ETH Zürich, D-MATH
HS 2017
Prof. Dr. Mario V. Wüthrich

Coordinator
A. Gabrielli

Non-Life Insurance: Mathematics and Statistics

Solution sheet 10

Solution 10.1 Tariffication Methods

In this exercise we work with $K = 2$ tariff criteria. The first criterion (vehicle type) has $I = 3$ risk characteristics:

$$\chi_{1,1} \text{ (passenger car)}, \quad \chi_{1,2} \text{ (delivery van)} \quad \text{and} \quad \chi_{1,3} \text{ (truck)}.$$

The second criterion (driver age) has $J = 4$ risk characteristics:

$$\chi_{2,1} \text{ (21 - 30 years)}, \quad \chi_{2,2} \text{ (31 - 40 years)}, \quad \chi_{2,3} \text{ (41 - 50 years)} \quad \text{and} \quad \chi_{2,4} \text{ (51 - 60 years)}.$$

The claim amounts $S_{i,j}$ for the risk classes $(i,j), 1 \le i \le 3, 1 \le j \le 4$, are given on the exercise sheet. We work with a multiplicative tariff structure. In particular, we use the model

$$\mathbb{E}[S_{i,j}] = v_{i,j}\, \mu\, \chi_{1,i}\, \chi_{2,j},$$

for all $1 \le i \le 3, 1 \le j \le 4$, where we set the number of policies $v_{i,j} = 1$. Moreover, in order to get a unique solution, we set $\mu = 1$ and $\chi_{1,1} = 1$. Therefore, there remains to find the risk characteristics $\chi_{1,2}, \chi_{1,3}, \chi_{2,1}, \chi_{2,2}, \chi_{2,3}, \chi_{2,4}$.

(a) In the method of Bailey & Simon, these risk characteristics are found by minimizing

$$X^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(S_{i,j} - v_{i,j}\, \mu\, \chi_{1,i}\, \chi_{2,j})^2}{v_{i,j}\, \mu\, \chi_{1,i}\, \chi_{2,j}} = \sum_{i=1}^{3} \sum_{j=1}^{4} \frac{(S_{i,j} - \chi_{1,i}\, \chi_{2,j})^2}{\chi_{1,i}\, \chi_{2,j}}.$$

Let $i \in \{2,3\}$. Then $\hat{\chi}_{1,i}$ is found by the solution of

$$\begin{aligned}
0 &\stackrel{!}{=} \frac{\partial}{\partial \chi_{1,i}} X^2 \\
&= \sum_{j=1}^{4} \frac{\partial}{\partial \chi_{1,i}} \frac{(S_{i,j} - \chi_{1,i}\, \chi_{2,j})^2}{\chi_{1,i}\, \chi_{2,j}} \\
&= \sum_{j=1}^{4} \frac{-2(S_{i,j} - \chi_{1,i}\, \chi_{2,j})\chi_{1,i}\, \chi_{2,j} - (S_{i,j} - \chi_{1,i}\, \chi_{2,j})^2}{\chi_{1,i}^2\, \chi_{2,j}} \\
&= \sum_{j=1}^{4} \frac{-2S_{i,j}\chi_{1,i}\, \chi_{2,j} + 2\chi_{1,i}^2\, \chi_{2,j}^2 - S_{i,j}^2 + 2S_{i,j}\chi_{1,i}\, \chi_{2,j} - \chi_{1,i}^2\, \chi_{2,j}^2}{\chi_{1,i}^2\, \chi_{2,j}} \\
&= \sum_{j=1}^{4} \frac{\chi_{1,i}^2\, \chi_{2,j}^2 - S_{i,j}^2}{\chi_{1,i}^2\, \chi_{2,j}} \\
&= \sum_{j=1}^{4} \chi_{2,j} - \frac{1}{\chi_{1,i}^2} \sum_{j=1}^{4} \frac{S_{i,j}^2}{\chi_{2,j}}.
\end{aligned}$$

Thus, we get

$$\hat{\chi}_{1,i} = \left(\frac{\sum_{j=1}^{4} S_{i,j}^2/\chi_{2,j}}{\sum_{j=1}^{4} \chi_{2,j}} \right)^{1/2}.$$

By an analogous calculation, one finds

$$\hat{\chi}_{2,j} = \left(\frac{\sum_{i=1}^{3} S_{i,j}^2 / \chi_{1,i}}{\sum_{i=1}^{3} \chi_{1,i}} \right)^{1/2},$$

for $j \in \{1,2,3,4\}$. For solving these equations, one has to apply a root-finding algorithm like for example the Newton-Raphson method. We get the following multiplicative tariff structure:

	21-30y	31-40y	41-50y	51-60y	$\hat{\chi}_{1,i}$
passenger car	2'176	1'751	1'491	1'493	1
delivery van	2'079	1'674	1'425	1'427	0.96
truck	2'456	1'977	1'684	1'686	1.13
$\hat{\chi}_{2,j}$	2'176	1'751	1'491	1'493	

We see that the risk characteristics for the classes passenger car and delivery van are close to each other, whereas for trucks we have a higher tariff. Moreover, an insured with age in the class 21 - 30 years gets a considerably higher tariff than an insured with age in the class 31 - 40 years. The smallest tariff is assigned to insureds with age in the classes 41 - 50 years and 51 - 60 years. Note that we have

$$\sum_{i=1}^{3}\sum_{j=1}^{4} \hat{\chi}_{1,i}\hat{\chi}_{2,j} = 21'320 > 21'300 = \sum_{i=1}^{3}\sum_{j=1}^{4} S_{i,j},$$

which confirms the (systematic) positive bias of the method of Bailey & Simon shown in Lemma 7.2 of the lecture notes.

(b) In the method of Bailey & Jung, which is also called method of marginal totals, the risk characteristics $\chi_{1,2}, \chi_{1,3}, \chi_{2,1}, \chi_{2,2}, \chi_{2,3}, \chi_{2,4}$ are found by solving the equations

$$\sum_{j=1}^{J} v_{i,j}\, \mu\, \chi_{1,i}\, \chi_{2,j} = \sum_{j=1}^{J} S_{i,j},$$
$$\sum_{i=1}^{I} v_{i,j}\, \mu\, \chi_{1,i}\, \chi_{2,j} = \sum_{i=1}^{I} S_{i,j}.$$

Since $I = 3, J = 4$ and we work with $v_{i,j} = 1$ and set $\mu = 1$, we get the equations

$$\sum_{j=1}^{4} \chi_{1,i}\, \chi_{2,j} = \sum_{j=1}^{4} S_{i,j},$$
$$\sum_{i=1}^{3} \chi_{1,i}\, \chi_{2,j} = \sum_{i=1}^{3} S_{i,j}.$$

Thus, for $i \in \{2,3\}$ and $j \in \{1,2,3,4\}$, we get

$$\hat{\chi}_{1,i} = \sum_{j=1}^{4} S_{i,j} \Big/ \sum_{j=1}^{4} \chi_{2,j},$$
$$\hat{\chi}_{2,j} = \sum_{i=1}^{3} S_{i,j} \Big/ \sum_{i=1}^{3} \chi_{1,i}.$$

Analogously to the method of Bailey & Simon, one has to solve this system of equations using a root-finding algorithm. We get the following multiplicative tariff structure:

	21-30y	31-40y	41-50y	51-60y	$\hat{\chi}_{1,i}$
passenger car	2'170	1'749	1'490	1'490	1
delivery van	2'076	1'673	1'425	1'425	0.96
truck	2'454	1'977	1'684	1'684	1.13
$\hat{\chi}_{2,j}$	2'170	1'749	1'490	1'490	

We see that the results are very close to those in part (a) where we applied the method of Bailey & Simon. However, now we have

$$\sum_{i=1}^{3}\sum_{j=1}^{4}\hat{\chi}_{1,i}\hat{\chi}_{2,j} = 21'300 = 21'300 = \sum_{i=1}^{3}\sum_{j=1}^{4}S_{i,j},$$

which comes as no surprise as we fitted the risk characteristics such that the above equality holds true.

(c) In the log-linear regression model we work with the stochastic model

$$X_{i,j} \overset{\text{def}}{=} \log\frac{S_{i,j}}{v_{i,j}} = \log S_{i,j} \sim \mathcal{N}(\beta_0 + \beta_{1,i} + \beta_{2,j}, \sigma^2),$$

where $\beta_0, \beta_{1,i}, \beta_{2,j} \in \mathbb{R}$ and $\sigma^2 > 0$, for all risk classes $(i,j), 1 \le i \le 3, 1 \le j \le 4$. The risk characteristics of the two tariff criteria vehicle type and driver age are now given by

$$\beta_{1,1} \text{ (passenger car)}, \quad \beta_{1,2} \text{ (delivery van)} \quad \text{and} \quad \beta_{1,3} \text{ (truck)},$$

and

$$\beta_{2,1} \text{ (21 - 30 years)}, \quad \beta_{2,2} \text{ (31 - 40 years)}, \quad \beta_{2,3} \text{ (41 - 50 years)} \quad \text{and} \quad \beta_{2,4} \text{ (51 - 60 years)}.$$

In order to get a unique solution, we set $\beta_{1,1} = \beta_{2,1} = 0$. Because this will simplify notation considerably, we write $\mathbf{X} = (X_1, \dots, X_M)'$ with $M = 12$ and

$$X_1 = X_{1,1}, \quad X_2 = X_{1,2}, \quad X_3 = X_{1,3}, \quad X_4 = X_{1,4}, \quad X_5 = X_{2,1}, \quad X_6 = X_{2,2},$$
$$X_7 = X_{2,3}, \quad X_8 = X_{2,4}, \quad X_9 = X_{3,1}, \quad X_{10} = X_{3,2}, \quad X_{11} = X_{3,3}, \quad X_{12} = X_{3,4}.$$

Moreover, we define

$$\boldsymbol{\beta} = (\beta_0, \beta_{1,2}, \beta_{1,3}, \beta_{2,2}, \beta_{2,3}, \beta_{2,4})' \in \mathbb{R}^{r+1},$$

where $r = 5$. Then, we assume that $\mathbf{X}$ has a multivariate Gaussian distribution

$$\mathbf{X} \sim \mathcal{N}(Z\boldsymbol{\beta}, \sigma^2 I),$$

where $I \in \mathbb{R}^{M\times M}$ denotes the identity matrix and $Z \in \mathbb{R}^{M\times(r+1)}$ is the so-called design matrix that satisfies

$$\mathbb{E}[\mathbf{X}] = Z\boldsymbol{\beta}.$$

For example for $m = 1$ we have

$$\mathbb{E}[X_m] = \mathbb{E}[X_1] = \mathbb{E}[X_{1,1}] = \beta_0 + \beta_{1,1} + \beta_{2,1} = \beta_0 = (1,0,0,0,0,0)\,\boldsymbol{\beta},$$

and for $m = 8$

$$\mathbb{E}[X_m] = \mathbb{E}[X_8] = \mathbb{E}[X_{2,4}] = \beta_0 + \beta_{1,2} + \beta_{2,4} = (1,1,0,0,0,1)\,\boldsymbol{\beta}.$$

Doing this for all $m \in \{1, \dots, 12\}$, we find the design matrix Z:

intercept (β_0)	van ($\beta_{1,2}$)	truck ($\beta_{1,3}$)	31-40y ($\beta_{2,2}$)	41-50y ($\beta_{2,3}$)	51-60y ($\beta_{2,4}$)
1	0	0	0	0	0
1	0	0	1	0	0
1	0	0	0	1	0
1	0	0	0	0	1
1	1	0	0	0	0
1	1	0	1	0	0
1	1	0	0	1	0
1	1	0	0	0	1
1	0	1	0	0	0
1	0	1	1	0	0
1	0	1	0	1	0
1	0	1	0	0	1

Here we would like to point out that we can also use R to find the design matrix, see the R-Code of Exercise 10.2. According to formula (7.9) of the lecture notes, the MLE $\hat{\boldsymbol{\beta}}^{\text{MLE}}$ of the parameter vector $\boldsymbol{\beta}$ is given by

$$\hat{\boldsymbol{\beta}}^{\text{MLE}} = [Z'(\sigma^2 I)^{-1} Z]^{-1} Z'(\sigma^2 I)^{-1} \mathbf{X} = (Z'Z)^{-1} Z' \mathbf{X}.$$

Note that $\hat{\boldsymbol{\beta}}^{\text{MLE}}$ does not depend on σ^2. Moreover, the design matrix Z has full column rank and, thus, $Z'Z$ is indeed invertible. See the R-Code given at the end of the solution to this exercise for the calculation of $\hat{\boldsymbol{\beta}}^{\text{MLE}}$. We get the following tariff structure:

$\hat{\beta}_0 = 7.688$	21-30y	31-40y	41-50y	51-60y	$\hat{\beta}_{1,i}$
passenger car	2'182	1'758	1'500	1'501	0
delivery van	2'063	1'663	1'417	1'419	-0.056
truck	2'444	1'970	1'680	1'682	0.113
$\hat{\beta}_{2,j}$	0	-0.216	-0.375	-0.374	

We see that the results are very close to those in parts (a) and (b) where we applied the method of Bailey & Simon and the method of Bailey & Jung. However, since we are now working in a stochastic framework, we also get standard errors and we can make statements about the statistical significance of the parameters. According to the R-output, we get the following p-values for the individual parameters:

	$\hat{\beta}_0$	$\hat{\beta}_{1,2}$	$\hat{\beta}_{1,3}$	$\hat{\beta}_{2,2}$	$\hat{\beta}_{2,3}$	$\hat{\beta}_{2,4}$
p-value	≈ 0	0.232	0.036	-0.005	0.0003	0.0003

R gets these p-values by applying a t-test individually to each parameter, whether they are equal to zero. While the p-values for $\hat{\beta}_0, \hat{\beta}_{1,3}, \hat{\beta}_{2,2}, \hat{\beta}_{2,3}, \hat{\beta}_{2,4}$ are smaller than 0.05 and, thus, these parameters are significantly different from zero, the p-value of $\hat{\beta}_{1,2}$ (delivery van) is fairly high. Hence, we might question if we really need the class delivery van.

In order to check whether there is statistical evidence that the classification into different types of vehicles could be omitted, we define the null hypothesis of the reduced model:

$$H_0 : \beta_{1,2} = \beta_{1,3} = 0,$$

i.e. we set $p = 2$ parameters equal to 0. Then we can perform the same analysis as above to get the MLE $\hat{\boldsymbol{\beta}}_{H_0}^{\text{MLE}}$. In particular, let Z_{H_0} be the design matrix Z without the second column van ($\beta_{1,2}$) and the third column truck ($\beta_{1,3}$). Then $\hat{\boldsymbol{\beta}}_{H_0}^{\text{MLE}}$ is given by

$$\hat{\boldsymbol{\beta}}_{H_0}^{\text{MLE}} = (Z_{H_0}' Z_{H_0})^{-1} Z_{H_0}' \mathbf{X}.$$

See the R-Code given below for the calculation of $\hat{\boldsymbol{\beta}}_{H_0}^{\text{MLE}}$. Now, for all $m \in \{1, \dots, 12\}$, we define the fitted value $\hat{X}_m^{\text{full}}$ of the full model and the fitted value $\hat{X}_m^{H_0}$ of the reduced model. In particular, we have

$$\hat{X}_m^{\text{full}} = \left[Z \hat{\boldsymbol{\beta}}^{\text{MLE}} \right]_m$$

and

$$\hat{X}_m^{H_0} = \left[Z_{H_0} \hat{\boldsymbol{\beta}}_{H_0}^{\text{MLE}} \right]_m,$$

where $[\cdot]_m$ denotes the m-th element of the corresponding vector, for all $m \in \{1, \dots, 12\}$. Moreover, we define

$$SS_{\text{err}}^{\text{full}} = \sum_{m=1}^{M} \left(X_m - \hat{X}_m^{\text{full}} \right)^2$$

and

$$SS_{\text{err}}^{H_0} = \sum_{m=1}^{M} \left(X_m - \hat{X}_m^{H_0} \right)^2.$$

According to formula (7.15) of the lecture notes, the test statistic

$$T = \frac{SS_{\text{err}}^{H_0} - SS_{\text{err}}^{\text{full}}}{SS_{\text{err}}^{\text{full}}} \frac{M - r - 1}{p} = 3 \frac{SS_{\text{err}}^{H_0} - SS_{\text{err}}^{\text{full}}}{SS_{\text{err}}^{\text{full}}}$$

has an F-distribution with degrees of freedeom given by $\text{df}_1 = p = 2$ and $\text{df}_2 = M - r - 1 = 6$. See the R-Code below for the calculation of T. We get

$$T \approx 8.336,$$

which corresponds to a p-value of approximately 1.85%. Thus, we can reject H_0 at significance level of 5%, i.e. there is no statistical evidence that the classification into different types of vehicles could be omitted.

(d) As we already mentioned above, the method of Bailey & Simon, the method of Bailey & Jung and the MLE method in the log-linear regression model all lead to approximately the same results. The only differences are, that with the method of Bailey & Jung we get coinciding marginal totals and with the log-linear regression model we are in a stochastic framework which allows for calculating parameter uncertainties and hypothesis testing.

```
### c)

### We apply the log-linear regression method to the observed
    claim amounts given on the exercise sheet

### Load the observed claim amounts into a matrix
S <- matrix(c
    (2000,2200,2500,1800,1600,2000,1500,1400,1700,1600,1400,1600)
    , nrow = 3)

### Define the design matrix Z
Z <- matrix(c(rep(1,12),rep(0,4),rep(1,4),rep(0,12),rep(1,4),
    rep(c(0,1,0,0),3),rep(c(0,0,1,0),3),rep(c(0,0,0,1),3)),nrow
    = 12)

### Store the design matrix Z (without the intercept term) and
    the dependent variable log(S_{i,j}) in one dataset
```

```
data <- cbind(Z[,-1],matrix(log(t(S)),nrow = 12))
data <- as.data.frame(data)
colnames(data) <- c("van", "truck", "X31_40y", "X41_50y", "X51_
    60y", "observation")

### Apply the regression model
linear.model1 <- lm(formula = observation ~ van + truck + X31_
    40y + X41_50y + X51_60y,data=data)

### Print the output of the regression model
summary(linear.model1)

### Fitted values
fitted(linear.model1)

### We can also get the parameters by applying the formula
    (7.9) of the lecture notes
solve(t(Z)%*%Z)%*%t(Z)%*%matrix(log(t(S)),nrow = 12)


### Apply the regression model under H_{0}
linear.model2 <- lm(formula = observation ~ X31_40y + X41_50y +
     X51_60y,data=data)

### Calculation of the test statistic T which has an F-
    distribution
T <- 3 * (sum((fitted(linear.model2) - data[,6])^2) - sum((
    fitted(linear.model1) - data[,6])^2)) / sum((fitted(linear.
    model1) - data[,6])^2)

### Calculation of the corresponding p-value
pf(T, 2, 6, lower.tail = FALSE)
```

Note that we could also define the covariates of factor type in R which then automatically implies that these covariates are of categorical type and R chooses the design matrix Z accordingly, see the R-Code for the solution of Exercise 10.2 given below.

Solution 10.2 Tariffication Methods

(a) In this exercise we work with $K = 3$ tariff criteria. The first criterion (vehicle class) has 2 risk characteristics:

$$\beta_{1,1} \text{ (weight over 60 kg and more than two gears)} \quad \text{and} \quad \beta_{1,2} \text{ (other)}.$$

The second criterion (vehicle age) also has 2 risk characteristics:

$$\beta_{2,1} \text{ (at most one year)} \quad \text{and} \quad \beta_{2,2} \text{ (more than one year)}.$$

The third criterion (geographic zone) has 3 risk characteristics:

$\beta_{3,1}$ (large cities), $\beta_{3,2}$ (middle-sized towns) and $\beta_{3,3}$ (smaller towns and countryside).

The observed number of claims N_{l_1,l_2,l_3}, the observed volumes v_{l_1,l_2,l_3} and the observed claim frequencies

$$\lambda_{l_1,l_2,l_3} = \frac{N_{l_1,l_2,l_3}}{v_{l_1,l_2,l_3}}$$

for the risk classes $(l_1, l_2, l_3), 1 \leq l_1 \leq 2, 1 \leq l_2 \leq 2, 1 \leq l_3 \leq 3$, are given on the exercise sheet. Now, for modelling purposes, we assume that all N_{l_1,l_2,l_3} are independent with

$$N_{l_1,l_2,l_3} \sim \text{Poi}(\lambda_{l_1,l_2,l_3} v_{l_1,l_2,l_3})$$

and define

$$X_{l_1,l_2,l_3} = \frac{N_{l_1,l_2,l_3}}{v_{l_1,l_2,l_3}}.$$

Then we use the model Ansatz

$$g(\lambda_{l_1,l_2,l_3}) = g\left(\mathbb{E}\left[\frac{N_{l_1,l_2,l_3}}{v_{l_1,l_2,l_3}}\right]\right) = g\left(\mathbb{E}\left[X_{l_1,l_2,l_3}\right]\right) = \beta_0 + \beta_{1,l_1} + \beta_{2,l_2} + \beta_{3,l_3},$$

where $\beta_0 \in \mathbb{R}$ and where we use the log-link function, i.e. $g(\cdot) = \log(\cdot)$. In order to get a unique solution, we set $\beta_{1,1} = \beta_{2,1} = \beta_{3,1} = 0$. Moreover, we define

$$\boldsymbol{\beta} = (\beta_0, \beta_{1,2}, \beta_{2,2}, \beta_{3,2}, \beta_{3,3})' \in \mathbb{R}^{r+1},$$

where $r = 4$. Similarly as in Exercise 10.1, (c), we will relabel the risk classes with the index $m \in \{1, \ldots, M\}$, where $M = 2 \cdot 2 \cdot 3 = 12$, define $\mathbf{X} = (X_1, \ldots, X_M)'$ and the design matrix $Z \in \mathbb{R}^{M \times (r+1)}$ that satisfies

$$\log \mathbb{E}[\mathbf{X}] = Z\boldsymbol{\beta},$$

where the logarithm is applied componentwise to $\mathbb{E}[\mathbf{X}]$. Let $m \in \{1, \ldots, 12\}$. According to Example 7.10 of the lecture notes, $X_m = N_m/v_m$ belongs to the exponential dispersion family with cumulant function $b(\cdot) = \exp\{\cdot\}$, $\theta_m = \log \lambda_m$, $w_m = v_m$ and dispersion parameter $\phi = 1$, i.e. we have

$$[Z\boldsymbol{\beta}]_m = \log \mathbb{E}[X_m] = \log \mathbb{E}\left[\frac{N_m}{v_m}\right] = \log \lambda_m = \theta_m,$$

where $[Z\boldsymbol{\beta}]_m$ denotes as above the m-th element of the vector $Z\boldsymbol{\beta}$. Thus, we assume that $X_1, \ldots, X_M$ are independent with

$$X_m \sim \text{EDF}(\theta_m = [Z\boldsymbol{\beta}]_m, \phi = 1, v_m, b(\cdot) = \exp\{\cdot\}),$$

for all $m \in \{1, \ldots, M\}$. According to Proposition 7.11 of the lecture notes, the MLE $\hat{\boldsymbol{\beta}}^{\text{MLE}}$ of $\boldsymbol{\beta}$ is the solution of

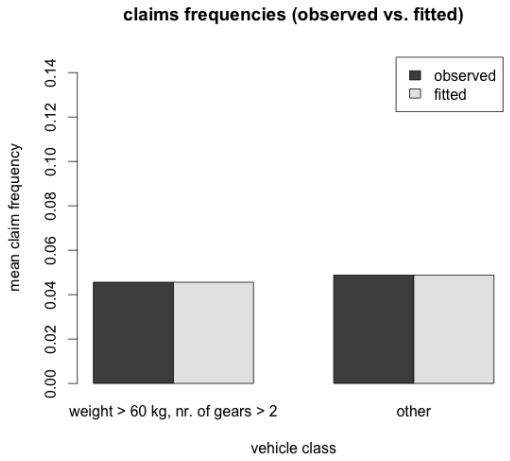
$$Z'V \exp\{Z\boldsymbol{\beta}\} = Z'V\mathbf{X}, \tag{1}$$

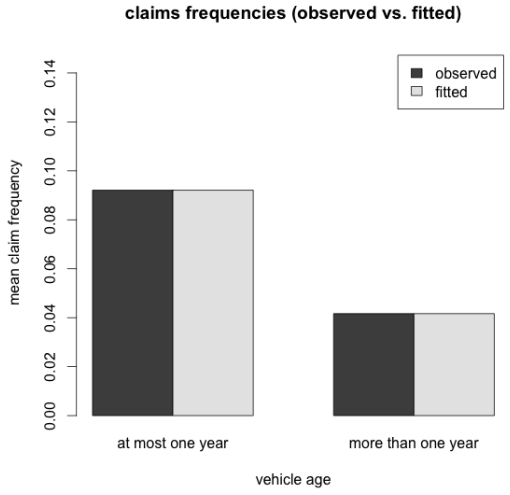
where the weight matrix V is given by $V = \text{diag}(v_1, \ldots, v_M)$. This equation has to be solved numerically. See the R-Code at the end of the solution to this exercise for the calculation of $\hat{\boldsymbol{\beta}}^{\text{MLE}}$. We get the following estimates:

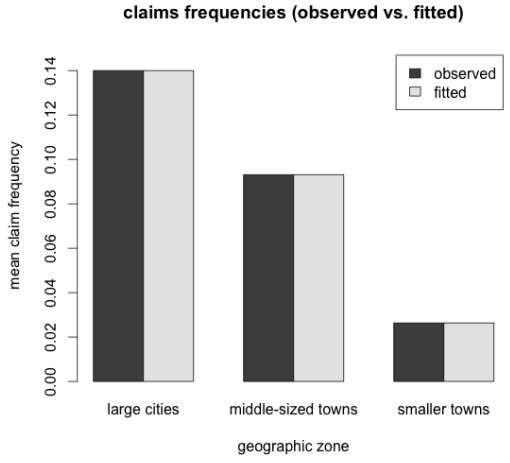
	$\hat{\beta}_0$	$\hat{\beta}_{1,2}$	$\hat{\beta}_{2,2}$	$\hat{\beta}_{3,2}$	$\hat{\beta}_{3,3}$
MLE	-1.435	-0.237	-0.502	-0.404	-1.657

We observe that insureds with a vehicle with weight over 60 kg and more than two gears tend to cause more claims than insureds with other vehicles. Analogoulsy, if the vehicle is at most one year old, we expect more claims than if it was older. Regarding the geographic zone, we see that driving in middle-sized towns leads to fewer claims than driving in large cities. Moreover, driving in smaller towns and countryside leads to even fewer claims than driving in middle-sized towns, where this difference is greater than the difference between large cities and middle-sized towns.

(b) The observed and the fitted claim frequencies against the vehicle class, the vehicle age and the geographical zone look as follows:

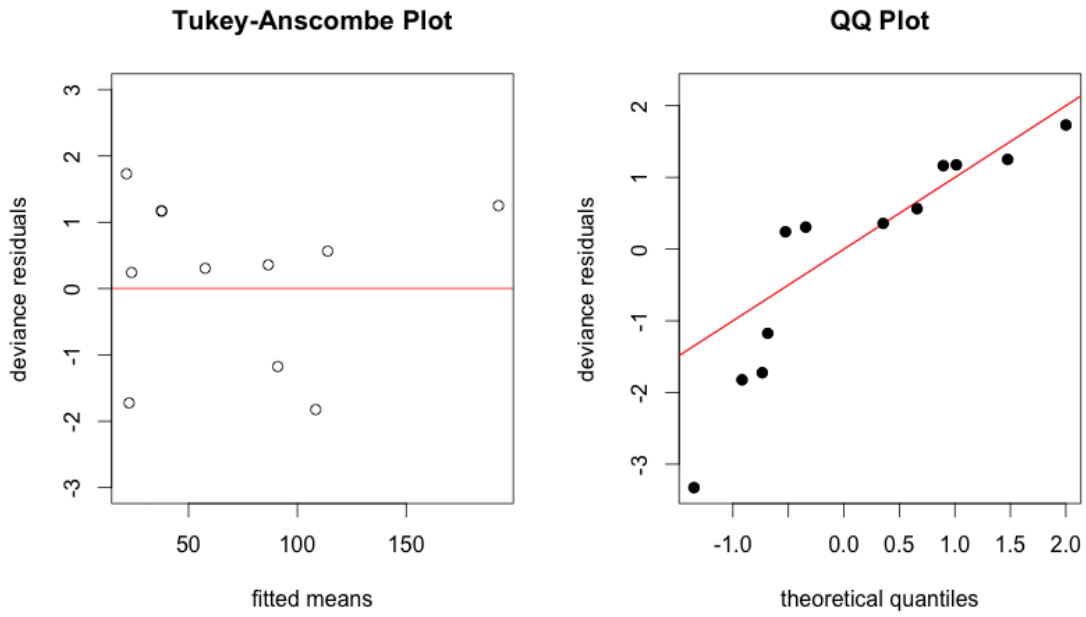






See the R-Code at the end of the solution to this exercise for creating the plots given above. Note that the observed and the fitted marginal claim frequencies are always the same. This is a direct consequence of equation (1) given above which ensures that the observed and the fitted total marginal sums are the same if we use the same volumes again. This is also the reason why in the marginal plot for the vehicle class we don't see that insureds with a vehicle with weight over 60 kg and more than two gears tend to cause more claims than insureds with other vehicles as expected after the discussion at the end of part (a). More precisely, for the vehicles with weight over 60 kg and more than two gears we have a smaller volume for the riskier classes with respect to the other tariff criteria vehicle age and geographic zone than for the other vehicles. This compensates for the fact that vehicles with weight over 60 kg and more than two gears tend to cause more claims than other vehicles, as seen at the end of part (a). For the other variables vehicle age and geographic zone we again see the same results as in part (a).

(c) The Tukey-Anscombe plot and the QQ plot look as follows:



See the R-Code at the end of the solution to this exercise for creating the plots given above. They are both not ideal, but considering that we only have 12 risk classes, we accept them.

(d) We will perform two tests in order to check if there is statistical evidence that the classification into the geographic zones could be omitted. Note that in part (a) we saw that we tend to have considerably fewer claims for drivers in smaller towns and countryside than for drivers in middle-sized towns. The same holds true in a weakened form for middle-sized towns and large cities. Thus, we would expect that the classification into the three different geographic zone is reasonable. Now we will investigate this. To start with, note that the logarithmic probability that a Poisson random variable with frequency parameter α attains the value k, for some $k \in \mathbb{N}$, is equal to

$$\log\left(\exp\{-\alpha\}\frac{\alpha^k}{k!}\right) = -\alpha + k\log\alpha - \log k!.$$

Thus, defining

$$\hat{\boldsymbol{\lambda}}^{\text{MLE}} = \exp\left\{Z\hat{\boldsymbol{\beta}}^{\text{MLE}}\right\},$$

with $\hat{\boldsymbol{\lambda}}^{\text{MLE}} = \left(\hat{\lambda}_1^{\text{MLE}}, \dots, \hat{\lambda}_M^{\text{MLE}}\right)$, the joint log-likelihood function $l_{\mathbf{X}}$ of $\mathbf{X}$ at $\hat{\boldsymbol{\lambda}}^{\text{MLE}}$ is given by

$$l_{\mathbf{X}}\left(\hat{\boldsymbol{\lambda}}^{\text{MLE}}\right) = \sum_{m=1}^{M} -\hat{\lambda}_m^{\text{MLE}} v_m + X_m v_m \log\left(\hat{\lambda}_m^{\text{MLE}} v_m\right) - \log\left[(X_m v_m)!\right].$$

Therefore, we get for the scaled deviance statistics $D^*\left(\mathbf{X}, \hat{\boldsymbol{\lambda}}^{\text{MLE}}\right)$:

$$\begin{aligned} D^*\left(\mathbf{X}, \hat{\boldsymbol{\lambda}}^{\text{MLE}}\right) &= 2\left[l_{\mathbf{X}}(\mathbf{X}) - l_{\mathbf{X}}\left(\hat{\boldsymbol{\lambda}}^{\text{MLE}}\right)\right] \\ &= 2\sum_{m=1}^{M} -X_m v_m + X_m v_m \log X_m + \hat{\lambda}_m^{\text{MLE}} v_m - X_m v_m \log \hat{\lambda}_m^{\text{MLE}} \\ &= 2\sum_{m=1}^{M} v_m\left(X_m \log X_m - X_m - X_m \log \hat{\lambda}_m^{\text{MLE}} + \hat{\lambda}_m^{\text{MLE}}\right). \end{aligned}$$

Moreover, since for the Poisson case we have $\phi = 1$, the scaled deviance statistics $D^*\left(\mathbf{X}, \hat{\boldsymbol{\lambda}}^{\text{MLE}}\right)$ and the deviance statistics $D\left(\mathbf{X}, \hat{\boldsymbol{\lambda}}^{\text{MLE}}\right)$ are the same. Now, in order to check whether there is statistical evidence that the classification into the geographic zones could be omitted, we define the null hypothesis

$$H_0 : \beta_{3,2} = \beta_{3,3} = 0.$$

Thus, in the reduced model, we set the above $p = 2$ variables equal to 0. Then we can recalculate $\hat{\boldsymbol{\beta}}_{H_0}^{\text{MLE}}$ for this reduced model and define

$$\hat{\boldsymbol{\lambda}}_{H_0}^{\text{MLE}} = \exp\left\{Z_{H_0}\hat{\boldsymbol{\beta}}_{H_0}^{\text{MLE}}\right\},$$

where Z_{H_0} is the design matrix in the reduced model. According to formula (7.22) of the lecture notes, the test statistic

$$\begin{aligned} \text{F} &= \frac{D\left(\mathbf{X}, \hat{\boldsymbol{\lambda}}_{H_0}^{\text{MLE}}\right) - D\left(\mathbf{X}, \hat{\boldsymbol{\lambda}}^{\text{MLE}}\right)}{D\left(\mathbf{X}, \hat{\boldsymbol{\lambda}}^{\text{MLE}}\right)} \frac{M - r - 1}{p} \\ &= \frac{7}{2} \frac{D\left(\mathbf{X}, \hat{\boldsymbol{\lambda}}_{H_0}^{\text{MLE}}\right) - D\left(\mathbf{X}, \hat{\boldsymbol{\lambda}}^{\text{MLE}}\right)}{D\left(\mathbf{X}, \hat{\boldsymbol{\lambda}}^{\text{MLE}}\right)} \end{aligned}$$

has approximately an F-distribution with degrees of freedom given by $\text{df}_1 = p = 2$ and $\text{df}_2 = M - r - 1 = 7$. See the R-Code below for the calculation of F. We get

$$\text{F} \approx 51.239,$$

which corresponds to a p-value of approximately 0.0066%. Thus, we can reject H_0 at significance level of 5%. According to formula (7.23) of the lecture notes, a second test statistic is given by

$$X^2 = D^*\left(\mathbf{X}, \hat{\boldsymbol{\lambda}}_{H_0}^{\text{MLE}}\right) - D^*\left(\mathbf{X}, \hat{\boldsymbol{\lambda}}^{\text{MLE}}\right).$$

The test statistic X^2 has approximately a χ^2-distribution with $\text{df} = p = 2$ degrees of freedom. See the R-Code below for the calculation of X^2. We get

$$X^2 \approx 389.882,$$

which corresponds to a p-value of approximately $2.179 \cdot 10^{-85}$, which is basically 0. Thus, we can reject H_0 at significance level of 5%. Since we can reject H_0 using two different test statistics, we can conclude that there is no statistical evidence that the classification into different types of vehicles could be omitted.

```
### a)

### We perform a GLM analysis for the claim frequencies

### Determine the design matrix Z
class <- factor(c(rep(1,6),rep(2,6)))
age <- factor(c(rep(1,3),rep(2,3),rep(1,3),rep(2,3)))
zone <- factor(c(rep(1:3,4)))
counts <- c(25,15,15,60,90,210,45,45,30,80,120,90)
volumes <- c(1,2,5,4,9,70,2,3,6,8,15,50) * 100
Z <- model.matrix(counts ~ class + age + zone)

### Store the design matrix Z (without the intercept term), the counts and the volumes in one dataset
data <- cbind(Z[,-1],counts,volumes)
data <- as.data.frame(data)

### Apply GLM
d.glm <- glm(counts ~ class2 + age2 + zone2 + zone3, data=data, offset = log(volumes), family = poisson())
d.glm



### b)

### Fitted number of claims
fitted(d.glm)

### Store the features, the observed number of claims and the fitted numer of claims in one data set
data2 <- cbind(class, age, zone, volumes, counts, fitted(d.glm))
data2 <- as.data.frame(data2)
colnames(data2)[5:6] <- c("observed","fitted")

### Marginal claim frequencies for the two class categories
library(plyr)
class.comp <- ddply(data2, .(class), summarise, volumes = sum(volumes), observed = sum(observed), fitted = sum(fitted))
barplot(t(as.matrix(class.comp[,3:4]/class.comp[,2])), beside = TRUE, names.arg = c("weight > 60 kg, nr. of gears > 2", "other"), main = "claims frequencies (observed vs. fitted)", ylim = c(0,0.15), xlab = "vehicle class", ylab = "mean claim frequency",legend.text = TRUE)

### Marginal claim frequencies for the two age categories
```

```
39 age.comp <- ddply(data2, .(age), summarise, volumes = sum(
       volumes), observed = sum(observed), fitted = sum(fitted))
40 barplot(t(as.matrix(age.comp[,3:4]/age.comp[,2])), beside =
       TRUE, names.arg = c("at most one year", "more than one year"
       ), main = "claims frequencies (observed vs. fitted)",ylim =
       c(0,0.15), xlab = "vehicle age", ylab = "mean claim
       frequency",legend.text = TRUE)
41
42 ### Marginal claim frequencies for the three zone categories
43 zone.comp <- ddply(data2, .(zone), summarise, volumes = sum(
       volumes), observed = sum(observed), fitted = sum(fitted))
44 barplot(t(as.matrix(zone.comp[,3:4]/zone.comp[,2])), beside =
       TRUE, names.arg = c("large cities", "middle-sized towns", "
       smaller towns"), main = "claims frequencies (observed vs.
       fitted)",ylim = c(0,0.15), xlab = "geographic zone", ylab =
       "mean claim frequency",legend.text = TRUE)
45
46
47
48 ### c)
49
50 par(mfrow = c(1, 2))
51
52 ### Calculate the deviance residuals
53 dev.red <- sign(data2$observed - data2$fitted) * sqrt(2 * data2
       $observed*(-log(data2$fitted / data2$observed) + data2$
       fitted / data2$observed - 1))
54
55 ### Tukey-Anscombe plot
56 plot(data2$fitted, dev.red, main = "Tukey-Anscombe Plot", xlab
       = "fitted means", ylab = "deviance residuals", ylim = c
       (-3,3))
57 abline(h = 0,col = "red")
58
59 ### QQ plot
60 library(mgcv)
61 qq.gam(d.glm, type = "deviance",rep = 1, pch=19, main = "QQ
       Plot")
62
63
64
65 ### d)
66
67 ### Calculate the deviance statistics of the full model
68 X <- data2$observed / data2$volumes
69 lambda.full <- data2$fitted / data2$volumes
70 D.full <- 2 * sum(data2$volumes * (X * log(X) - X - X * log(
       lambda.full) + lambda.full))
71
72 ### Fit the reduced model
73 d.glm.2 <- glm(counts ~ class2 + age2, data=data, offset = log(
       volumes), family = poisson())
```

```
d.glm.2

### Calculate the deviance statistics of the reduced model
lambda.reduced <- fitted(d.glm.2) / data2$volumes
D.reduced <- 2 * sum(data2$volumes * (X * log(X) - X - X * log(
    lambda.reduced) + lambda.reduced))

### Calculate the test statistic F
F <- 7 / 2 * (D.reduced - D.full) / D.full

### Calculation of the corresponding p-value
pf(F, 2, 7, lower.tail = FALSE)

### Calculate the test statistic X^2
X.2 <- D.reduced - D.full

### Calculation of the corresponding p-value
pchisq(X.2, 2, lower.tail = FALSE)
```

Solution 10.3 Tweedie's Compound Poisson Model

(a) We can write S as

$$S = \sum_{i=1}^{N} Y_i,$$

where $N \sim \text{Poi}(\lambda v)$, $Y_1, Y_2, \dots \overset{\text{i.i.d.}}{\sim} G$ and N and $(Y_1, Y_2, \dots)$ are independent. Since G is the distribution function of a gamma distribution, we have $G(0) = 0$ and, thus,

$$\mathbb{P}[S = 0] = \mathbb{P}[N = 0] = \exp\{-\lambda v\}.$$

Let $x \in (0, \infty)$. Then the density f_S of S at x can be calculated as

$$f_S(x) = \frac{d}{dx} \mathbb{P}[S \le x],$$

where we have

$$\begin{aligned}
\mathbb{P}[S \le x] &= \sum_{n=0}^{\infty} \mathbb{P}[S \le x, N = n] \\
&= \sum_{n=0}^{\infty} \mathbb{P}[S \le x \,|\, N = n] \, \mathbb{P}[N = n] \\
&= \mathbb{P}[S \le x \,|\, N = 0] \, \mathbb{P}[N = 0] + \sum_{n=1}^{\infty} \mathbb{P}[S \le x \,|\, N = n] \, \mathbb{P}[N = n] \\
&= \mathbb{P}[N = 0] + \sum_{n=1}^{\infty} \mathbb{P}\left[\sum_{i=1}^{n} Y_i \le x\right] \mathbb{P}[N = n].
\end{aligned}$$

Since $Y_1, Y_2, \ldots \overset{\text{i.i.d.}}{\sim} \Gamma(\gamma, c)$, we get

$$\sum_{i=1}^{n} Y_i \sim \Gamma(n\gamma, c).$$

By writing f_n for the density function of $\Gamma(n\gamma, c)$, for all $n \in \mathbb{N}$, we get

$$\begin{aligned}
f_S(x) &= \frac{d}{dx}\left(\mathbb{P}[N=0] + \sum_{n=1}^{\infty} \mathbb{P}\left[\sum_{i=1}^{n} Y_i \le x\right] \mathbb{P}[N=n]\right) \\
&= \sum_{n=1}^{\infty} \frac{d}{dx}\, \mathbb{P}\left[\sum_{i=1}^{n} Y_i \le x\right] \mathbb{P}[N=n] \\
&= \sum_{n=1}^{\infty} f_n(x)\, \mathbb{P}[N=n] \\
&= \sum_{n=1}^{\infty} \frac{c^{n\gamma}}{\Gamma(n\gamma)} x^{n\gamma-1} \exp\{-cx\} \exp\{-\lambda v\} \frac{(\lambda v)^n}{n!} \\
&= \exp\{-(cx+\lambda v)\} \sum_{n=1}^{\infty} (\lambda v c^{\gamma})^n \frac{1}{\Gamma(n\gamma) n!} x^{n\gamma-1} \\
&= \exp\left\{-(cx+\lambda v) + \log\left[\sum_{n=1}^{\infty} (\lambda v c^{\gamma})^n \frac{1}{\Gamma(n\gamma) n!} x^{n\gamma-1}\right]\right\},
\end{aligned}$$

for all $x \in (0, \infty)$. Note that one can show that interchanging summation and differentiation above is indeed allowed. However, the proof is omitted here.

(b) Let $X \sim f_X$ belong to the exponential dispersion family with $w, \phi, \theta, b(\cdot)$ and $c(\cdot,\cdot,\cdot)$ as given on the exercise sheet. Then we have

$$\frac{x\theta}{\phi/w} = -xv \frac{(\gamma+1)\left(\frac{\lambda v \gamma}{c}\right)^{-\frac{1}{\gamma+1}}}{\frac{\gamma+1}{\lambda\gamma}\left(\frac{\lambda v\gamma}{c}\right)^{\frac{\gamma}{\gamma+1}}} = -x\lambda v\gamma \left(\frac{\lambda v\gamma}{c}\right)^{-1} = -cx,$$

for all $x \ge 0$, and

$$\frac{b(\theta)}{\phi/w} = v \frac{\frac{\gamma+1}{\gamma}\left(\frac{-\theta}{\gamma+1}\right)^{-\gamma}}{\frac{\gamma+1}{\lambda\gamma}\left(\frac{\lambda v\gamma}{c}\right)^{\frac{\gamma}{\gamma+1}}} = \lambda v \frac{\left(\frac{\lambda v\gamma}{c}\right)^{\frac{\gamma}{\gamma+1}}}{\left(\frac{\lambda v\gamma}{c}\right)^{\frac{\gamma}{\gamma+1}}} = \lambda v.$$

Moreover, since

$$\begin{aligned}
\frac{(\gamma+1)^{\gamma+1}}{\gamma}\left(\frac{\phi}{w}\right)^{-\gamma-1} &= \frac{(\gamma+1)^{\gamma+1}}{\gamma}\left[\frac{\gamma+1}{\lambda v\gamma}\left(\frac{\lambda v\gamma}{c}\right)^{\frac{\gamma}{\gamma+1}}\right]^{-\gamma-1} \\
&= \frac{1}{\gamma}(\lambda v\gamma)^{\gamma+1}\left(\frac{\lambda v\gamma}{c}\right)^{-\gamma} \\
&= \frac{1}{\gamma}\lambda v\gamma c^{\gamma} \\
&= \lambda v c^{\gamma},
\end{aligned}$$

we have

$$\begin{aligned} c(x,\phi,w) &= \log\left(\sum_{n=1}^{\infty}\left[\frac{(\gamma+1)^{\gamma+1}}{\gamma}\left(\frac{\phi}{w}\right)^{-\gamma-1}\right]^n \frac{1}{\Gamma(n\gamma)n!}\,x^{n\gamma-1}\right) \\ &= \log\left[\sum_{n=1}^{\infty}\left(\lambda v c^{\gamma}\right)^n \frac{1}{\Gamma(n\gamma)n!}\,x^{n\gamma-1}\right], \end{aligned}$$

for all $x > 0$. By putting together the above terms, we get

$$\begin{aligned} f_X(x;\theta,\phi) &= \exp\left\{\frac{x\theta - b(\theta)}{\phi/w} + c(x,\phi,w)\right\} \\ &= \exp\left\{-(cx+\lambda v) + \log\left[\sum_{n=1}^{\infty}(\lambda v c^{\gamma})^n \frac{1}{\Gamma(n\gamma)n!}x^{n\gamma-1}\right]\right\} \\ &= f_S(x), \end{aligned}$$

for all $x > 0$, and

$$f_X(0;\theta,\phi) = \exp\left\{\frac{0\cdot\theta - b(\theta)}{\phi/w} + c(0,\phi,w)\right\} = \exp\{-\lambda v\} = \mathbb{P}[S=0].$$

We conclude that S indeed belongs to the exponential dispersion family. Note that with this result at hand one might be tempted to estimate the shape parameter γ of the claim size distribution and then to do a GLM analysis directly on the compound claim size S. However, there are two reasons to rather perform a separate GLM analysis of the claim frequency and the claim severity instead: First, claim frequency modelling is usually more stable than claim severity modelling and often much of the differences between tariff cells are due to the claim frequency. Second, a separate analysis of the claim frequency and the claim severity allows more insight into the differences between the tariffs.