametric D-vine regression models in the next section. Therefore, we only consider the discrete covariates from the Bad Driver Data in our nonparametric D-vine regression models.

We present the nonparametric model `DVReg2_nonpar_orig`, which predicts the response variable on original scale *standardized claims*, and the nonparametric model `DVReg2_nonpar_ln`, which predicts the transformed response variable *ln(standardized claims)*. The order of the variables in the models is presented in Table 38 and the normalized contour plots of the fitted pair copulas of the models in Figure 29 and Figure 30, respectively.

Order	**DVReg2_nonpar_orig**	**DVReg2_nonpar_ln**
1	standr_claims	ln_standr_claims
2	vh_cyl	vh_weight
3	pol_usage2	drv_age_lic1
4	gender2	vh_cyl
5	pol_coverage2	vh_speed
6	claim_ind	vh_sale_end
7	drv_age1	claim_ind
8	vh_sale_end	pol_coverage2
9	vh_fuel	vh_fuel
10	vh_weight	
11	pol_payd	

Table 38: Order of variables in the D-vine regression models `DVReg2_nonpar_orig` and `DVReg2_nonpar_ln`.

In the model `DVReg2_nonpar_orig`, more than the half of the covariates are ordinal, and they are included early in the model, which differs from the model `DVReg2_nonpar_ln` where the most influential covariates are continuous. All of the discrete covariates present in the model `DVReg2_nonpar_ln` are included in the model `DVReg2_nonpar_orig` too, and their common continuous covariates are `vh_weight`, `vh_cyl` and `vh_sale_end`.

Based on Figure 29, few of the pair copulas in the model `DVReg2_nonpar_orig` look particularly nonparametric; one example is the pair copula (`pol_usage2`, `pol_coverage2`;`gender`) in the second tree of the D-vine. Additionally, we can notice that not many normalized contour plots in the model indicate high Kendall's taus, at least not in the first tree, which is given by the last row of Figure 29. As expected, the first covariate in the model is the covariate `vh_cyl`, which indicated highest dependence with the response `standr_claims` (in absolute terms) in Figure 26. We are interested in comparing the fitted pair copulas with continuous marginals in the first tree of the model, (`standr_claims`, `vh_cyl`) and (`drv_age1`, `vh_sale_end`), with their respective empirical pair copulas from Figure 26. The fitted pair copulas look similar to their empirical pair copulas, with only the fitted normalized contour plot of (`standr_claims`, `vh_cyl`) looking different than the empirical one for higher contour levels.

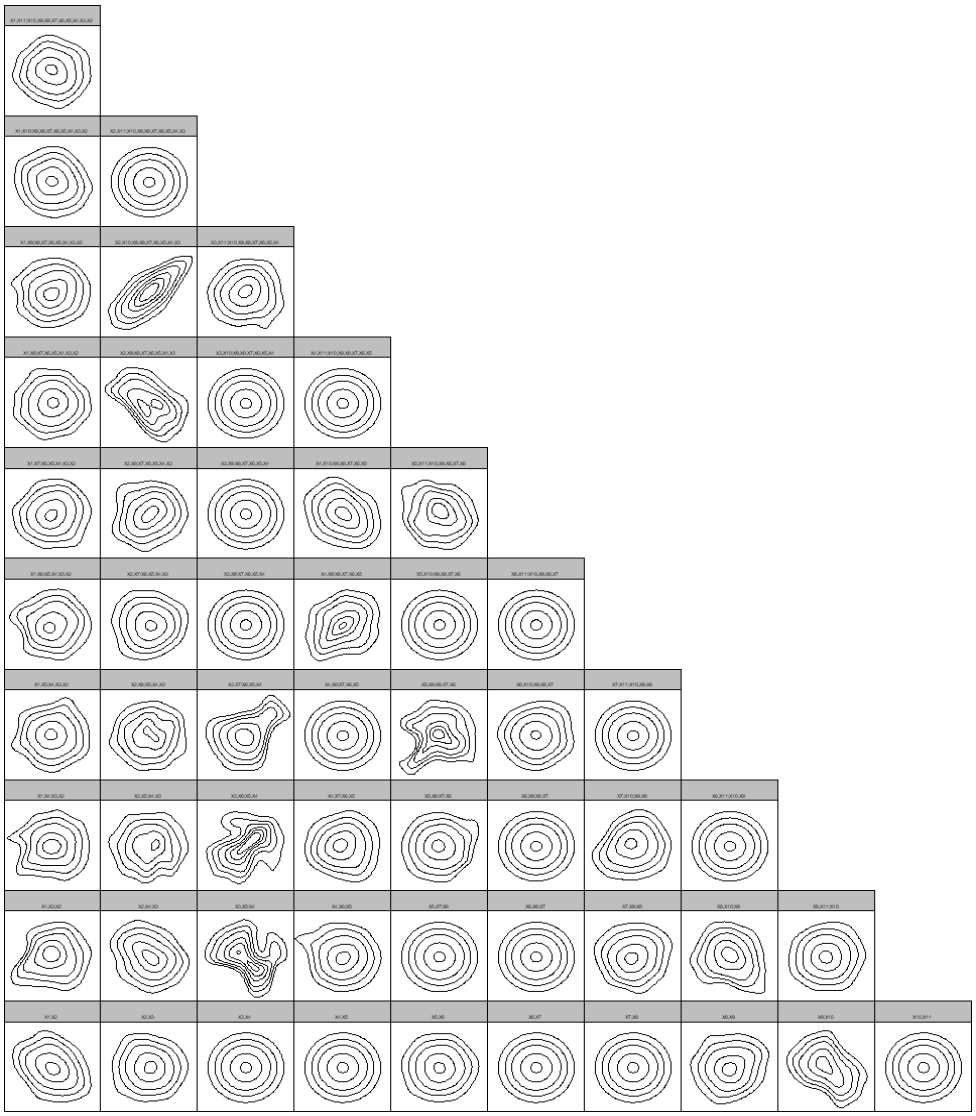

Figure 29: Normalized fitted contour plots for the pair copulas of the D-vine regression model `DVReg2_nonpar_orig`, where the variables of the model are presented by X1=standr_claims, X2=vh_cyl, X3=pol_usage2, X4=gender2, X5=pol_coverage2, X6=claim_ind, X7=drv_age1, X8=vh_sale_end, X9=vh_fuel, X10=vh_weight and X11=pol_payd, depending on their order in Table 38.

In Figure 30 we can see the normalized contour plots of the fitted pair copulas in the model `DVReg2_nonpar_ln`, which predicts the transformed response *ln(standardized claims)*. The advantage of using the transformed response variable over the original one is that it guarantees positive values of the fitted values for *standardized claims*, which we can obtain from the fitted values of the model using the exponential function. Compared to the pair copulas of the model `DVReg2_nonpar_orig`, only one of the fitted pair copulas of `DVReg2_nonpar_ln` looks strongly nonparametric, the copula of the pair (`vh_speed`, `vh_sale_end`). By comparing the normalized fitted contour plots of the pair copulas with continuous marginals in the first tree of the model with their respective empirical contour plots, we can see that the fitted normalized contour plot of the pairs look similar to their empirical normalized contour plots, except for the pair (`vh_speed`, `vh_sale_end`).

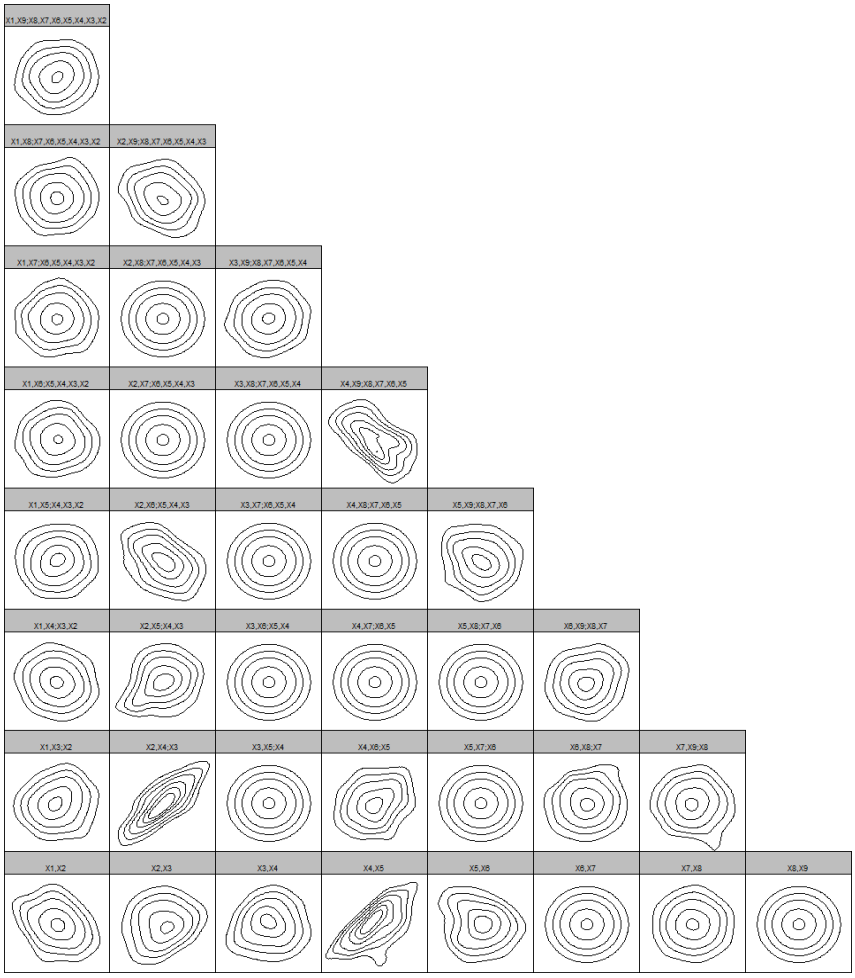

Figure 30: Normalized fitted contour plots for the pair copulas of the D-vine regression model `DVReg2_nonpar_ln`, where the variables of the model are given by X1=ln_standr_claims, X2=vh_weight, X3=drv_age_lic1, X4=vh_cyl, X5=vh_speed, X6=vh_sale_end, X7=claim_ind, X6=pol_coverage2 and X7=vh_fuel, depending on their order in Table 38.

Parametric D-vine quantile regression

As previously discussed, the R function `vinereg` (Nagler (2022)) does not allow for ordinal covariates when the data is on pseudo data scale. Therefore, in this section we only analyse the continuous marginals of the data, where we discard the covariates `pol_duration`, `pol_sit_duration` and `pol_diff_duration`, since they were transformed to ordinal covariates in the nonparametric case. As a parametric estimation approach for the continuous marginals we use finite distribution mixture of either normal or skew-normal distributions, which is implemented in the R package `mixsmsn` (Prates, Lachos and Cabral (2013)). This approach is beneficial when the data is skewed and multimodal. The choice of distribution in the mixture and the number of modes is based on the smallest BIC value of several marginal estimations. The resulting estimators of each continuous marginal with their respective parameters are presented in Table 39 and visualized in Figure 31.

Variable	Distribution of the mixture's components	Number of components	Parameters						
				Component i					
				1	2	3	4	5	6
pol_bonus	Skew-normal	6	μ_i	0.507	0.775	0.587	0.925	1.074	0.673
			σ_i^2	0.001	0.006	0.001	0.003	0.027	0.004
			λ_i	22.016	0.943	0.734	0.581	4.411	0.982
			p_i	0.239	0.145	0.315	0.091	0.011	0.198
drv_age1	Skew-normal	2	μ_i	52.055	25.984				
			σ_i^2	229.189	140.109				
			λ_i	0.910	2.529				
			p_i	0.336	0.664				
drv_age_lic1	Skew-normal	2	μ_i	6.825	28.303				
			σ_i^2	97.320	204.527				
			λ_i	2.049	1.242				
			p_i	0.683	0.317				
vh_age	Skew-normal	2	μ_i	0.945	6.148				
			σ_i^2	20.170	43.318				
			λ_i	8.852	4.376				
			p_i	0.63	0.37				
vh_cyl	Skew-normal	4	μ_i	1238.917	1397.853	1858.883	1909.683		
			σ_i^2	26338.147	18465.116	15180.417	706867.914		
			λ_i	-3.612	1.216	1.566	4.383		
			p_i	0.147	0.429	0.345	0.079		
vh_din	Skew-normal	3	μ_i	70.500	124.118	81.660			
			σ_i^2	58.665	5356.762	990.761			
			λ_i	-1.040	3.054	1.425			
			p_i	0.250	0.057	0.693			
vh_sale_begin	Skew-normal	4	μ_i	9.697	4.383	13.631	2.352		
			σ_i^2	6.852	8.700	52.037	0.702		
			λ_i	1.312	1.502	3.509	-0.245		
			p_i	0.334	0.480	0.093	0.093		
vh_sale_end	Skew-normal	4	μ_i	9.332	11.483	0.863	4.944		
			σ_i^2	5.529	30.884	1.943	3.511		
			λ_i	0.963	2.667	27.200	1.087		
			p_i	0.176	0.086	0.328	0.410		
vh_speed	Normal	3	μ_i	183.348	160.923	186.761			
			σ_i^2	122.340	65.253	594.618			
			p_i	0.370	0.328	0.302			
vh_weight	Normal	3	μ_i	960.569	1618.37	1271.034			
			σ_i^2	11883.951	157207.96	31408.245			
			p_i	0.296	0.094	0.610			
standr_claims	Skew-normal	3	μ_i	0.150	0.003	0.048			
			σ_i^2	0.019	0.0007	0.004			
			λ_i	4.730	17.361	3.361			
			p_i	0.056	0.683	0.261			
ln_standr_claims	Normal	3	μ_i	-3.673	-2.522	-5.039			
			σ_i^2	0.189	0.533	0.284			
			p_i	0.363	0.432	0.205			

Table 39: Finite mixture estimators for the continuous marginals in Bad Driver training data set and their respective parameters.

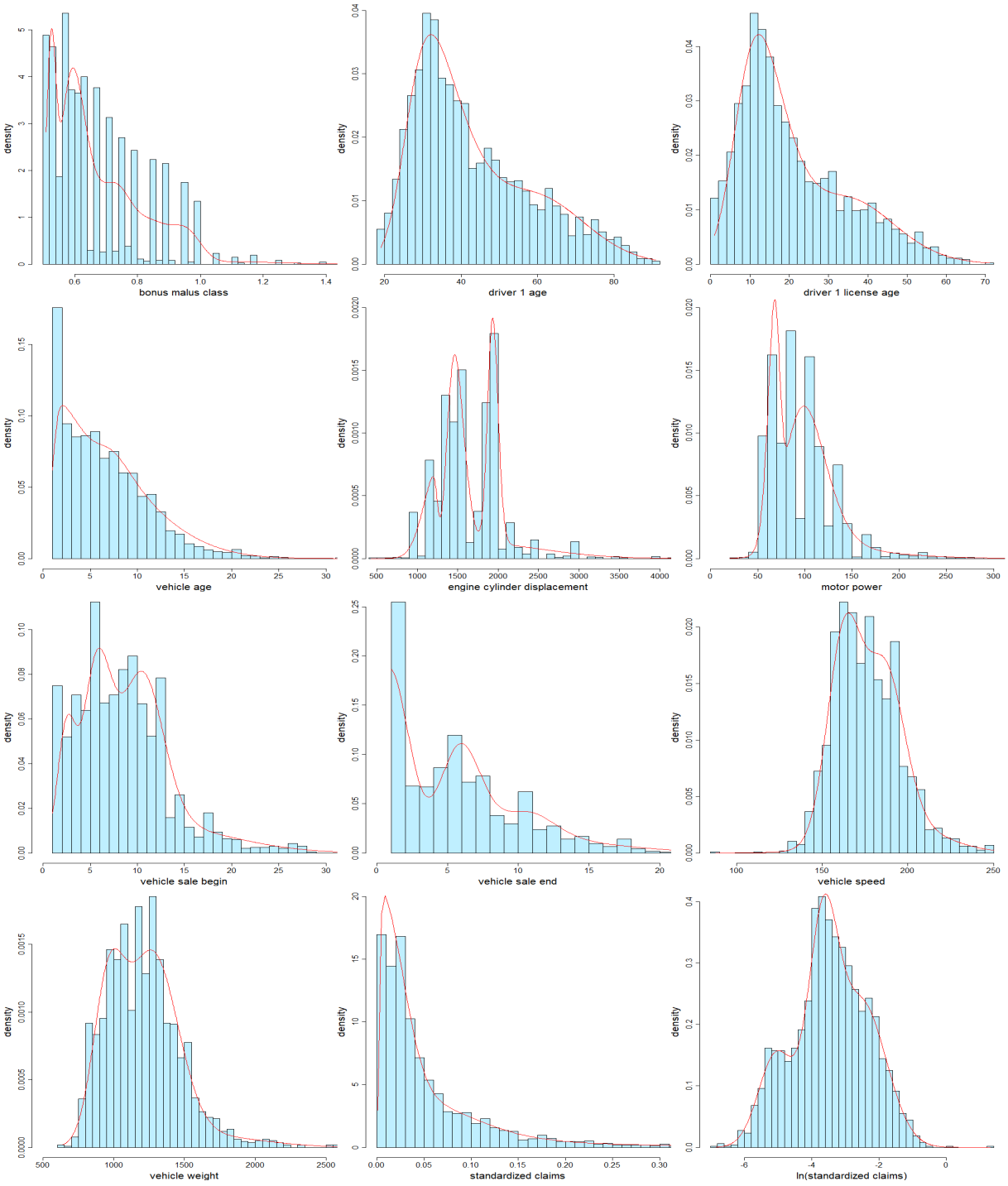

Figure 31: Histograms of the continuous marginals in Bad Driver training data set. The red line denotes their parametric estimators from Table 39.

Using these estimators, we transform the data to copula scale data and the histograms of its marginals are presented in Figure 32. We expect them to look uniformly distributed, which in the case of *bonus malus class*, *vehicle age*, *engine cylinder displacement*, *motor power*, *vehicle sale begin* and *vehicle sale end* do not look so uniform, however these pseudo data histograms look much better than the pseudo data histograms obtained by

the nonparametric `kde1d` estimators which are presented in Figure 27.

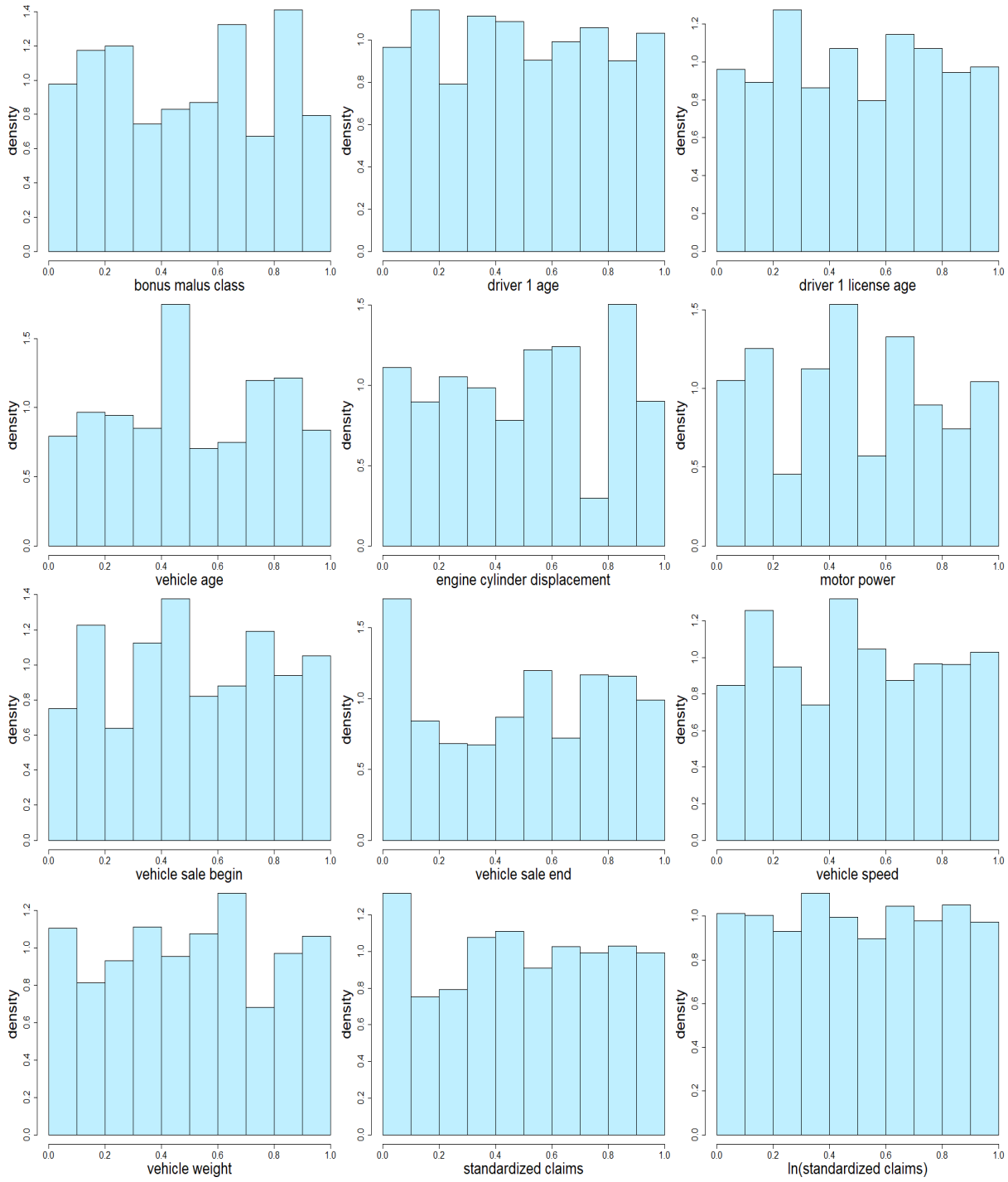

Figure 32: Histograms of the continuous marginals of the copula scale data in Bad Driver training data set obtained from their parametric estimators in Table 39,

Now we present the parametric models `DVReg2_par_orig` and `DVReg2_par_ln` fitted on the pseudo data from Figure 32, with response variables `standr_claims` and `ln_standr_claims` respectively, and parametric bivariate copulas. The models have the same covariates with the same order, presented in Table 40.

Order	Variable
2	vh_cyl
3	drv_age1
4	vh_weight
5	pol_bonus
6	vh_age
7	vh_sale_end

Table 40: Order of variables in the D-vine regression model `DVReg2_par_orig`, where the first variable is standr_claims, and the D-vine regression model `DVReg2_par_ln` where the first variable is ln_standr_claims.

Tree	Edge	Conditioned	Conditioning	Var_types	Family	Rotation	Parameters	Df	Tau	LogLik
1	1	1,2;		c,c	gaussian	0	-0.27	1	-0.17	91.07
1	2	2,3;		c,c	bb8	90	1.07,0.97	2	-0.03	5.12
1	3	3,4;		c,c	clayton	0	0.13	1	0.06	15.97
1	4	4,5;		c,c	bb8	90	1.49,0.74	2	-0.09	23.03
1	5	5,6;		c,c	bb8	0	1.07,0.92	2	0.02	2.14
1	6	6,7;		c,c	bb8	0	6.92,0.95	2	0.73	2214.03
2	1	1,3;	2	c,c	frank	0	0.55	1	0.06	10.20
2	2	2,4;	3	c,c	t	0	0.78, 13.68	2	0.57	1109.18
2	3	3,5;	4	c,c	bb8	90	1.54,0.97	2	-0.20	164.30
2	4	4,6;	5	c,c	bb8	90	1.28,1.00	2	-0.13	81.23
2	5	5,7;	6	c,c	indep	0		0	0	0
3	1	1,4;	2,3	c,c	gaussian	0	-0.09	1	-0.06	9.82
3	2	2,5;	3,4	c,c	gaussian	0	-0.04	1	-0.03	2.16
3	3	3,6;	4,5	c,c	bb8	270	1.14, 0.80	2	-0.03	3.35
3	4	4,7;	5,6	c,c	bb7	270	1.04,0.05	2	-0.04	9.49
4	1	1,5;	2,3,4	c,c	clayton	180	0.05	1	0.02	3.49
4	2	2,6;	3,4,5	c,c	bb8	0	4.68,0.56	2	0.33	305.39
4	3	3,7;	4,5,6	c,c	joe	0	1.03	1	0.02	3.37
5	1	1,6;	2,3,4,5	c,c	frank	0	-0.29	1	-0.03	2.88
5	2	2,7;	3,4,5,6	c,c	gumbel	90	1.02	1	-0.02	9.33
6	1	1,7;	2,3,4,5,6	c,c	gaussian	0	0.05	1	0.03	4.25

Table 41: Fitted pair copulas of the D-vine regression model `DVReg2_par_orig` with their family parameters, Kendall's τ and log likelihood.

Tree	Edge	Conditioned	Conditioning	Var_types	Family	Rotation	Parameters	Df	Tau	LogLik
1	1	1,2;		c,c	gaussian	0	**-0.28**	1	**-0.18**	**96.12**
1	2	2,3;		c,c	bb8	90	1.07,0.97	2	-0.03	5.12
1	3	3,4;		c,c	clayton	0	0.13	1	0.06	15.97
1	4	4,5;		c,c	bb8	90	1.49,0.74	2	-0.09	23.03
1	5	5,6;		c,c	bb8	0	1.07,0.92	2	0.02	2.14
1	6	6,7;		c,c	bb8	0	6.92,0.95	2	0.73	2214.03
2	1	1,3;	2	c,c	frank	0	**0.57**	1	0.06	**10.70**
2	2	2,4;	3	c,c	t	0	0.78, 13.68	2	0.57	1109.18
2	3	3,5;	4	c,c	bb8	90	1.54,0.97	2	-0.20	164.30
2	4	4,6;	5	c,c	bb8	90	1.28,1.00	2	-0.13	81.23
2	5	5,7;	6	c,c	indep	0		0	0	0
3	1	1,4;	2,3	c,c	gaussian	0	-0.09	1	-0.06	**10.23**
3	2	2,5;	3,4	c,c	gaussian	0	-0.04	1	-0.03	2.16
3	3	3,6;	4,5	c,c	bb8	270	1.14, 0.80	2	-0.03	3.35
3	4	4,7;	5,6	c,c	bb7	270	1.04,0.05	2	-0.04	9.49
4	1	1,5;	2,3,4	c,c	clayton	180	0.05	1	0.02	**2.90**
4	2	2,6;	3,4,5	c,c	bb8	0	4.68,0.56	2	0.33	305.39
4	3	3,7;	4,5,6	c,c	joe	0	1.03	1	0.02	3.37
5	1	1,6;	2,3,4,5	c,c	frank	0	**-0.30**	1	-0.03	**2.91**
5	2	2,7;	3,4,5,6	c,c	gumbel	90	1.02	1	-0.02	9.33
6	1	1,7;	2,3,4,5,6	c,c	gaussian	0	**0.06**	1	**0.04**	**4.35**

Table 42: Fitted pair copulas of the D-vine regression model `DVReg2_par_ln` with their family parameters, Kendall's τ and log likelihood. The values that differ from the fitted pair copulas of the model `DVReg2_par_orig` are identified in gray.

In Table 41 and 42, we present the fitted pair copulas of the models `DVReg2_par_orig` and `DVReg2_par_ln`. The fitted pair copula families in both models are the same, with only few values that slightly differ (identified in gray in Table 42). The normalized contour plots of both models are the same and are presented in Figure 33. The pairs (`vh_age`, `vh_sale_end`) and (`vh_cyl`, `vh_weight;drv_age1`) have high Kendall's tau values. By comparing the fitted normalized contour plots of the pairs in the first tree of the D-vine regression models with their respective empirical contour plots, we can see that the fitted normalized contour plot of the pair (`standr_claims`, `vh_cyl`) (respectively the pair (`ln_standr_claims`, `vh_cyl`)) does not look like the empirical contour plot for the high contour levels. Additionally, the normalized fitted contour plot of the pair (`vh_age`, `vh_sale_end`) does not fully explain the structure of the empirical normalized contour plot, since the empirical contour plot shows nonparametric properties. The other fitted pair copulas in the first tree however, are very similar to their respective empirical pair copulas.

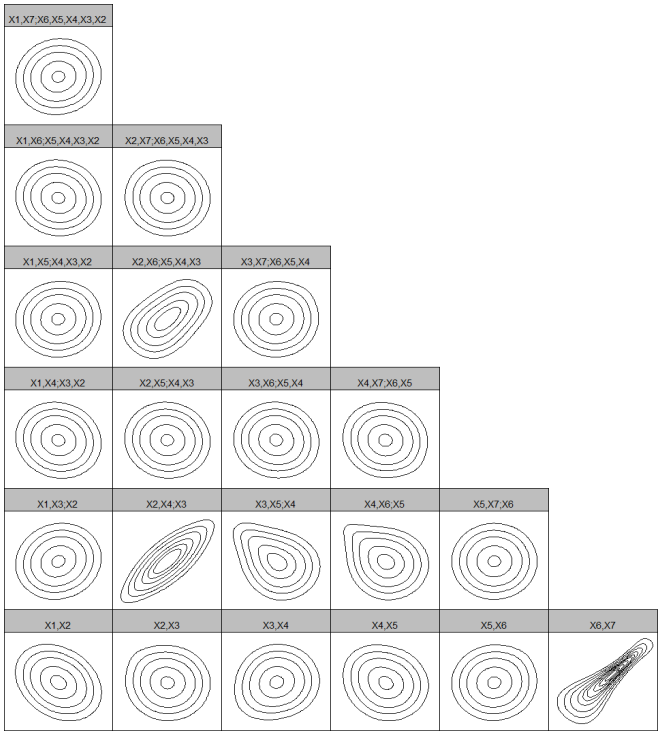

Figure 33: Normalized fitted contour plots for the pair copulas of the D-vine regression models `DVReg2_par_orig` and `DVReg2_par_ln`, where the variable X1 is standr_claims and ln_standr_claims respectively, and the variables X2=vh_cyl, X3=drv_age1, X4=vh_weight, X5=pol_bonus, X6=vh_age, X7=vh_sale_end are ordered as in Table 40.

In the next step, which is comparison of the different regression models fitted on Bad Driver training data set, we consider all four D-vine regression models, `DVReg2_nonpar_orig`, `DVReg2_nonpar_ln`, `DVReg2_par_orig` and `DVReg2_par_ln`. In particular, we would like to know whether the parametric or nonparametric D-vine regression models perform better and whether the choice of the response variable makes a difference in the models' explanatory power, which we can investigate by comparing the models `DVReg2_par_orig` and `DVReg2_par_ln`.

5.4 Comparison of the models

In this section we proceed with the comparison of the performance of the different regression models on Bad Driver Data. In particular, we are interested whether the fitted D-vine regression models perform better in severity modelling than the commonly used lognormal and gamma regression models, as well as linear quantile regression models.
We fit linear quantile regression models using both the original response and the transformed response variable, which have the same covariates as the best D-vine regression model with the respective response variable. More precisely, based on Table 43, the D-vine model `DVReg2_nonpar_orig` shows better values than `DVReg_par_orig` for the log likelihood, the training and test error, and the interval score for the 90% prediction interval. Similarly, the model `DVReg2_nonpar_ln` has better values than `DVReg2_par_ln` for all performance measures, except for the measures which use penalty for complexity i.e. AIC and BIC. Therefore we fit a linear quantile regression model `LQR2_orig` using the original response variable *standaridzed claims* and the covariates included in the model `DVReg2_nonpar_orig`, and the linear quantile regression model `LQR2_ln` using the original response variable *ln(standaridzed claims)* and the covariates included in the model `DVReg2_nonpar_ln`.
As performance measures for the different regression methods we use the log likelihood, AIC, BIC, training and test error and interval scores on level $\alpha = 0.05$ and $\alpha = 0.1$, all of which are calculated on the original scale *standardized claims*. These measures were extensively studied in Section 2.8 for every regression model separately. The log likelihood, AIC, BIC and the training error are calculated on the Bad Driver training data set, whereas the test error and the interval scores on the Bad Driver test data set. For the calculation of AIC and BIC for the lognormal, gamma and linear quantile regression models as number of parameters we use the number of estimated parameters in the model. For the nonparametric D-vine regression model, as number of parameters we use the sum of the degrees of freedom of each pair copula in the model, whereas for the parametric D-vine regression model we sum the number of parameters used to estimate the marginals included in the model and the number of parameters used to fit the bivariate copulas of the model. The models with smaller AIC and BIC values, as well as larger log likelihood values are considered to show a better performance on the data set.

	df	**LogLik**	**AIC**	**BIC**	**Training error**	**Test error**	**Interval score $\alpha = 0.05$**	**Interval score $\alpha = 0.1$**
LogLM2_main_red	19	4661.44	-9284.88	**-9175.40**	0.010709	0.007603	**0.00966**	0.01655
LogLM2_inter_red	24	4667.29	**-9286.58**	-9148.29	0.010693	0.007610	0.00970	0.01668
GamReg2_main_red	14	4491.27	-8954.55	-8873.88	0.009995	**0.006723**	0.01080	0.01721
GamReg2_inter_red	24	4515.82	-8983.64	-8845.35	**0.009865**	0.006821	0.01486	0.01964
DVReg2_nonpar_orig	444	4514.92	-8142.65	-5586.56	0.010327	0.007074	0.01022	0.01597
DVReg2_nonpar_ln	406	**4884.49**	-8957.32	-6618.87	0.010529	0.007423	0.00994	**0.01560**
DVReg2_par_orig	123	4322.59	-8399.17	-7690.42	0.010707	0.007573	0.00997	0.01629
DVReg2_par_ln	120	4685.94	-9131.88	-8440.42	0.010678	0.007550	0.01008	0.01628
LQReg2_orig	11	3562.30	-7102.61	-7039.23	0.010747	0.007513	0.00979	0.01588
LQReg2_ln	9	4535.58	-9053.16	-9001.30	0.010776	0.007618	0.01004	0.01613

Table 43: Comparison of different regression models on Bad Driver Data based on log likelihood, AIC, BIC, training error, test error and interval score values. The best values per column are identified in gray.

For the calculation of training and test error of the D-vine and linear quantile regression models, as fitted values we use the conditional 0.5 quantiles. Additionally, for the calculation of the interval score values for the lognormal and gamma regression, as upper and lower limits of the interval we use the upper and lower limits of the respective prediction interval, whose formula can be found in Section 2.8. For the linear quantile regression models, when computing the interval scores, no quantile crossing was present, therefore we can use the formula for the interval score given in Section 2.8.
Based on Table 43, we can see that the D-vine quantile regression model `DVReg2_nonpar_ln` shows best performance according to log likelihood and interval score values on level $\alpha = 0.05$. However, the nonparametric D-vine regression models do not lag much behind gamma regression models when it comes to training and test errors, too. Both the parametric and nonparametric D-vine regression models perform better when the transformed response is used. This can be especially noticed by comparing the models `DVReg2_par_orig` and `DVReg2_par_ln`, which have the same order of covariates, just different response variables. The model `DVReg2_par_ln` performs better in every performance measure, except for interval score on $\alpha = 0.05$ level. Additionally, the linear quantile regression models perform worse than their respective D-vine regression models, `DVReg2_nonpar_orig` and `DVReg2_nonpar_ln`, according to most performance measures. When it comes to the BIC values all D-vine regression models are lagging behind (especially the nonparametric ones due to their high complexity).
The gamma regression models show worse performance than the lognormal regression models in every performance measure, except in the training and testing error. Additionally, we can see that the interaction terms do not improve much the lognormal and gamma regression models with main effects.
Finally, in the following analysis we consider the models `LogLM2_main_red`, `GamReg2_main_red`, `DVReg2_nonpar_orig` and the respective linear quantile regression model `LQReg2_orig`, as well as `DVReg2_nonpar_ln` and the respective linear quantile regression model `LQReg2_ln`. Although the linear quantile regression models do not have better performance values than their respective D-vine regression models, we consider them in our analysis as they contain the same covariates as the D-vine models, therefore we can easily compare them and analyze the different behaviour of D-vine quantile regression and linear quantile regression on Bad Driver Data.

LogLM2_main_red	GamReg2_main_red	DVReg2_nonpar_orig	DVReg2_nonpar_ln
pol_bonus	pol_bonus	vh_cyl	vh_weight
poly(drv_age_lic1,3)	poly(drv_age1,4)	pol_usage2	drv_age_lic1
poly(vh_cyl,2)	vh_cyl	gender2	vh_cyl
poly(vh_din,2)	poly(vh_din,3)	pol_coverage2	vh_speed
poly(vh_weight, 5)	vh_weight	claim_ind	vh_sale_end
pol_coverage2	vh_fuel	drv_age1	claim_ind
vh_fuel	class_make2	vh_sale_end	pol_coverage2
claim_ind		vh_fuel	vh_fuel
class_make2		vh_weight	
		pol_payd	

Table 44: Covariates included in each of the models `LogLM2_main_red`, `GamReg2_main_red`, `DVReg2_nonpar_orig` and `DVReg2_nonpar_ln`. The model `LQReg2_orig` has the same covariates as `DVReg2_nonpar_orig` and the model `LQReg2_ln` has the same covariates as `DVReg2_nonpar_ln`.

Based on Table 44, the continuous covariates `vh_cyl` and `vh_weight` and the discrete covariate `vh_fuel`, are included in all six models. In Figure 34 we present the histograms of the fitted values on original *standardized claims* scale on Bad Driver training data set for the models `LogLM2_main_red`, `GamReg2_main_red`, `DVReg2_nonpar_orig`, `LQReg2_orig`, `DVReg2_nonpar_ln` and `LQReg2_ln`. The gamma regression model predicts better the tail, but it overestimates the small values of *standardized claims*. The other models underestimate the tails, but predict well the small values of *standardized claims*. Additionally, the gamma regression model has larger range of the fitted values of *standardized claims* compared to the other models. However, all of the histograms indicate that the models cannot predict the response *standardized claims* very precisely, given our data.

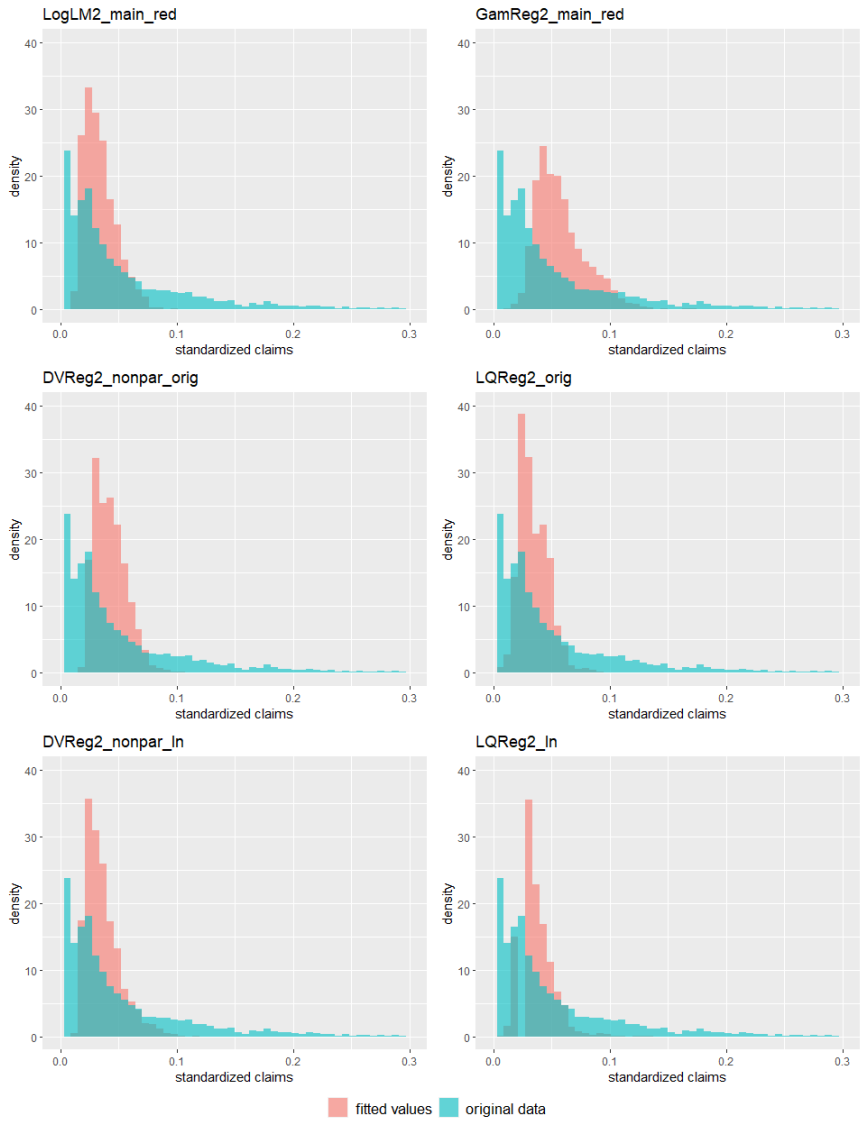

Figure 34: Histograms of the fitted values of *standardized claims* of the models `LogLM2_main_red`, `GamReg2_main_red`, `DVReg2_nonpar_orig`, `LQReg2_orig`, `DVReg2_nonpar_ln` and `LQReg2_ln` on the Bad Driver training data set.

Similarly, we present the histograms of the predicted values on the Bad Driver test data set in Figure 35. Whe testing the models on new data we can see that they perform similarly as on the training data set. In particular, the model `GamReg2_main_red` predicts the tails the best, but severely overestimates the small values of the response. In comparison, the model `DVReg2_nonpar_ln` predicts well part of the tail, while predicting the small values as precisely as all other models (except for the `GamReg2_main_red`). Once again, given these histograms we can see that the prediction power of the models is not very satisfactory, especially not for the big values of *standardized claims*.

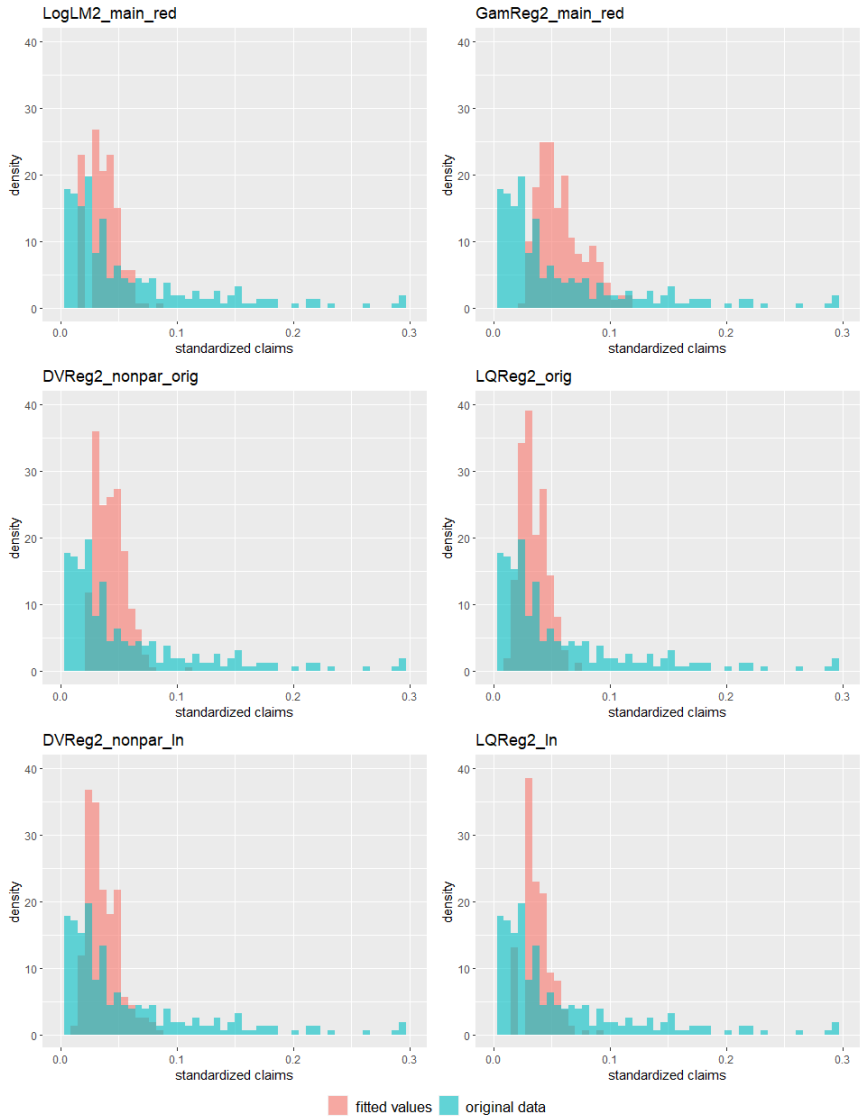

Figure 35: Histograms of the predicted values of *standardized claims* of the models `LogLM2_main_red`, `GamReg2_main_red`, `DVReg2_nonpar_orig`, `LQReg2_orig`, `DVReg2_nonpar_ln` and `LQReg2_ln` on the Bad Driver test data set.

Now we would like to investigate the models' prediction intervals on the Bad Driver test data set. In Figure 36 we present the 90% prediction intervals of the observations in the test data set for each of these six models.

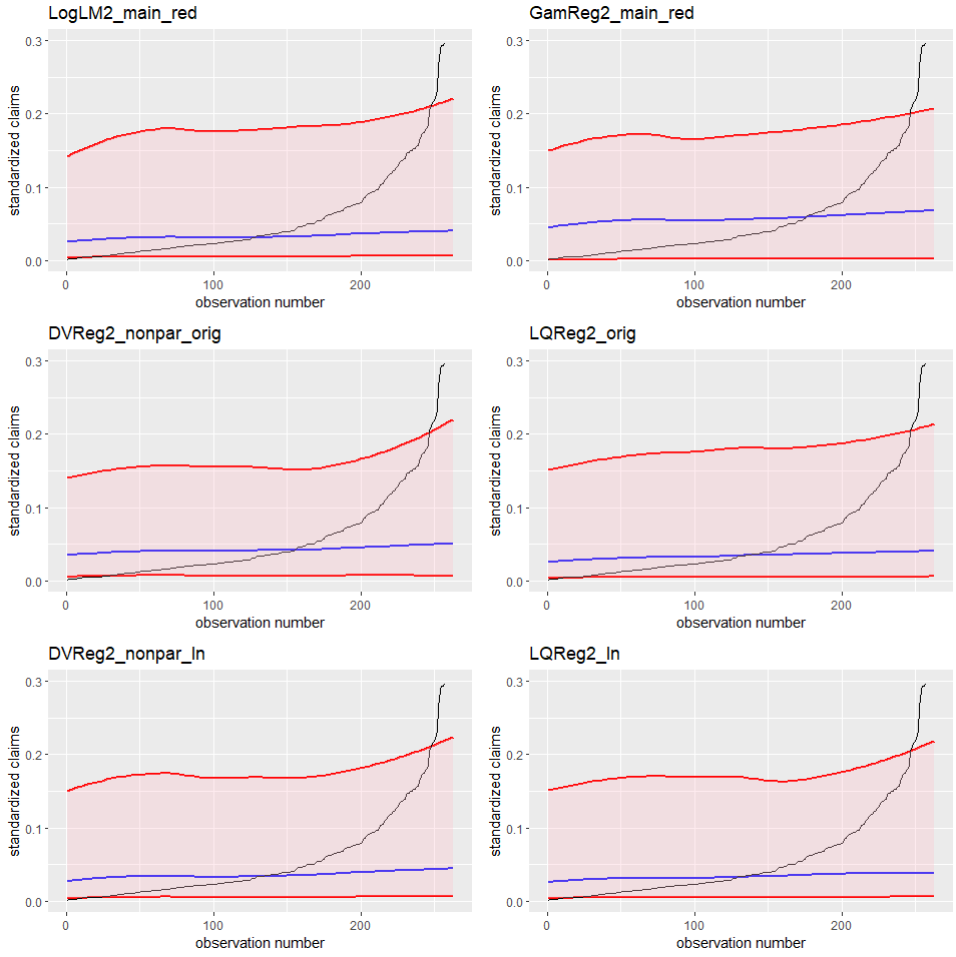

Figure 36: 90% prediction intervals of *standardized claims* on the Bad Driver test data set for the models `LogLM2_main_red`, `GamReg2_main_red`, `DVReg2_nonpar_orig`, `LQReg2_orig`, `DVReg2_nonpar_ln` and `LQReg2_ln`. The black line denotes the original values of *standardized claims*, the blue line denotes the smoothed line of the predicted values and the red area denotes the prediction interval, where the red lines are the smoothed limits of the prediction interval. The observations are sorted in increasing order based on the original values of *standardized claims*.

The observations in Figure 36 are ordered based on their original values of the response *standardized claims*, which on the plots are presented with the black line. The blue line denotes the predicted values of the respective observations and the red lines denote the lower and upper limit of the prediction interval. Most of the original values of *standardized claims* belong in the 90% prediction intervals of the models, where the exception are the values of *standardized claims* larger than 0.2. The prediction intervals of all six models look similar, where only the prediction interval of the model `DVReg2_nonpar_orig` looks narrower than the others. The predicted values of *standardized claims* in the gamma model `GamReg2_main_red` are larger than the other five models. This results in the

same conclusions we got from Figure 35, which are that the models `LogLM2_main_red`, `DVReg2_par_orig`, `LQReg2_orig`, `DVReg2_nonpar_ln` and `LQReg2_ln` underestimate the bigger values of *standardized claims*, whereas the gamma regression model overestimates the smaller values of *standardized claims*.

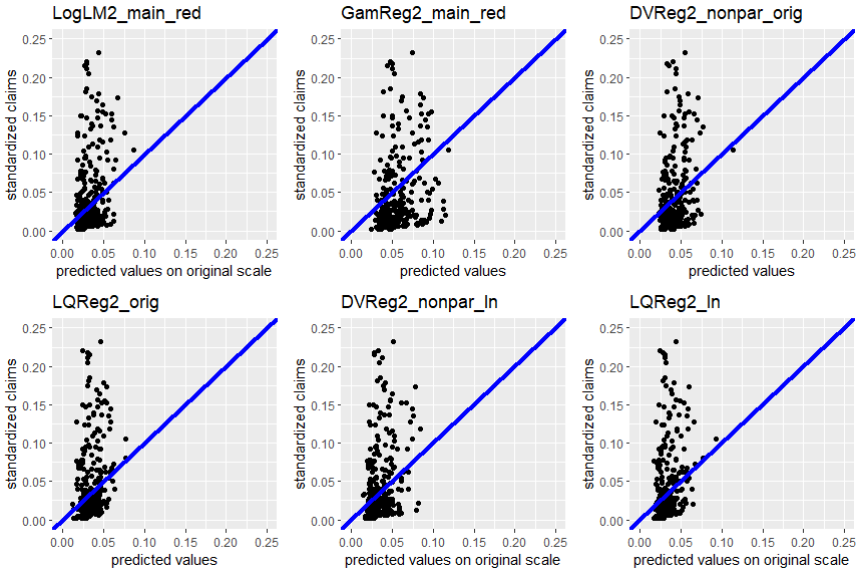

Figure 37: Plots of the fitted values on original *standardized claims* scale against the original values of *standardized claims* on the Bad Driver test data set for the models `LogLM2_main_red`, `GamReg2_main_red`, `DVReg2_nonpar_orig`, `LQReg2_orig`, `DVReg2_nonpar_ln` and `LQReg2_ln`. The blue line denotes the 45° line.

To confirm the statement of underestimation of *standardized claims* in all models except the gamma regression model, in Figure 37 we can see the plots of the predicted values of *standardized claims* against the original values on the test data set. Intuitevely, we would expect the black points on the plot to be scattered around the blue 45° line, which is not the case for any of the models. For the gamma regression model, the bigger values of *standardized claims* are predicted better than the other five models and the predicted values have larger range. For the other models, the range of the predicted values of *standardized claims* is much smaller than the range of the original data and they heavily underestimate the big values of *standardized claims*.
Finally, as the models `DVReg2_ nonpar_orig` and `LQReg2_orig`, as well as the models `DVReg2_nonpar_ln` and `LQReg2_ln` have the same covariates, we would like to investigate the marginal effects of the continuous covariates in the models on the predicted quantiles on the Bad Driver training data set. The marginal effect of a continuous covariate is presented by a plot of the continuous covariate x_i against the fitted conditional quantiles $\hat{q}^i_\alpha, i = 1, \ldots, n_{tr}$, where all other covariates are set to their observed value. The marginal effect plots of the models for three different quantile levels 0.1, 0.5 and 0.9. are presented in Figures 38 and 39.

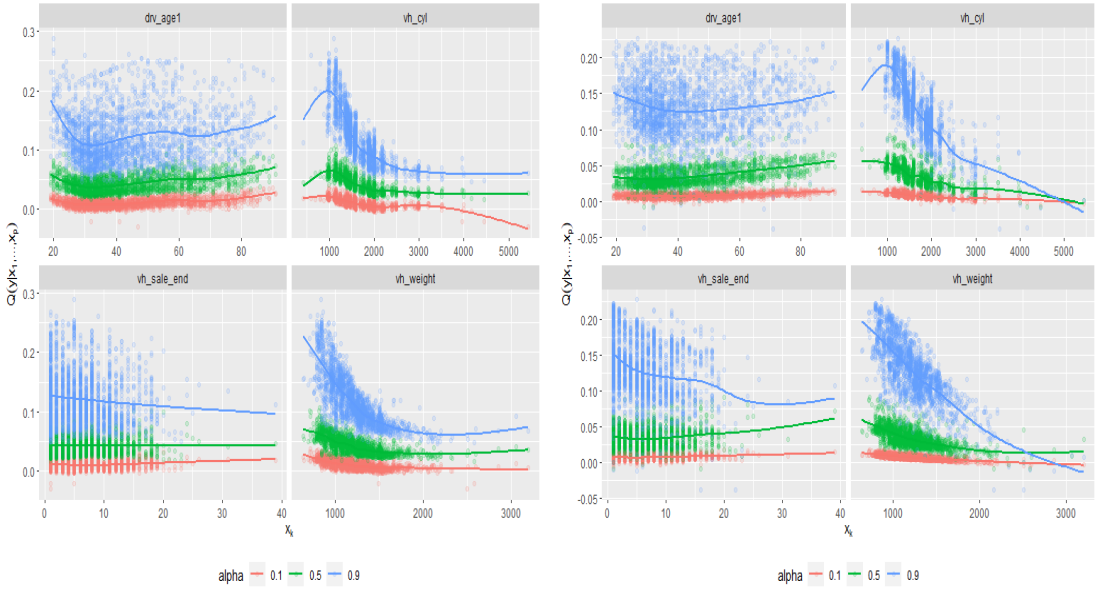

Figure 38: Marginal effect plots for the continuous covariates on the predicted quantiles on Bad Driver training data set for three different quantile levels. *Left*:`DVReg2_nonpar_orig`, *right*: `LQReg2_orig`.

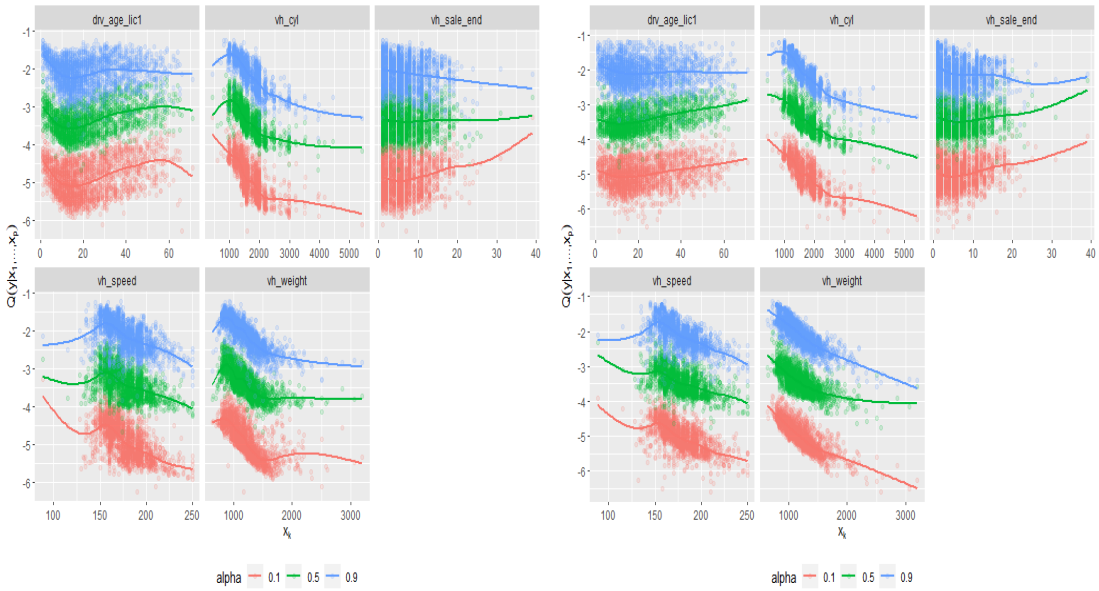

Figure 39: Marginal effect plots for the continuous covariates on the predicted quantiles on Bad Driver training data set for three different quantile levels.*Left*:`DVReg2_nonpar_ln`,*right*:`LQReg2_ln`.

In Figure 38, on the marginal effect plots of `vh_cyl` and `vh_weight` of the model `LQReg2_orig` we can spot quantile crossing, which cannot appear in a D-vine regression model. We can also notice that some of the predicted quantiles in the `LQReg2_orig` model have negative values, which is a disadvantage of using the response variable on original scale. The

effects of the continuous covariates in the model `LQReg2_orig` however are much more linear than the effects of the continuous covariates in the model `DVReg2_par_orig`, with only the covariate `vh_cyl` indicating strong nonlinearity. Similarly, the marginal effect plots of the model `LQReg2_ln` in Figure 38 look more linear than the model `DVReg2_nonpar_ln`, however, most of the continuous covariates in these two models indicate nonlinear effects. In this case no quantile crossing is present on the plots. This nonlinearity in the marginal effect plots of the D-vine regression models may be one the reasons why D-vine regression models perform better than the linear quantile regression models.
Finally, we can conclude that although the D-vine regression models are more complex than the lognormal, gamma and linear quantile regression models, they give satisfactory results as a modelling method in predicting *standardized claims* in Bad Driver Data compared to the other regression methods.

6 Conclusion

This thesis is concerned with four different statistical methods for modelling third party liability motor insurance claims. The lognormal and gamma regression are standard approach in modelling claims severity. We also studied two quantile regression methods, the linear quantile regression and D-vine quantile regression, which are more robust methods for outliers and provide an opportunity to investigate the tails.
On both Good and Bad Driver Data, D-vine quantile regression results in highest log likelihood among all four regression methods. However, when it comes to AIC and BIC criteria, this regression method lags behind, due to its higher complexity compared to the other regression methods. The fitted D-vine quantile regression models also show low training error, test error and interval scores in comparison to the lognormal, gamma and linear quantile regression models.
Compared to the lognormal regression models, gamma regression models have better values only for the training and test errors. The advantage of the gamma models is that the range of predicted values of the original response is larger, indicating that they predict the larger values of the response better, compared to the other models which were concentrated more in a smaller window of predicted values of the response.
In both data sets, linear quantile regression models perform worse in almost every performance measure than their respective D-vine quantile regression models. That being said, D-vine quantile regression shows larger predictive power than linear quantile regression on the given data sets.
The parametric D-vine regression is easier to interpret and less complex than the nonparametric D-vine regression. However, the parametric D-vine regression is less flexible in capturing nonparametric bivariate copula structures and requires more work in the preprocessing step of the estimation of the marginals. It also does not allow for ordinal covariates in the model, compared to the nonparametric D-vine regression. In this thesis, the nonparametric D-vine quantile regression model performs better than the parametric one for Bad Driver Data. An important finding is that for both D-vine quantile regression and linear quantile regression, the models which use the transformed response as a response variable show better performance measures than the models which use the response on original scale.
Before we conclude the work of this thesis, it is important to keep in mind that our fitted models did not show big predictive power for *standardized claims*, given the data sets. This can directly be seen by the R^2_{adj} values for the lognormal models on the both Good and Bad Driver Data, which are smaller than 15%. Having that said, we cannot make any strong conclusions about D-vine regression performance compared to the other methods. However, considering the fact that D-vine regression requires less data analysis prior modelling and automatically includes only the important covariates in the model, we can say that this regression method may be beneficial for further application in insurance risk modelling. In particular, it provides a good overview of dependence structures between the important covariates and the response, and yields an opportunity to investigate the tails of the response.

Bibliography

Ahsanullah, M. (2017). *Characterizations of Univariate Continuous Distributions*. Springer.

Azzalini, A. (2005). The skew-normal distribution and related multivariate families. *Scandinavian journal of statistics*, *32*(2), 159–188.

Basso, R. M., Lachos, V. H., Cabral, C. R. B., & Ghosh, P. (2010). Robust mixture modeling based on scale mixtures of skew-normal distributions. *Computational Statistics & Data Analysis*, *54*(12), 2926–2941.

Bernard, C., & Czado, C. (2015). Conditional quantiles and tail dependence. *Journal of Multivariate Analysis*, *138*, 104–126.

Cabral, C. R. B., Lachos, V. H., & Prates, M. O. (2012). Multivariate mixture modeling using skew-normal independent distributions. *Computational Statistics & Data Analysis*, *56*(1), 126–142.

Chang, B., & Joe, H. (2019). Prediction based on conditional distributions of vine copulas. *Computational Statistics & Data Analysis*, *139*, 45–63.

Czado, C. (2019). *Analyzing Dependent Data With Vine Copulas: A Practical Guide With R*. Springer.

Czado, C., & Schmidt, T. (2011). *Mathematische Statistik*. Springer-Verlag.

De Jong, P., & Heller, G. Z. (2008). *Generalized Linear Models for Insurance Data*. Cambridge University Press.

Fahrmeir, L., Kneib, T., Lang, S., & Marx, B. (2013). *Regression: Models, Methods and Applications*. Berlin: Springer-Verlag.

Fahrmeir, L., & Turz, G. (2001). *Multivariate Statistical Modelling Based on Generalized Linear Models* (Second edition). Springer, New York.

Geenens, G. (2014). Probit transformation for kernel density estimation on the unit interval. *Journal of the American Statistical Association*, *109*(505), 346–358.

Gneiting, T., & Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association*, *102*(477), 359–378.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (Vol. 2). Springer.

Jiang, J. (2010). *Large Sample Techniques for Statistics*. Springer.

Joe, H. (2014). *Dependence Modeling with Copulas*. CRC Press.

Koenker, R. (2005). *Quantile Regression*. Cambridge University Press.

Koenker, R., & Bassett Jr, G. (1978). Regression quantiles. *Econometrica: journal of the Econometric Society*, 33–50.

Kraus, D., & Czado, C. (2017). D-vine copula based quantile regression. *Computational Statistics & Data Analysis*, *110*, 1–18.

Lefebvre, M. (2006). *Applied Probability and Statistics*. Springer.

McCullagh, P., & Nelder, J. A. (1989). *Generalized Linear Models*. Chapman & Hall.

Nagler, T. (2014). *Kernel methods for vine copula estimation* (Master's thesis). Technische Universität München. Garching b. Müunchen.

Nagler, T. (2022). *vinereg: D-Vine Quantile Regression*. R package version 0.8.3.

Nagler, T., & Vatter, T. (2022). *kde1d: Univariate Kernel Density Estimation*. R package version 1.0.4.

Nelder, J. A., & Wedderburn, R. W. (1972). Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)*, *135*(3), 370–384.

Nelsen, R. B. (2006). *An Introduction to Copulas*. New York: Springer.

Olive, D. J. (2017). *Linear Regression*. Springer.

Panagiotelis, A., Czado, C., & Joe, H. (2012). Pair copula constructions for multivariate discrete data. *Journal of the American Statistical Association*, *107*(499), 1063–1072.

Pewsey, A. (2000). Problems of inference for Azzalini's skewnormal distribution. *Journal of applied statistics*, *27*(7), 859–870.

Prates, M. O., Lachos, V. H., & Cabral, C. R. B. (2013). mixsmsn: Fitting finite mixture of scale mixture of skew-normal distributions. *Journal of Statistical Software*, *54*, 1–20.

Sklar, A. (1959). Fonctions de répartition à n dimensions et leurs marges. *Publications de l'Institut de Statistique de l'Université de Paris*, *8*, 229–231.

Tepegjozova, M. (2019). *D- and C-vine quantile regression for large data sets* (Master's thesis). Technische Universität München. Garching b. Müunchen.

Tepegjozova, M., Zhou, J., Claeskens, G., & Czado, C. (2022). Nonparametric C- and D-vine based quantile regression. *Dependence Modeling*, *10*(1), 1–21.

Venables, W. N., & Ripley, B. D. (2002). *Modern Applied Statistics with S* (Fourth edition). Springer, New York.

Wasef Hattab, M. (2016). A derivation of prediction intervals for gamma regression. *Journal of Statistical Computation and Simulation*, *86*(17), 3512–3526.