Chapter 13
Appendix B: Data and Examples

This appendix presents and describes the data sets used.

13.1 French Motor Third Party Liability Data

We consider a French motor third party liability (MTPL) claims data set. This data set is available through the R library CASdatasets[1] being hosted by Dutang–Charpentier [113]. The specific data sets chosen from CASdatasets are called FreMTPL2freq and FreMTPL2sev, the former contains the insurance policy and claim frequency information and the latter the corresponding claim severity information.[2]

Before we can work with this data set we perform data cleaning. It has been pointed out by Loser [259] that the claim counts on the insurance policies with policy IDs $\leq$ 24500 in FreMTPL2freq do not seem to be correct because these claims do not have claim severity counterparts in FreMTPL2sev. For this reason we work with the claim counts extracted from the latter file. In Listing 13.1 we give the code used for data cleaning.[3] In this code we merge FreMTPL2freq with the aggregated severities on each insurance policy and the corresponding claim counts are received from FreMTPL2sev, this is done on lines 2–11 of Listing 13.1. A

[1] CASdatasets website: http://cas.uqam.ca/.

[2] We use CASdatasets version 1.0–8 which has been packaged on 2018-05-20. This version uses for the 22 French regions the labels R11, ..., R94. In later versions of CASdatasets these labels have been replaced by the region names, in this transformation the labels R31 (Nord-Pas-de-Calais) and R41 (Lorraine) have been merged to one region called Nord-Pas-de-Calais. We believe that this is an error and therefore prefer to work with an older version of CASdatasets. This older version can be downloaded in R with library(OpenML), library(farff), freMTPL2freq <- getOMLDataSet(data.id = 41214)$data

[3] The code in Listing 13.1 is a modified version of the R code provided by Loser [259].

M. V. Wüthrich, M. Merz, *Statistical Foundations of Actuarial Learning and its Applications*, Springer Actuarial, https://doi.org/10.1007/978-3-031-12409-9_13

further inspection of the data indicates that policies with more than 5 claims may be data error because they all seem to belong to the same driver (and they have very short exposures).[4] For this reason we drop these records on line 12. On line 13 we censor exposures at one accounting year (since these policies are active within one calendar year). Finally, on lines 15–16 we re-level the `VehBrands`.[5] All subsequent analysis is based on this cleaned data set.

Listing 13.1 Data cleaning applied to the French MTPL data set

```
 1  #
 2  data(freMTPL2freq)
 3  dat <- freMTPL2freq[, -2]
 4  dat$VehGas <- factor(dat$VehGas)
 5  data(freMTPL2sev)
 6  sev <- freMTPL2sev
 7  sev$ClaimNb <- 1
 8  dat0 <- aggregate(sev, by=list(IDpol=sev$IDpol), FUN = sum)[c(1,3:4)]
 9  names(dat0)[2] <- "ClaimTotal"
10  dat <- merge(x=dat, y=dat0, by="IDpol", all.x=TRUE)
11  dat[is.na(dat)] <- 0
12  dat <- dat[which(dat$ClaimNb <=5),]
13  dat$Exposure <- pmin(dat$Exposure, 1)
14  sev <- sev[which(sev$IDpol %in% dat$IDpol), c(1,2)]
15  dat$VehBrand <- factor(dat$VehBrand, levels=c("B1","B2","B3","B4","B5","B6",
16                                               "B10","B11","B12","B13","B14"))
```

Listing 13.2 Excerpt of the French MTPL data set

```
 1  'data.frame':   678007 obs. of  13 variables:
 2   $ IDpol      : num  1 3 5 10 11 13 15 17 18 21 ...
 3   $ Exposure   : num  0.1 0.77 0.75 0.09 0.84 0.52 0.45 0.27 0.71 0.15 ...
 4   $ Area       : Factor w/ 6 levels "A","B","C","D",..: 4 4 2 2 2 5 5 3 3 2 ...
 5   $ VehPower   : int  5 5 6 7 7 6 6 7 7 7 ...
 6   $ VehAge     : int  0 0 2 0 0 2 2 0 0 0 ...
 7   $ DrivAge    : int  55 55 52 46 46 38 38 33 33 41 ...
 8   $ BonusMalus: int  50 50 50 50 50 50 50 68 68 50 ...
 9   $ VehBrand   : Factor w/ 11 levels "B1","B2","B3",..: 9 9 9 9 9 9 9 9 9 9 ...
10   $ VehGas     : Factor w/ 2 levels "Diesel","Regular": 2 2 1 1 1 2 2 1 1 1 ...
11   $ Density    : int  1217 1217 54 76 76 3003 3003 137 137 60 ...
12   $ Region     : Factor w/ 22 levels "R11","R21","R22",..: 18 18 3 15 15 8 8 ...
13   $ ClaimTotal: num  0 0 0 0 0 0 0 0 0 0 ...
14   $ ClaimNb    : num  0 0 0 0 0 0 0 0 0 0 ...
15  ####
16  'data.frame':   26383 obs. of  2 variables:
17   $ IDpol       : int  1552 1010996 4024277 4007252 4046424 4073956 4012173 ...
18   $ ClaimAmount: num  995 1128 1851 1204 1204 ...
```

Listing 13.2 gives an excerpt of the cleaned French MTPL data set, lines 2–14 give the insurance policy and claim counts information, and lines 17–18

[4] Short exposure policies may also belong to a commercial car rental company.

[5] The data set `FreMTPLfreq` of `CASdatasets` is a subset of `FreMTPL2freq` with slightly changed feature components, for instance, the former data set contains car brand names in a more aggregated version than the latter, see Table 13.2, below.

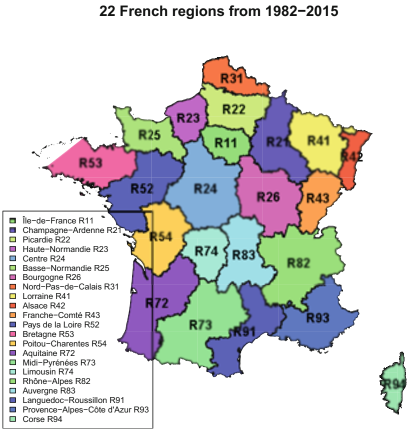
display the individual claim amounts. We have 9 feature components on lines 4–12 (1 component is binary, 3 components are categorical, and 5 components are continuous), an exposure variable on line 3, and claim information on lines 13–14 and 18. In total we have 26’383 claims on 678’007 insurance policies.

We start by giving a descriptive analysis of the data, this closely follows Noll et al. [287]. We have the following insurance policy information:

1. `IDpol`: policy number (unique identifier);
2. `Exposure`: total exposure in yearly units (years-at-risk) and within (0, 1];
3. `Area`: area code (categorical, ordinal with 6 levels);
4. `VehPower`: power of the car (continuous);
5. `VehAge`: age of the car in years;
6. `DrivAge`: age of the (most common) driver in years;
7. `BonusMalus`: bonus-malus level between 50 and 230 (with entrance level 100);
8. `VehBrand`: car brand (categorical, nominal with 11 levels), see also Table 13.2;
9. `VehGas`: diesel or regular fuel car (binary);
10. `Density`: density of population per km^2 at the location of the living place of the driver;
11. `Region`: regions in France (prior to 2016), see also Fig. 13.1 (categorical).

We start by describing the `Exposure`. The `Exposure` measures the duration of an insurance policy in yearly units; sometimes it is also called *years-at-risk*. The shortest exposure in our data set is 0.0027 which corresponds to 1 day, and the longest exposure is 1 which corresponds to 1 year. Figure 13.2 (lhs, middle) shows a histogram and a boxplot of these exposures. In view of the histogram we conclude that roughly 1/4 of all policies have a full exposure of 1 calendar year, and all other policies are only partly exposed during the calendar year. From a practical insurance point of view this high ratio of partly exposed policies seems rather

Fig. 13.1 The 22 regions in France between 1982 and 2015



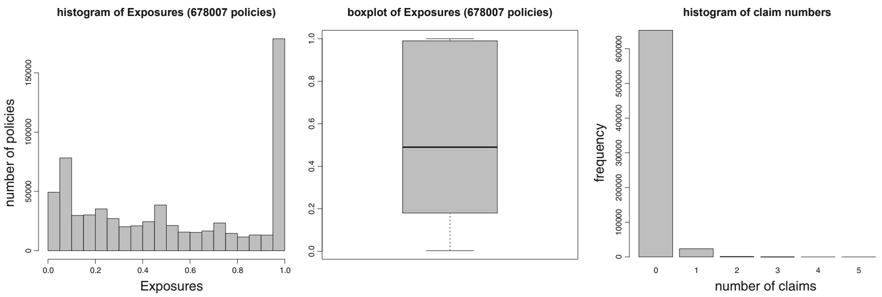


Fig. 13.2 (lhs) Histogram of `Exposure`, (middle) boxplot of `Exposure`, (rhs) number of observed claims `ClaimNb` of the French MTPL data

Table 13.1 Split of the portfolio w.r.t. the number of claims

Number of claims	0	1	2	3	4	5
Number of policies	653’069	23’571	1’298	62	5	2
Total exposure	341’090	16’315	909	42	2	1

unusual. A further inspection of the data indicates that policy renewals during the year account for two separate records in the data set. Of course, such split policies should be merged to one yearly policy. Unfortunately, we do not have the necessary information to perform this merger, therefore, we need to work with the data as it is. In Table 13.1 and Fig. 13.2 (rhs) we split the portfolio w.r.t. the number of claims. On 653’069 insurance policies (amounting to a total exposure of 341’090 years-at-risk) we do not have any claim, and on the remaining 24’938 policies (17’269 years-at-risk) we have at least one claim. The overall portfolio claim frequency (w.r.t. `Exposure`) is $\overline{\lambda} = 7.35\%$.

We study the split of this overall frequency $\overline{\lambda} = 7.35\%$ across the different feature levels. This empirical analysis is crucial for the model choice in regression modeling.[6] For the empirical analysis we provide 3 different types of graphs for each feature component (where applicable), these are given in Figs. 13.3, 13.4, 13.5, 13.6, 13.7, 13.8, 13.9, 13.10, and 13.11. The first graph (lhs) gives the split of the total exposure to the different feature levels, the second graph (middle) gives the average feature value in each French region (green meaning low and red meaning high),[7] and the third graph (rhs) gives the observed average frequency per feature level. This observed frequency is obtained by dividing the total number of claims by the total exposure per feature level. The frequencies are complemented by confidence bounds of two standard deviations (shaded area). These confidence bounds correspond to twice the estimated standard deviations. The standard deviations are estimated under

[6] The empirical analysis in these notes differs from Noll et al. [287] because data cleaning has been done differently here, we refer to Listing 13.1.

[7] We acknowledge the use of UNESCO (1987) database through UNEP/GRID-Geneva for the French map.

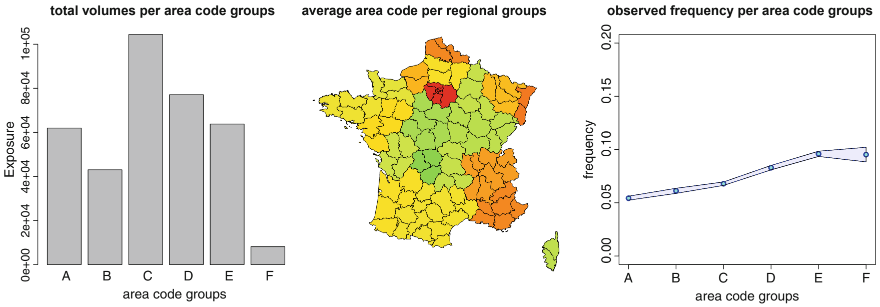


Fig. 13.3 (lhs) Histogram of exposures per `Area` code, (middle) average `Area` code per `Region`, we map $(A, \ldots, F) \mapsto (1, \ldots, 6)$, (rhs) observed frequency per `Area` code

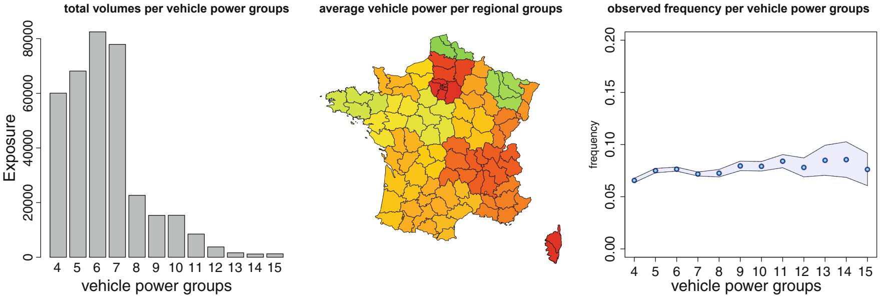


Fig. 13.4 (lhs) Histogram of exposures per `VehPower`, (middle) average `VehPower` per `Region`, (rhs) observed frequency per `VehPower`

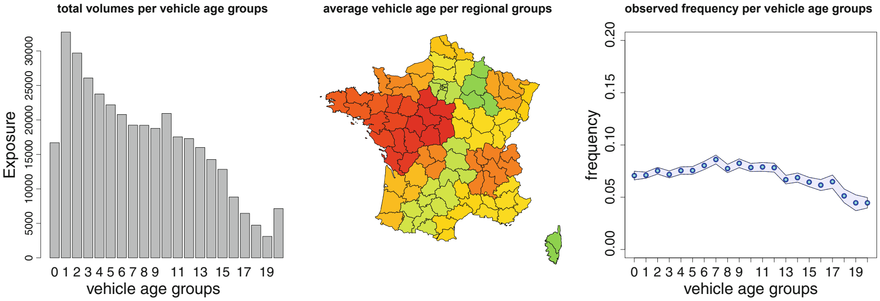


Fig. 13.5 (lhs) Histogram of exposures per `VehAge` (censored at 20), (middle) average `VehAge` per `Region`, (rhs) observed frequency per `VehAge`

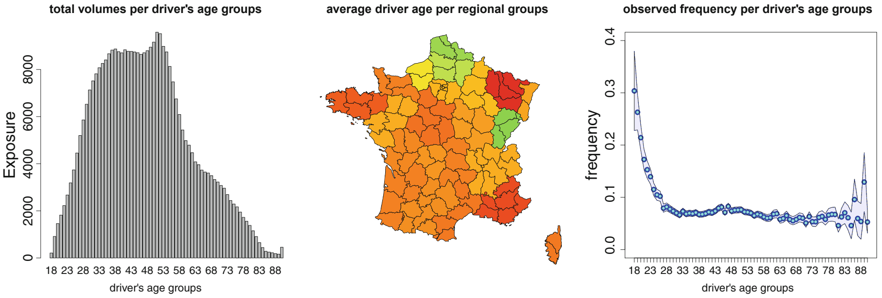


Fig. 13.6 (lhs) Histogram of exposures per `DrivAge` (censored at 90), (middle) average `DrivAge` per `Region`, (rhs) observed frequency per `DrivAge` (y-scale is different compared to the other frequency plots)

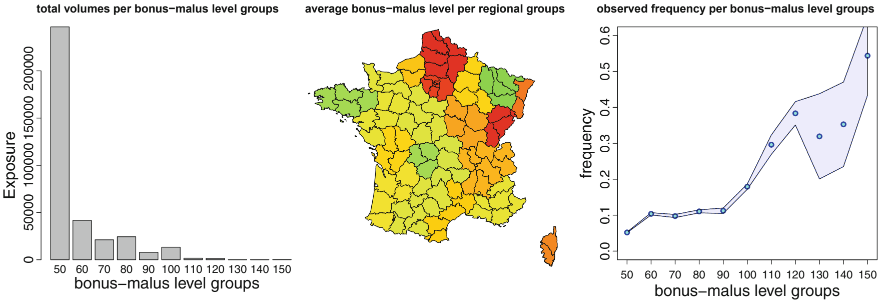


Fig. 13.7 (lhs) Histogram of exposures per `BonusMalus` level (censored at 150), (middle) average `BonusMalus` level per `Region`, (rhs) observed frequency per `BonusMalus` level (y-scale is different compared to the other frequency plots)

a Poisson assumption, thus, they are obtained by $\pm 2\sqrt{\overline{\lambda}_k/\texttt{Exposure}_k}$, where $\overline{\lambda}_k$ is the observed frequency and $\texttt{Exposure}_k$ is the total exposure for a given feature level k. We note that in all frequency plots the y-axis ranges from 0% to 20%, except in the `BonusMalus` plot where the maximum is set to 60%, and the `DrivAge` plot where the maximum is set to 40%. From these plots we conclude that some levels have only a small underlying `Exposure`; `BonusMalus` leads to the highest variability in frequencies followed by `DrivAge`; and there is quite some heterogeneity.

Table 13.2 gives the assignment of the different `VehBrand` levels to car brands. This list has been compiled from the two data sets `FreMTPLfreq` and `FreMTPL2freq` contained in the R package `CASdatasets` [113], see Footnote 5.

Next, we analyze collinearity between the feature components. For this we calculate Pearson's correlation and Spearman's Rho for the continuous feature components, see Table 13.3. In general, these correlations are low, except for `DrivAge` vs. `BonusMalus`. Of course, the latter is very sensible because a `BonusMalus`

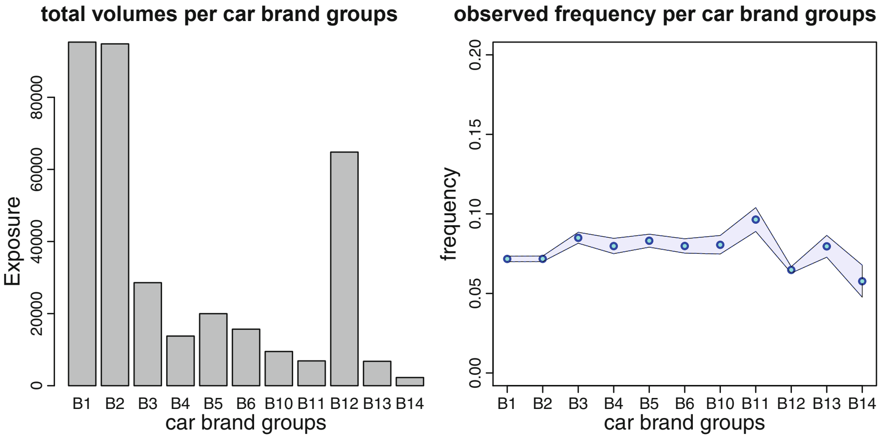


Fig. 13.8 (lhs) Histogram of exposures per `VehBrand`, (rhs) observed frequency per `VehBrand`; for `VehBrand` assignment we refer to Table 13.2

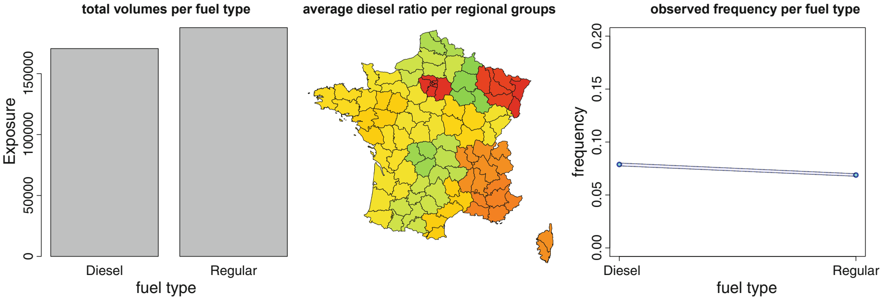


Fig. 13.9 (lhs) Histogram of exposures per `VehGas`, (middle) average `VehGas` per `Region` (diesel is green and regular red), (rhs) observed frequency per `VehGas`

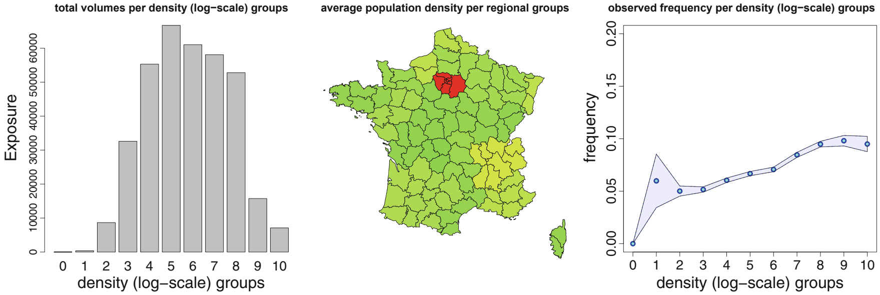


Fig. 13.10 (lhs) Histogram of exposures per population `Density` (on log-scale), (middle) average population `Density` per `Region`, (rhs) observed frequency per population `Density`; in general, we always consider `Density` on the log-scale

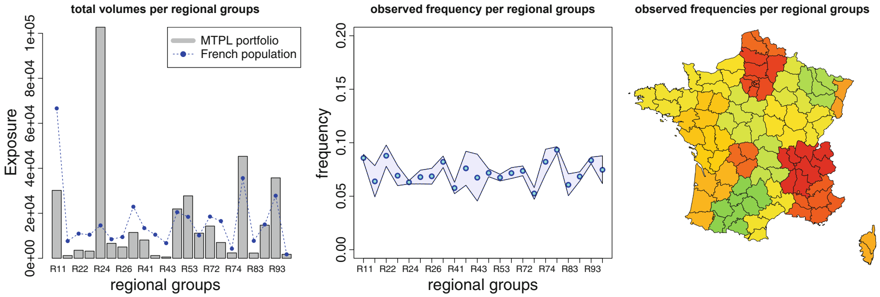


Fig. 13.11 (lhs) Histogram of exposures `Exposure`, and (middle, rhs) observed claim frequencies per `Region` in France (prior to 2016)

Table 13.2 `VehBrand` assignment

Renault, Nissan and Citroën	B1 / B2
Volkswagen, Audi, Skoda and Seat	B3
Opel, General Motors and Ford	B4 / B5
Fiat	B6
Mercedes, Chrysler and BMW	B10 / B11
Japanese (except Nissan) and Korean cars	B12
Other cars	B13 / B14

Table 13.3 Correlations in feature components: top-right shows Pearson's correlation; bottom-left shows Spearman's Rho; `Density` is considered on the log-scale; significant correlations are boldface

	`VehPower`	`VehAge`	`DrivAge`	`BonusMalus`	`Density`
`VehPower`		−0.01	0.03	−0.08	0.01
`VehAge`	0.00		−0.06	0.08	−0.10
`DrivAge`	0.04	−0.08		**−0.48**	−0.05
`BonusMalus`	−0.07	0.08	**−0.57**		0.13
`Density`	−0.01	−0.10	−0.05	0.14	

level below 100 needs a certain number of driving years without claims. We give the corresponding boxplot in Fig. 13.12 (lhs) which confirms this negative correlation. Figure 13.12 (rhs) gives the boxplot of log-`Density` vs. `Area` code. From this plot we conclude that the area code has likely been set w.r.t. the log-`Density`. For our regression models this means that we can drop the area code information, and we should only work with `Density`. Nevertheless, we will use the area code to show what happens in case of collinear feature components, i.e., if we replace $(A, \ldots, F) \mapsto (1, \ldots, 6)$.

Figure 13.13 illustrates each continuous feature component w.r.t. the different `VehBrands`. Vehicle brands `B10` and `B11` (Mercedes, Chrysler and BMW) have more `VehPower` than other cars, `B10` being more likely a diesel car, and vehicle brand `B12` (Japanese and Korean cars) has comparably new cars in more densely populated French regions.

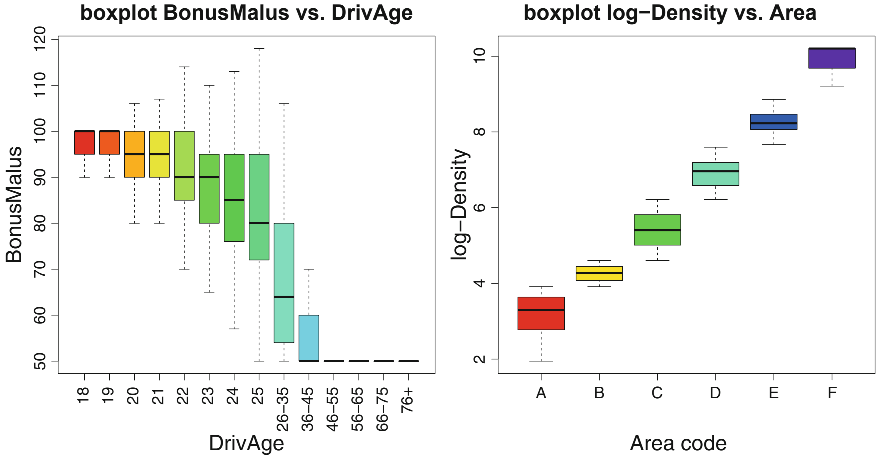


Fig. 13.12 Boxplots (lhs) `BonusMalus` vs. `DrivAge`, (rhs) log-`Density` vs. `Area` code; these plots are inspired by Fig. 2 in Lorentzen–Mayer [258]

More formally, the strength of dependence between categorical variables can be measured by Cramér's V. Cramér's V is based on the χ^2-test of independence on contingency tables. We briefly explain this. Assume we have two-dimensional categorical features $\boldsymbol{x} = (x_1, x_2) \in \mathcal{X}$ having m_1 and m_2 levels, respectively. Let $p_{\boldsymbol{x}}$ describe the probability on $\mathcal{X}$ that a randomly chosen insurance policy takes feature $\boldsymbol{x}$, and let p_{x_1} and p_{x_2} be the marginal distributions of $p_{\boldsymbol{x}}$. If the two components of $\boldsymbol{x}$ are independent with these two marginals, then we have special (independence) distribution

$$\pi_{\boldsymbol{x}} = p_{x_1} p_{x_2} \qquad \text{for all } \boldsymbol{x} = (x_1, x_2) \in \mathcal{X}.$$

The χ^2-test for independence now analyzes $p_{\boldsymbol{x}}$ vs. $\pi_{\boldsymbol{x}}$. Assume we have n observations. Denote by $n_{\boldsymbol{x}} = n_{x_1,x_2}$ the number of instances that have feature $\boldsymbol{x} = (x_1, x_2)$, and let $n_{x_1,\cdot}$ and $n_{\cdot,x_2}$ be the corresponding marginal observations. The χ^2-test statistics is given by

$$\chi^2 = \sum_{\boldsymbol{x}=(x_1,x_2)\in\mathcal{X}} \frac{\left(n_{\boldsymbol{x}} - \frac{n_{x_1,\cdot}\, n_{\cdot,x_2}}{n}\right)^2}{\frac{n_{x_1,\cdot}\, n_{\cdot,x_2}}{n}}.$$

Under the null hypothesis of having independence between the components of $\boldsymbol{x}$, the test statistics χ^2 converges in distribution to a χ^2-distribution with $(m_1 m_2 - 1)$ degrees of freedom if we let the number of independently drawn instances go to infinity. Seven different proofs of this statement are given in Benhamou–Melot [30].

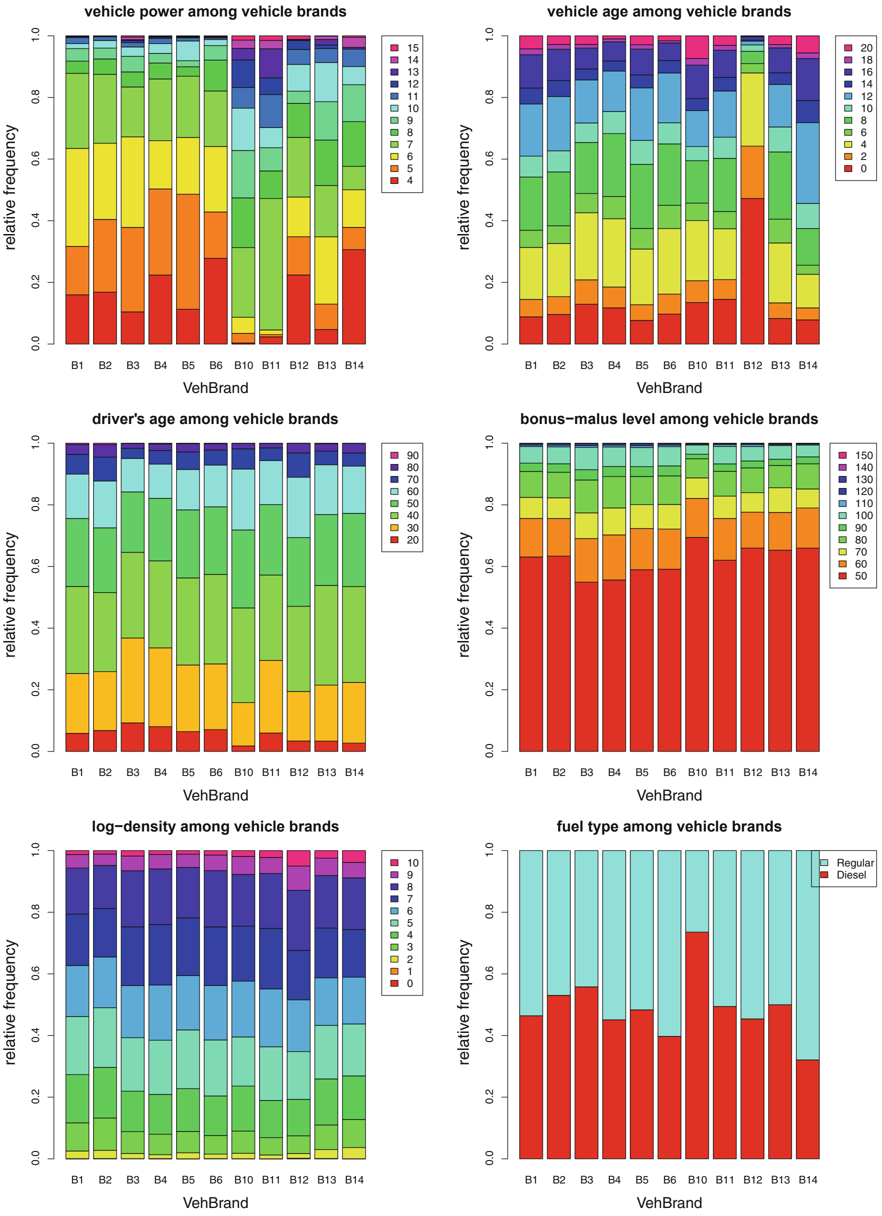


Fig. 13.13 Distribution of the variables `VehPower`, `VehAge`, `DrivAge`, `BonusMalus`, log-`Density`, `VehGas` for each car brand `VehBrand`, individually

Table 13.4 Cramér's V for the categorical feature components vs. the categorized continuous components

	VehPower	VehAge	DrivAge	BonusMalus	log-Density	VehGas	Region
VehBrand	0.16	0.17	0.06	0.03	0.05	0.12	0.13
Region	0.04	0.09	0.05	0.04	0.24	0.09	
Area					0.87		

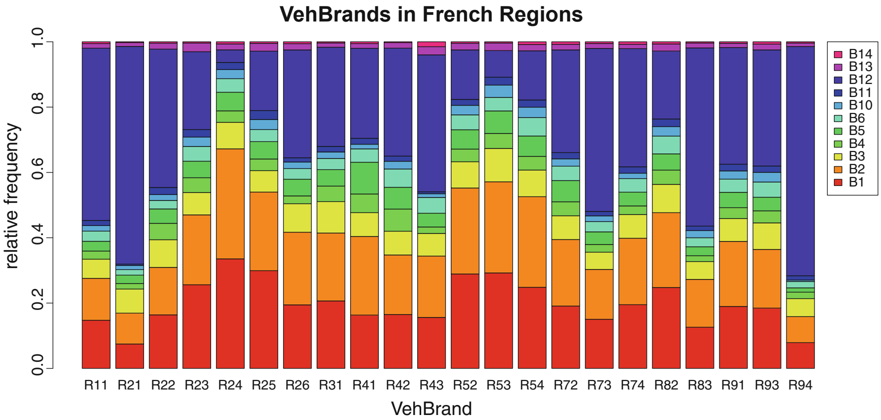


Fig. 13.14 VehBrands in the different French Regions

We scale the test statistics to the interval [0, 1] by dividing it by the comonotonic (maximal dependent) case and by the sample size n. This motivates Cramér's V

$$V = \sqrt{\frac{\chi^2/n}{\min\{m_1 - 1, m_2 - 1\}}} \in [0, 1].$$

Section 7.2.3 of Cohen [78] gives a rule of thumb for small, medium and large dependence. Cohen [78] calls the association between x_1 and x_2 small if $V\sqrt{\min\{m_1 - 1, m_2 - 1\}}$ is less 0.1, it is of medium strength for $V\sqrt{\min\{m_1 - 1, m_2 - 1\}}$ of size 0.3, and it is a large effect if this value is around 0.5. Our results are presented in Table 13.4. Clearly, there is some association between VehBrand and both VehPower and VehAge, this can also be seen from Fig. 13.13, for the remaining variables the dependence is somewhat weaker. Not surprisingly, Cramér's V shows the largest value between Region and log-Density.

In Fig. 13.14 we show the VehBrands in the different French Regions, Cramér's V is 0.13 for these two categorical variables, multiplying with $\sqrt{11 - 1}$ gives a value bigger than 0.4 which is a considerable association according to Cohen [78]. We note that in some regions the French car brands B1 and B2 are very dominant, whereas on the Isle of Corse (R94) 80% of the cars in our portfolio are Japanese

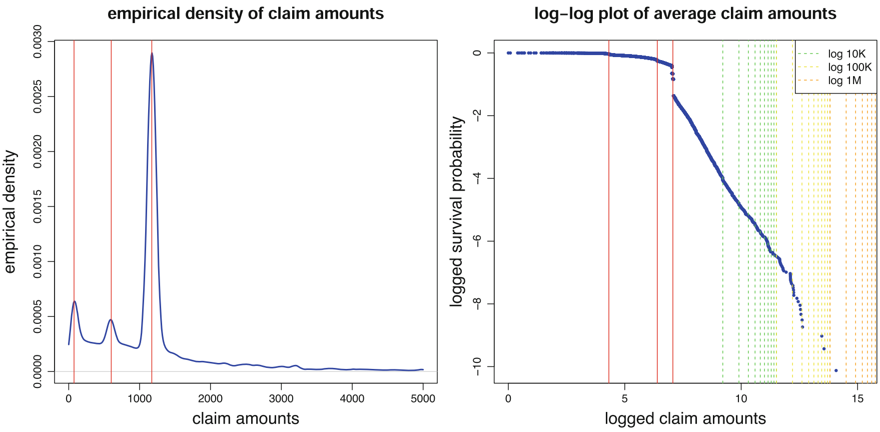


Fig. 13.15 Empirical density and log-log plots of the observed claim amounts

or Korean cars B12. Our portfolio has its biggest exposure in Region R24, see Fig. 13.11, in this region French cars are predominant.

Next, we study the claim sizes of this French MTPL example. Figure 13.15 shows the empirical density plot and the log-log plot. These two plots already illustrate the main difficulty we often face in claim size modeling. From the empirical density plot we observe that there are many payments of fixed size (red vertical lines) which do not match any absolutely continuous distribution function assumption. The log-log plot shows heavy-tailedness because we observe asymptotically a straight line with negative slope on the log-scale, this indicates regularly varying tails and, thus, the EDF is not a suitable model on the original observation scale.

Figure 13.16 gives the boxplots of the claim sizes per feature level (we omit the claims outside the whiskers because heavy-tailedness would distort the picture). The empirical mean in orange is much bigger than the median in red color, which also expresses the heavy-tailedness. From these plots we conclude that the claim sizes seem less sensitive in feature values which may question the use of a regression model for claim sizes.

Figure 13.17 shows the density plots for different feature levels. Interestingly, it seems that the features determine the sizes of the modes, for instance, if we focus on Area, Fig. 13.17 (top-left), we see that the area codes mainly influence the sizes of the modes. This may be interpreted by modes corresponding to different claim types which occur at different frequencies among the area codes.

13.2 Swedish Motorcycle Data

Our second example considers the Swedish motorcycle data which originally has been used in Ohlsson–Johansson [290]. It is available through the R library

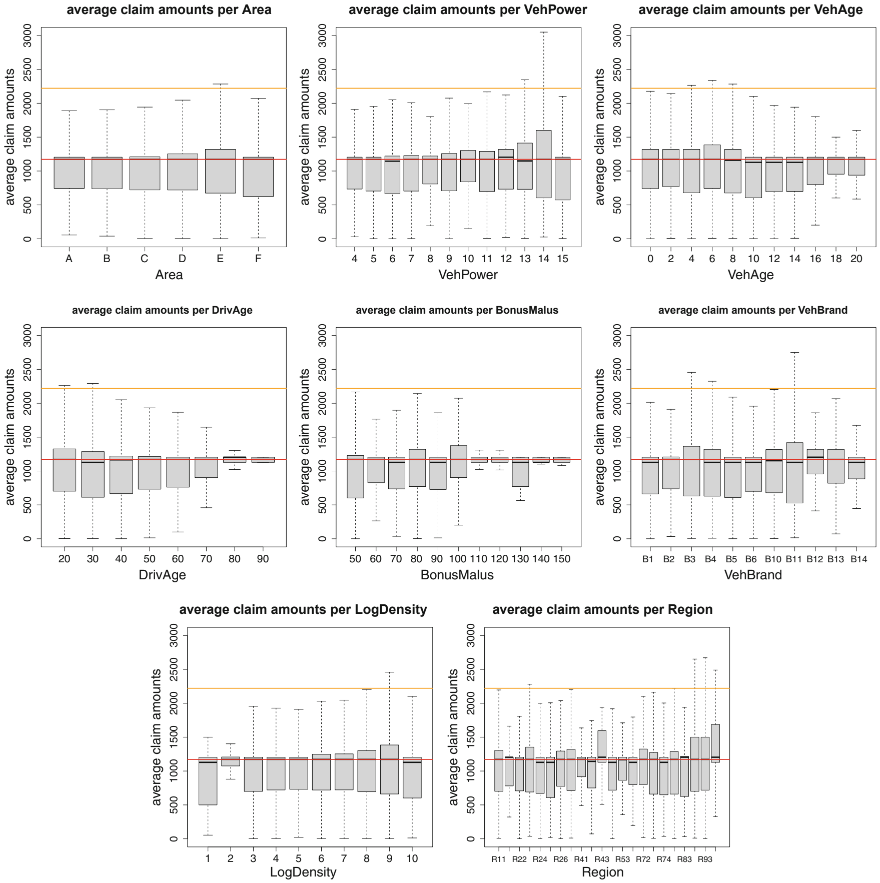


Fig. 13.16 Boxplots of claim sizes per feature level: these plots omit the claims outside the whiskers; red color shows the median and orange color the empirical mean

`CASdatasets` [113], and it is called `swmotorcycle`. Listing 13.3 shows the data cleaning that we have used, and Listing 13.4 gives an excerpt of the cleaned data.

We briefly describe the data. The data considers comprehensive insurance for motorcycles. This covers loss or damage of motorcycles other than collision, e.g., caused by theft, fire or vandalism. The data considers aggregated claims on feature levels for years 1994–1998. We have claims on 656 out of the 62’036 different features, thus, only slightly more than 1% of all feature combinations suffer a claim in the considered period.

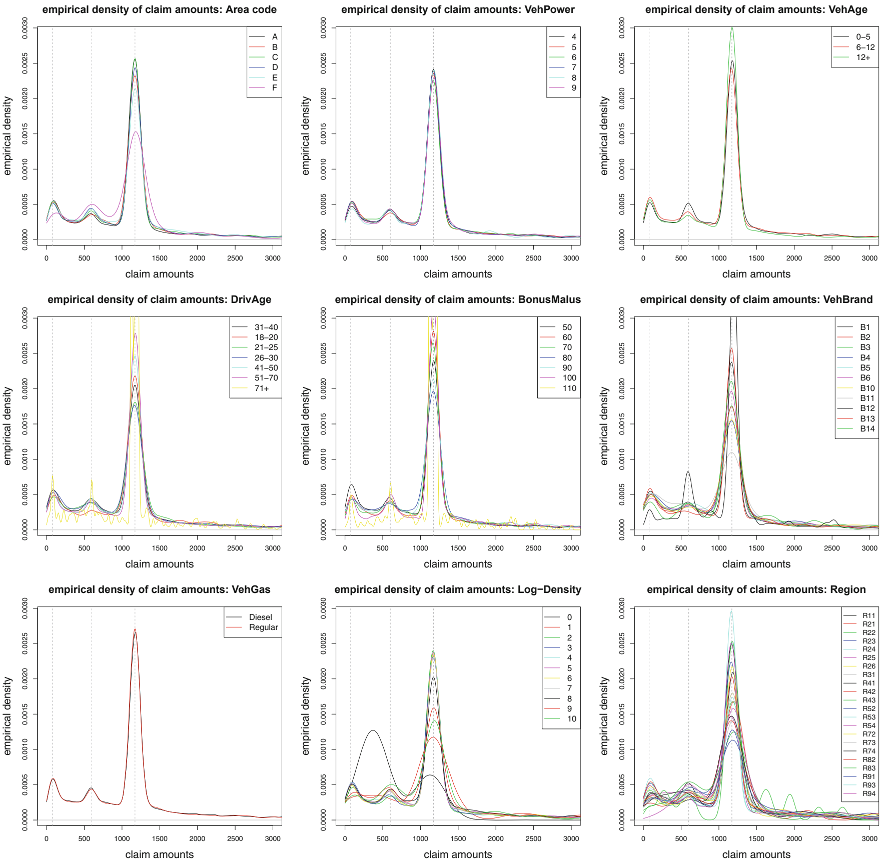


Fig. 13.17 Empirical claim size densities split w.r.t. the different levels of the feature components

We start by describing the available variables on lines 2–10 of Listing 13.4:

1. `OwnerAge`: age of motorcycle owner in $\{18, \dots, 70\}$ years (we censor at 70 because of scarcity of data above);
2. `Gender`: gender of motorcycle owner either being `Female` or `Male`;
3. `Area`: 7 geographical Swedish zones being (1) central parts of Sweden's three largest cities, (2) suburbs and middle-sized towns, (3) lesser towns except those in zones (5)–(7), (4) small towns and countryside except those in zones (5)–(7), (5) Northern towns, (6) Northern countryside, and (7) Gotland (Sweden's largest island);
4. `RiskClass`: 7 ordered motorcycle classes received from the so-called EV ratio defined as (Engine power in kW × 100) / (Vehicle weight in kg + 75kg);
5. `VehAge`: age of motorcycle in $\{0, \dots, 30\}$ years (we censor at 30);
6. `BonusClass`: ordered bonus-malus class from 1 to 7, entry level is 1;

Listing 13.3 Data cleaning applied to the Swedish motorcycle data set

```
library(CASdatasets)
data(swmotorcycle)
mcdata <- swmotorcycle
mcdata$Gender <- as.factor(mcdata$Gender)
mcdata$Area   <- as.factor(mcdata$Area)
mcdata$Area   <- factor(mcdata$Area,levels(mcdata$Area)[c(1,7,3,6,5,4,2)])
mcdata$Area   <- c("Zone 1","Zone 2","Zone 3","Zone 4","Zone 5",
                                     "Zone 6","Zone 7")[as.integer(mcdata$Area)]
mcdata$Area       <- as.factor(mcdata$Area)
mcdata$RiskClass  <- as.factor(mcdata$RiskClass)
mcdata$RiskClass  <- factor(mcdata$RiskClass,
                     levels(mcdata$RiskClass)[c(1,6,7,3,4,5,2)])
mcdata$RiskClass  <- as.integer(mcdata$RiskClass)
mcdata$BonusClass <- as.integer(as.factor(mcdata$BonusClass))
#
mcdata  <- mcdata[which(mcdata$OwnerAge>=18),]  # only minimal age 18
mcdata$OwnerAge <- pmin(70, mcdata$OwnerAge)    # set maximal age 70
mcdata$VehAge <- pmin(30, mcdata$VehAge)        # set maximal motorcycle age 30
mcdata <- mcdata[which(mcdata$Exposure>0),]     # only positive exposures
```

Listing 13.4 Excerpt of the Swedish motorcycle data set

```
'data.frame':   62036 obs. of  9 variables:
 $ OwnerAge   : num  18 18 18 18 18 18 18 18 18 18 ...
 $ Gender     : Factor w/ 2 levels "Female","Male": 1 1 1 1 1 1 1 1 1 1 ...
 $ Area       : Factor w/ 7 levels "Zone 1","Zone 2",..: 1 1 1 1 2 2 2 3 ...
 $ RiskClass  : int  1 2 3 3 1 1 3 1 1 1 ...
 $ VehAge     : num  8 11 9 9 11 12 24 4 6 6 ...
 $ BonusClass : int  2 2 3 4 1 1 2 1 1 2 ...
 $ Exposure   : num  1 0.778 0.499 0.501 0.929 ...
 $ ClaimNb    : int  0 0 0 0 0 0 0 0 0 0 ...
 $ ClaimAmount: int  0 0 0 0 0 0 0 0 0 0 ...
```

7. `Exposure`: total exposure in yearly units, these exposures are aggregated for given feature combinations, resulting in total exposures $[0.0274, 31.3397]$, the shortest entry referring to 10 days and the longest one to more than 31 years;
8. `ClaimNb`: number of claims N_i for a given feature;
9. `ClaimAmount`: total claim amount for a give feature (aggregated over all claims).

We start with a descriptive and exploratory analysis of the Swedish motorcycle data of Listing 13.4. We have $n = 62'036$ different feature combinations with positive `Exposure`. This `Exposure` is aggregated over individual policies with a fixed feature combination. We denote by N_i the number of claims on feature i, this corresponds to `ClaimNb`, and the total claim amount `ClaimAmount` is denoted by $S_i = \sum_{j=1}^{N_i} Z_{i,j}$, where $Z_{i,j}$ are the individual claim sizes on feature i (in case of claims). The empirical claim frequency is $\bar{\lambda} = \sum_{i=1}^{n} N_i / \sum_{i=1}^{n} v_i = 1.05\%$, and the average claim size is $\bar{\mu} = \sum_{i=1}^{n} S_i / \sum_{i=1}^{n} N_i = 24'641$ Swedish crowns SEK.

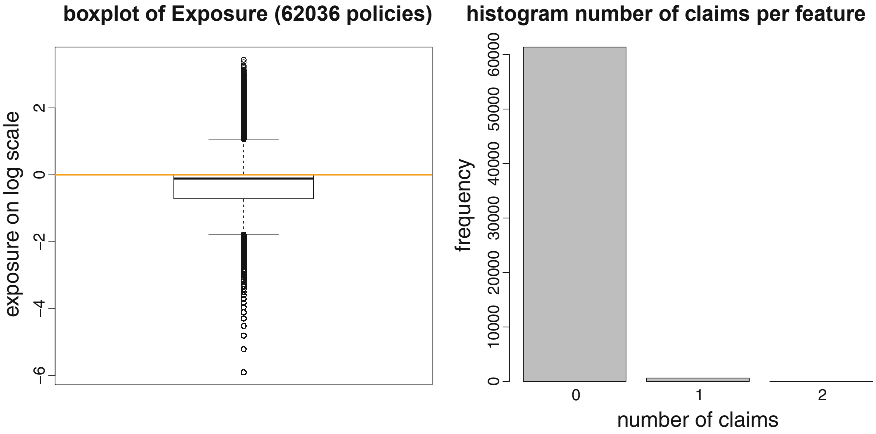


Fig. 13.18 (lhs) Boxplot of `Exposure` on the log-scale (the horizontal line corresponds to 1 accounting year), (rhs) histogram of the number of observed claims `ClaimNb` per feature of the Swedish motorcycle data

Figure 13.18 shows the boxplot over all `Exposures` and the claim counts on all insurance policies. We note that insurance claims are rare events for this product, because the empirical claim frequency is only $\bar{\lambda} = 1.05\%$.

Figures 13.19 and 13.20 give the marginal total exposures (split by gender), the marginal claim frequencies and the marginal average claim amounts for the covariate components `OwnerAge`, `Area`, `RiskClass`, `VehAge` and `BonusClass`. We observe that we have a very imbalanced portfolio between genders, only 11% of the total exposure is coming from females. The empirical claim frequency of females is 0.86% and the one of males is 1.08%. We note that the female claim frequency comes from (only) 61 claims (based on an exposure for female of 7'094 accounting years, versus 57'679 for male). Therefore, it is difficult to analyze females separately, and all marginal claim frequencies and claim sizes in Figs. 13.19 and 13.20 (middle and rhs) are analyzed jointly for both genders. If we run a simple Poisson GLM that only involves `Gender` as feature component, it turns out that the female frequency is 20% lower than the male frequency (remember we have the balance property on each dummy variable, see Example 5.12), but this variable should not be kept in the model on a 5% significance level. The same holds for claim amounts.

The empirical marginal frequencies in Figs. 13.19 and 13.20 (middle) are complemented with confidence bands of ±2 standard deviations. From the plots we conclude that we should keep the explanatory variables `OwnerAge`, `Area`, `RiskClass` and `VehAge`, but the variable `BonusClass` does not seem to have any predictive power. At the first sight, this seems surprising because the bonus class encodes the past claims history. The reason that the bonus class is not needed for our claims is that we consider comprehensive insurance for motorcycles covering loss or damage of motorcycles other than collision (for instance, caused by theft, fire or vandalism), and the bonus class encodes collision claims.

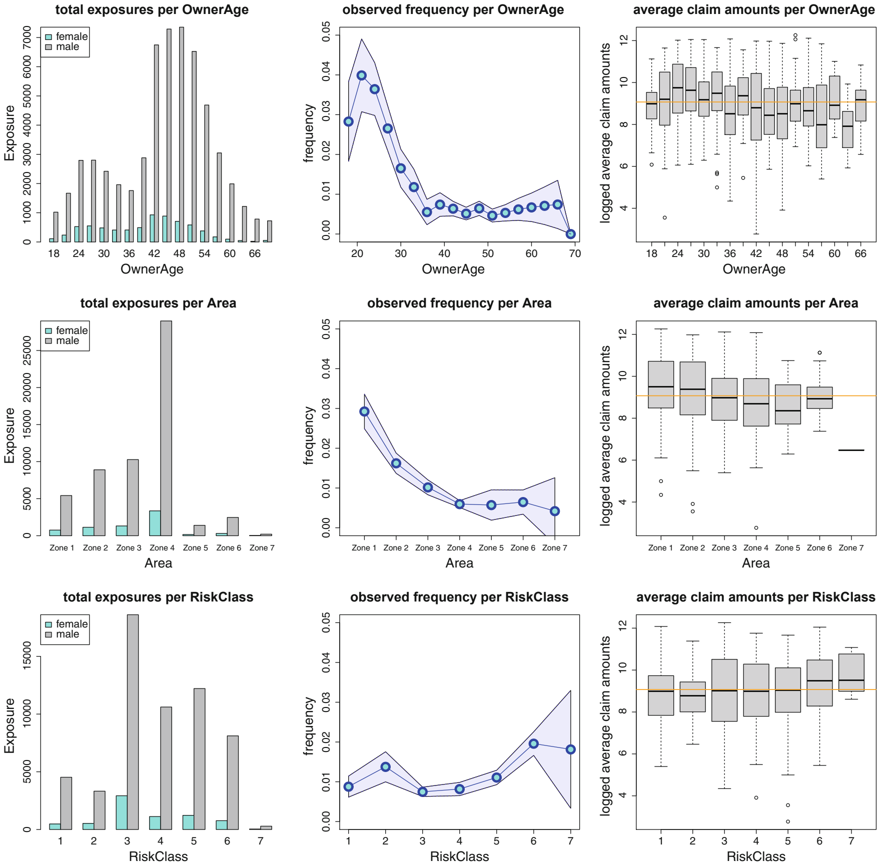


Fig. 13.19 (Top, middle and bottom rows) `OwnerAge`, `Area`, `RiskClass`: (lhs) histogram of exposures (split by gender), (middle) observed claim frequency, (rhs) boxplot of observed average claim amounts $\bar{\mu}_i = S_i/N_i$ of features with $N_i > 0$ (on log-scale)

For a regression analysis Zones 5 to 7 should be merged because of small exposures and a similar behavior, the same applies to `RiskClass` 6 and 7, and `VehAge` above 20.

Figure 13.21 shows the correlations between the features: (top) correlations between continuous features, (bottom), dependence between continuous features and the categorical `Area` features. We have some dependence, for instance, in `Zone 1` (three largest Swedish cities) the motorcycles are more light (`RiskClass`) and less old. Older people drive less heavy motorcycles that are more old, and older motorcycles are less heavy.

Figure 13.22 gives the empirical density, empirical distribution and log-log plot of average claim amounts $\bar{\mu}_i = S_i/N_i$. From the log-log plot we conclude that the average claim amounts are not heavy-tailed for this motorcycle insurance product.

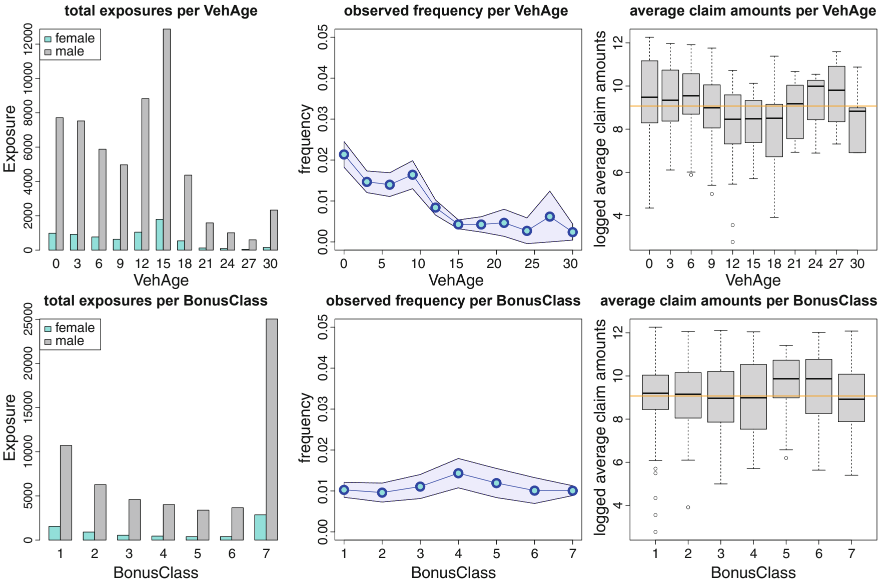


Fig. 13.20 (Top and bottom rows) `VehAge`, `BonusClass`: (lhs) histogram of exposures (split by gender), (middle) observed claim frequency, (rhs) boxplot of observed average claim amounts $\bar{\mu}_i = S_i/N_i$ of features with $N_i > 0$ (on log-scale)

13.3 Wisconsin Local Government Property Insurance Fund

The third example considers property insurance claims of the Wisconsin Local Government Property Insurance Fund (LGPIF). This data[8] has been made available through the book project of Frees [135],[9] and is also used in Lee et al. [236]. The Wisconsin LGPIF is an insurance pool that is managed by the Wisconsin Office of the Insurance Commissioner. This fund provides insurance protection to local governmental institutions such as counties, schools, libraries, airports, etc. It insures property claims for buildings and motor vehicles, and it excludes certain natural and man made perils like flood, earthquakes or nuclear accidents. We give a description of the data (we have applied some data cleaning to the original data).

The special feature of this data is that we have a short claim description on line 11 of Listing 13.5. This description will allow us to better understand the claim type beyond just knowing the hazard type that has been affected.

Figure 13.23 gives the empirical density (upper-truncated at 50’000) and the log-log plot of the observed LGPIF claim amounts. Most claims are below 10’000, however, the log-log plot shows clearly that the data is heavy-tailed, the largest claim being

[8] https://github.com/OpenActTexts/Loss-Data-Analytics/tree/master/Data.

[9] https://ewfrees.github.io/Loss-Data-Analytics/.

	OwnerAge	RiskClass	VehAge
OwnerAge		-11%	7%
RiskClass	-10%		-19%
VehAge	6%	-12%	

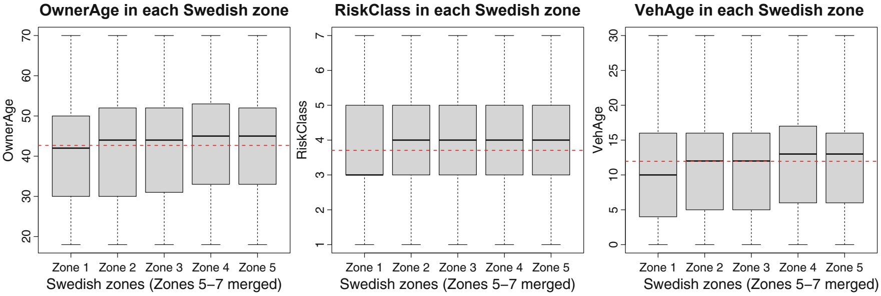


Fig. 13.21 (Top) Correlations: top-right shows Pearson's correlation; bottom-left shows Spearman's Rho; (bottom) boxplots of `OwnerAge`, `RiskClass`, `VehAge` versus `Area` (where Zones 5–7 have been merged)

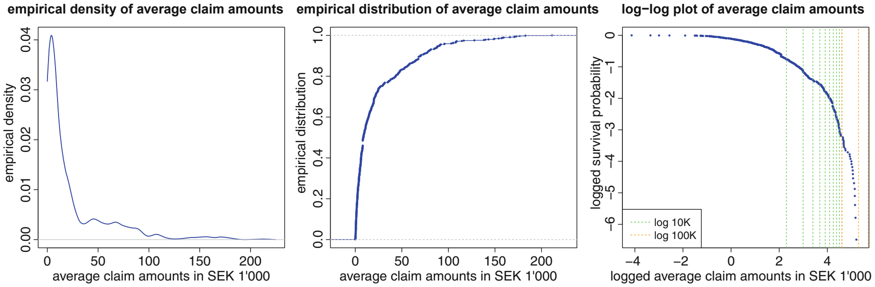


Fig. 13.22 (lhs) Empirical density (middle) empirical distribution and (rhs) log-log plot of average claim amounts $\bar{\mu}_i = S_i/N_i$ of features with $N_i > 0$

12'922'218 and 13 claims being above 1 million. These claims are further described by the features given in Listing 13.5.

In our example we will not focus on modeling the claim sizes, but we rather aim at predicting the hazard types from the claim descriptions. There are 9 different hazard types: Fire, Lightning, Hail, Wind, WaterW, WaterNW, Vehicle, Vandalism and Misc. The last label contains all claims that cannot be allocated to one of the previous hazard types, and WaterW refers to weather related water claims and WaterNW to the non-weather related ones. If we only focus on this latter problem we have more data available as there is a training data set and a validation data

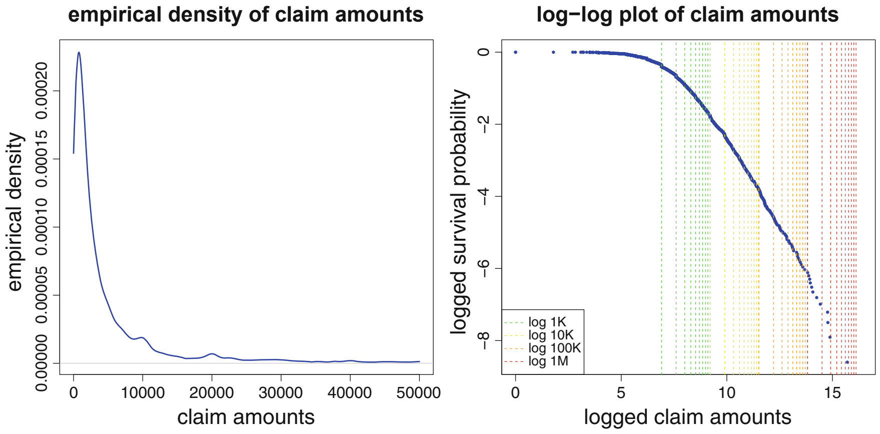


Fig. 13.23 (lhs) Empirical density (upper-truncated at 50'000), (rhs) log-log plot of the observed LGPIF claim amounts

Listing 13.5 Excerpt of the Wisconsin LGPIF data set

```
'data.frame':   5424 obs. of  10 variables:
 $ PolicyNum   : int 120002 120003 120003 120003 120003 120003 120003 ...
 $ Year        : int 2010 2007 2008 2007 2009 2010 2007 2007 2009 2007 ...
 $ Claim       : num 6839 2085 8775 600 34610 ...
 $ Deduct      : int 1000 5000 5000 5000 5000 5000 5000 5000 5000 5000 ...
 $ EntityType  : Factor w/ 6 levels "City","County",..: 2 2 2 2 2 2 2 2 2 2 ...
 $ CoverageCode: Factor w/ 13 levels "CE","CF","CS",..: 12 12 11 11 11 12 ...
 $ Fire5       : int 4 0 0 0 0 0 0 0 0 0 ...
 $ CountyCode  : Factor w/ 72 levels "ADA","ASH","BAR",..: 2 3 3 3 3 3 3 3...
 $ Hazard      : Factor w/ 9 levels "Fire","Hail",..: 3 3 5 5 9 6 3 3 3 3 ...
 $ Description : chr  "lightning damage" "lightning damage at Comm. Center" ...
```

set with hazard types and claim descriptions.[10] In total we have 6'031 such claim descriptions, see Listing 13.6, which are studied in our text recognition Chap. 10.

Listing 13.6 Excerpt of the Wisconsin LGPIF claim descriptions

```
'data.frame':   6031 obs. of  2 variables:
 Hazard     : Factor w/ 9 levels "Fire","Hail",..: 1 3 3 5 5 9 3 6 ...
 Description: chr "fire damage at Town Hall"
                  "lightning damage at water tower" ...
```

[10] https://github.com/OpenActTexts/Loss-Data-Analytics/tree/master/Data.

13.4 Swiss Accident Insurance Data

Our next example considers Swiss accident insurance data.[11] This data set is not publicly available. Swiss accident insurance is compulsory for employees, i.e., by law each employer has to sign an insurance contract to protect the employees against accidents. This insurance cover includes both work and leisure accidents, and it covers medical expenses and daily allowance. Listing 13.7 gives an excerpt of the data. Line BU indicates whether we have a workplace or a leisure accident, line 10 gives the medical expenses and line 12 shows the allowance expenses. In the subsequent analysis we only consider medical expenses.

Listing 13.7 Excerpt of the Swiss accident insurance data set

```
1   'data.frame':   339500 obs. of  11 variables:
2    $ Id        : int  1 2 3 4 5 6 7 8 9 10 ...
3    $ BU        : Factor w/ 2 levels "1","2": 1 1 2 2 2 1 2 2 2 1 ...
4    $ Sector    : Factor w/ 24 levels "5","12","13",..: 5 10 13 7 12 13 4 21 1 ...
5    $ AccQuart : int  3 2 1 3 4 4 1 2 1 3 ...
6    $ RepDel    : num  0 0 0 0 1 0 0 0 0 0 ...
7    $ Age       : num  45 20 20 20 60 55 30 25 20 20 ...
8    $ InjType   : Factor w/ 19 levels "1","2","3","4",..: 7 6 4 13 16 2 6 4 4 ...
9    $ InjPart   : Factor w/ 35 levels "1","2","3","4",..: 20 28 28 20 14 23 2 ...
10   $ Claim     : num  562 6675 700 57 2382 ...
11   $ NumbPaym : num  2 2 2 1 1 3 1 1 1 1 ...
12   $ Allowance: num  2345 5554 21 0 395 ...
```

Sector indicates the labor sector of the insured company, AccQuart gives the accident quarter since leisure claims have a seasonal component, RepDel gives the reporting delay in yearly units, Age is the age of the injured (in 5 years buckets), and InjType and InjPart denote the injury type and the injured body part. Figure 13.24 gives the empirical density (upper-truncated at 10’000) and the log-log plot of the observed Swiss accident insurance claim amounts. Most claims are below 5’000, however, the log-log plot shows some heavy-tailedness, the largest claim exceeding 1’300’000 CHF.

Figure 13.25 shows the average claim amounts split w.r.t. the different feature components (top) Sector, AccQuart, RepDel, (bottom) Age, InjType, InjPart, and moreover, split by work and leisure accidents (in cyan and gray in the colored version). Typically, leisure accidents are more numerous and more expensive on average than accidents at the work place. From Fig. 13.25 (top, left) we observe considerable variability in average claim sizes between the different labor sectors (cyan bars), whereas average leisure claim sizes (gray bars) are similar

[11] https://www.unfallstatistik.ch/.

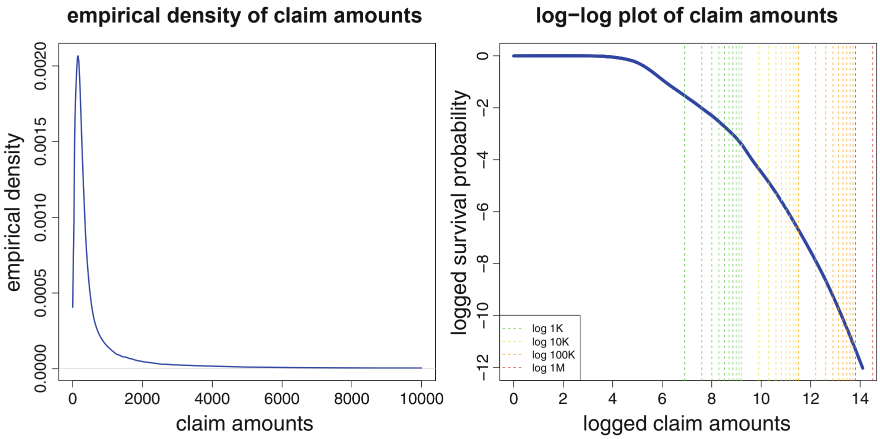


Fig. 13.24 (lhs) Empirical density (upper-truncated at 10’000), (rhs) log-log plot of the observed Swiss accident insurance claim amounts

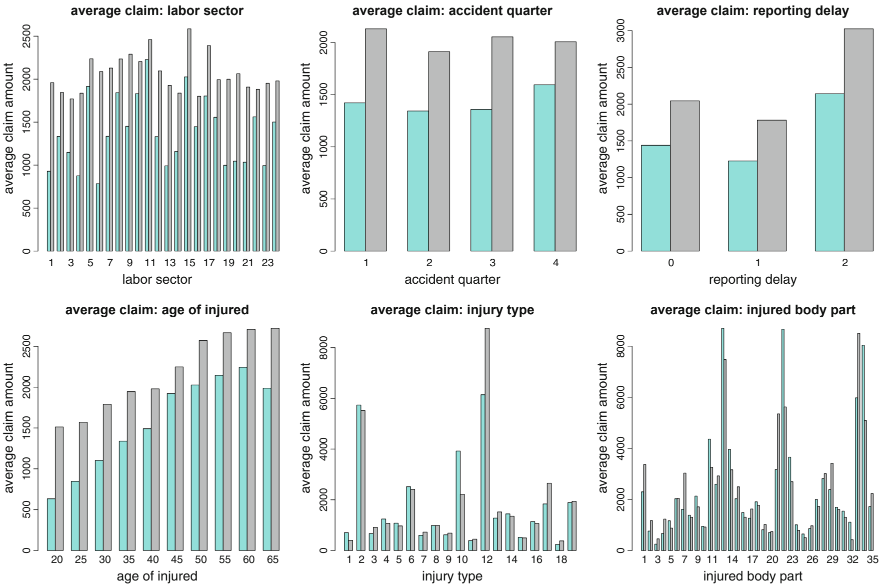


Fig. 13.25 Average claim amounts split w.r.t. the different feature components (top) `Sector`, `AccQuart`, `RepDel`, (bottom) `Age`, `InjType`, `InjPart`, and split by work and leisure accidents (cyan/gray in the colored version)

across the different labor sectors. Average claim sizes considerably differ between injury types and injured body parts (bottom, middle and right), but they do not differ between work and leisure claims.