ETH Zurich
D-MATH
Dr. C. Buser
Prof. Dr. M. Wüthrich

August 21, 2018

Examination

Data Analytics for Non-Life Insurance Pricing

Please fill in the following table

Last name	
First name	
Programme of study	**MATH** □ **SAV** □ **Other** □
Matriculation number	

Leave blank

Question	**Maximum**	**Points**	**Check**
1	10		
2	10		
3	10		
4	10		
Total	**40**		

Instructions

Duration of exam: 120 min.

Closed book examination: no notes, no books, no calculator, no smartphones, etc., allowed.

Important:

- ◇ Please put your student card (or an identification card for SAV students) on the table.
- ◇ Only pen and paper are allowed on the table. Please do **not** write with a **pencil** or a **red** or **green** pen. Moreover, please do not use **whiteout**.
- ◇ Start by reading all questions and answer the ones which you think are easier first, before proceeding to the ones you expect to be more difficult. Do not spend too much time on one question but try to solve as many questions as possible.
- ◇ Take a new sheet for each question and write your name on every sheet.
- ◇ All results have to be **explained/argued** by indicating intermediate steps in the respective calculations. You can use known formulas from the lecture without derivation.
- ◇ Simplify your results as far as possible.
- ◇ Some of the subquestions can be solved independently of each other.

⋆⋆⋆ Good luck! ⋆⋆⋆

Question 1 (10 points)

Assume we have n observations given by

$$\mathcal{D} = \{(Y_1, \boldsymbol{x}_1), \ldots, (Y_n, \boldsymbol{x}_n)\}.$$

Assume that Y_i are independent and Bernoulli distributed for $i = 1, \ldots, n$ with

$$Y_i = \begin{cases} 1 & \text{with probability } p(\boldsymbol{x}_i), \\ 0 & \text{with probability } 1 - p(\boldsymbol{x}_i), \end{cases}$$

for a given (but unknown) regression function $p : \mathcal{X} \to (0,1)$.

(a) Choose a homogeneous regression function, i.e. $p(\boldsymbol{x}) \equiv p \in (0,1)$ for all $\boldsymbol{x} \in \mathcal{X}$. Give the resulting log-likelihood function and derive the maximum likelihood estimator $\widehat{p}$ for p.

(b) Calculate the resulting in-sample deviance statistics and give sufficient conditions for the observations $\mathcal{D}$ such that the resulting estimated distribution is non-degenerate.

(c) Assume that $\boldsymbol{x} \in \mathcal{X}$ is a one-dimensional continuous real-valued feature, i.e. $\mathcal{X} = \mathbb{R}$. Define a generalized linear model for the estimation of the regression function $p : \mathcal{X} \to (0,1)$ using 4 (non-empty) categorical classes. Calculate the resulting maximum likelihood estimator. *Hint:* Use for data compression in the categorical classes the property that the sum of i.i.d. Bernoulli distributed random variables provides a random variable with a well-known distribution function.

(d) Assume that $\boldsymbol{x} \in \mathcal{X}$ is a one-dimensional continuous real-valued feature, i.e. $\mathcal{X} = \mathbb{R}$. Define a generalized linear model for the estimation of the regression function $p : \mathcal{X} \to (0,1)$ directly using the continuous feature $\boldsymbol{x}$. Give the design matrix and calculate the resulting maximum likelihood estimator (as far as possible).

(e) Comparing the results of items (a), (c) and (d) we obtain the following in-sample losses and out-of-sample losses.

	in-sample loss	out-of-sample loss
(a) homogeneous model	0.2320	0.2360
(c) categorical feature	0.2050	0.2200
(d) continuous feature	0.2100	0.2180

Discuss the two error measures (in-sample loss and out-of-sample loss) and make a model choice (with justification).

Solution 1

(a) The likelihood function $L_{\mathcal{D}}(p)$ of the data $\mathcal{D}$ is given by

$$L_{\mathcal{D}}(p) \;=\; \prod_{i=1}^{n} p^{Y_i}(1-p)^{1-Y_i}.$$

Thus, for the log-likelihood we get

$$l_{\mathcal{D}}(p) \;\overset{\text{def}}{=}\; \log(L_{\mathcal{D}}(p)) \;=\; \sum_{i=1}^{n} Y_i \log(p) + (1-Y_i)\log(1-p).$$

In order to determine the maximum likelihood estimator $\widehat{p}$ for p, we take the derivative of $l_{\mathcal{D}}(p)$ with respect to p and set it equal to 0. We have

$$\frac{\partial l_{\mathcal{D}}(p)}{\partial p} \;=\; \sum_{i=1}^{n} Y_i \frac{1}{p} - (1-Y_i)\frac{1}{1-p},$$

which is equal to 0 if and only if

$$\sum_{i=1}^{n} Y_i(1-p) = \left(n - \sum_{i=1}^{n} Y_i\right) p \quad\Longleftrightarrow\quad p = \frac{1}{n}\sum_{i=1}^{n} Y_i.$$

For the second derivative of $l_{\mathcal{D}}(p)$ we get

$$\frac{\partial^2 l_{\mathcal{D}}(p)}{\partial p^2} \;=\; -\sum_{i=1}^{n} Y_i \frac{1}{p^2} - (1-Y_i)\frac{1}{(1-p)^2} < 0.$$

We conclude that the log-likelihood function is concave in p, and the maximum likelihood estimator $\widehat{p}$ is given by

$$\widehat{p} = \frac{1}{n}\sum_{i=1}^{n} Y_i.$$

(b) By definition, the resulting (scaled) deviance statistics is obtained by twice the difference between the log-likelihood of the saturated model and the log-likelihood of the considered model. In the saturated model we have one parameter (p_i) per observation (i) and look for the MLE. By similar arguments as shown in a), in the saturated model we get $p_i = Y_i$, for all $i = 1, \ldots, n$. We write $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$. The log-likelihood of the saturated model is then given by

$$l_{\mathcal{D}}(\boldsymbol{Y}) = \sum_{i=1}^{n} Y_i \log(Y_i) + (1-Y_i)\log(1-Y_i) = 0,$$

where we define $Y_i \log(Y_i) = 0$ if $Y_i = 0$ and $(1-Y_i)\log(1-Y_i) = 0$ if $Y_i = 1$. For the (scaled) deviance statistics we then have

$$D^*(\boldsymbol{Y}, \widehat{p}) = 2(l_{\mathcal{D}}(\boldsymbol{Y}) - l_{\mathcal{D}}(\widehat{p})) = -2 l_{\mathcal{D}}(\widehat{p}) = -2\sum_{i=1}^{n} Y_i \log(\widehat{p}) + (1-Y_i)\log(1-\widehat{p}).$$

This deviance statistics is non-degenerate if and only if not all of the Y_i's are equal to the same value. Equivalently, if and only if there exist $i \neq j \in \{1, \ldots, n\}$ with $Y_i \neq Y_j$.

(c) We define a partition of $\mathbb{R}$ into 4 disjoint intervals I_1, I_2, I_3 and I_4, such that for all $j = 1, \dots, 4$ there exists an $i \in \{1, \dots, n\}$ with $x_i \in I_j$, i.e. we have at least one observation in every interval. We define, for all $j = 1, \dots, 4$,

$$v_j = \sum_{i=1}^{n} 1_{\{x_i \in I_j\}}$$

and

$$N_j = \sum_{i=1}^{n} 1_{\{x_i \in I_j\}} Y_i.$$

The quantities $v_1, \dots, v_4$ describe the volume of the four intervals and $N_1, \dots, N_4$ the sum of the Bernoulli successes in each interval. Let $p_1, \dots, p_4$ denote the unknown success parameters of the four categorical classes. As we have independent observations and the sum of independent Bernoulli random variables with the same success parameter follows a binomial distribution, we get the model

$$N_j \sim \mathrm{Bin}(v_j, p_j).$$

For each class we now have a logistic regression model. As the features consist only of the class affiliation, we model

$$p_j(\beta_j) = \frac{e^{\beta_j}}{1 + e^{\beta_j}},$$

for all $j = 1, \dots, 4$, where the parameter $\boldsymbol{\beta} = (\beta_1, \dots, \beta_4)$ has to be estimated. We note that

$$\frac{\partial p_j(\beta_j)}{\partial \beta_j} = \frac{e^{\beta_j}(1 + e^{\beta_j}) - e^{\beta_j} e^{\beta_j}}{(1 + e^{\beta_j})^2} = \frac{e^{\beta_j}}{(1 + e^{\beta_j})^2} = p_j(\beta_j)(1 - p_j(\beta_j)) > 0.$$

For the likelihood function $L_{\mathcal{D}}(\boldsymbol{\beta})$ (with respect to the unknown parameter $\boldsymbol{\beta}$) we have

$$L_{\mathcal{D}}(\boldsymbol{\beta}) = \prod_{j=1}^{4} \binom{v_j}{N_j} p_j(\beta_j)^{N_j} (1 - p_j(\beta_j))^{v_j - N_j}.$$

The log-likelihood function $l_{\mathcal{D}}(\boldsymbol{\beta})$ is then given by

$$l_{\mathcal{D}}(\boldsymbol{\beta}) = \log(L_{\mathcal{D}}(\boldsymbol{\beta})) = \sum_{j=1}^{4} \log\left(\binom{v_j}{N_j}\right) + N_j \log(p_j(\beta_j)) + (v_j - N_j) \log(1 - p_j(\beta_j)).$$

Similarly as in (a), we calculate the maximum likelihood estimator by setting the derivative of the log-likelihood function to 0. We have, for all $j = 1, \dots, 4$,

$$\begin{aligned}\frac{\partial l_{\mathcal{D}}(\boldsymbol{\beta})}{\partial \beta_j} &= N_j \frac{1}{p_j(\beta_j)} \frac{\partial p_j(\beta_j)}{\partial \beta_j} - (v_j - N_j) \frac{1}{1 - p_j(\beta_j)} \frac{\partial p_j(\beta_j)}{\partial \beta_j} \\ &= N_j(1 - p_j(\beta_j)) - (v_j - N_j) p_j(\beta_j).\end{aligned}$$

For all $j = 1, \dots, 4$, this is equal to 0 if and only if

$$p_j(\beta_j) = \frac{N_j}{v_j}.$$

For the second derivative of $l_{\mathcal{D}}(\boldsymbol{\beta})$ we get, for all $j = 1, \dots, 4$,

$$\frac{\partial^2 l_{\mathcal{D}}(\boldsymbol{\beta})}{\partial \beta_j^2} = -N_j \frac{\partial p_j(\beta_j)}{\partial \beta_j} - (v_j - N_j) \frac{\partial p_j(\beta_j)}{\partial \beta_j} < 0,$$

as at least one of the two terms N_j and $v_j - N_j$ is strictly positive. For all $j = 1, \ldots, 4$, we conclude that the log-likelihood function is concave in β_j, and the maximum likelihood estimator $\widehat{p}_j(\beta_j)$ is given by

$$\widehat{p}_j(\beta_j) = \frac{N_j}{v_j}.$$

Note that the logistic regression model is not necessary here, i.e. we can directly estimate p_j by $\widehat{p}_j = N_j/v_j$.

(d) We use a logistic regression approach and model

$$p_{\boldsymbol{\beta}}(x) = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}},$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1)$ is the unknown model parameter. We get the design matrix

$$X = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}.$$

The likelihood function $L_{\mathcal{D}}(\boldsymbol{\beta})$ of the data $\mathcal{D}$ is given by

$$L_{\mathcal{D}}(\boldsymbol{\beta}) \;=\; \prod_{i=1}^{n} p_{\boldsymbol{\beta}}(x_i)^{Y_i}(1-p_{\boldsymbol{\beta}}(x_i))^{1-Y_i}.$$

Thus, for the log-likelihood we get

$$l_{\mathcal{D}}(\boldsymbol{\beta}) \;=\; \log(L_{\mathcal{D}}(\boldsymbol{\beta})) \;=\; \sum_{i=1}^{n} Y_i \log(p_{\boldsymbol{\beta}}(x_i)) + (1-Y_i)\log(1-p_{\boldsymbol{\beta}}(x_i)).$$

Again, we calculate the maximum likelihood estimator by setting the derivative of the log-likelihood function to 0. We have, for $j = 0, 1$,

$$\frac{\partial l_{\mathcal{D}}(\boldsymbol{\beta})}{\partial \beta_j} = \sum_{i=1}^{n} \left(\frac{Y_i}{p_{\boldsymbol{\beta}}(x_i)} - \frac{1-Y_i}{1-p_{\boldsymbol{\beta}}(x_i)} \right) \frac{\partial p_{\boldsymbol{\beta}}(x_i)}{\partial \beta_j}.$$

For $j = 0$ we get

$$\frac{\partial p_{\boldsymbol{\beta}}(x_i)}{\partial \beta_0} = \frac{e^{\beta_0+\beta_1 x_i}(1+e^{\beta_0+\beta_1 x_i}) - (e^{\beta_0+\beta_1 x_i})^2}{(1+e^{\beta_0+\beta_1 x_i})^2} = p_{\boldsymbol{\beta}}(x_i)(1-p_{\boldsymbol{\beta}}(x_i)),$$

for all $i = 1, \ldots, n$. Thus, we get

$$\frac{\partial l_{\mathcal{D}}(\boldsymbol{\beta})}{\partial \beta_0} = \sum_{i=1}^{n} Y_i(1-p_{\boldsymbol{\beta}}(x_i)) - (1-Y_i)p_{\boldsymbol{\beta}}(x_i) > 0.$$

This is equal to 0 if and only if

$$\sum_{i=1}^{n} Y_i = \sum_{i=1}^{n} p_{\boldsymbol{\beta}}(x_i). \tag{1}$$

For the second derivative of $l_{\mathcal{D}}(\boldsymbol{\beta})$ with respect to β_0 we get

$$\frac{\partial^2 l_{\mathcal{D}}(\boldsymbol{\beta})}{\partial \beta_0^2} = \sum_{i=1}^{n} -Y_i \frac{\partial p_{\boldsymbol{\beta}}(x_i)}{\partial \beta_0} - (1-Y_i)\frac{\partial p_{\boldsymbol{\beta}}(x_i)}{\partial \beta_0} < 0.$$

For $j = 1$ we get

$$\frac{\partial p_{\boldsymbol{\beta}}(x_i)}{\partial \beta_1} = \frac{e^{\beta_0+\beta_1 x_i} x_i (1 + e^{\beta_0+\beta_1 x_i}) - (e^{\beta_0+\beta_1 x_i})^2 x_i}{(1 + e^{\beta_0+\beta_1 x_i})^2} = p_{\boldsymbol{\beta}}(x_i)(1 - p_{\boldsymbol{\beta}}(x_i)) x_i,$$

for all $i = 1, \dots, n$. Thus, we get

$$\frac{\partial l_{\mathcal{D}}(\boldsymbol{\beta})}{\partial \beta_1} = \sum_{i=1}^{n} Y_i (1 - p_{\boldsymbol{\beta}}(x_i)) x_i - (1 - Y_i) p_{\boldsymbol{\beta}}(x_i) x_i.$$

This is equal to 0 if and only if

$$\sum_{i=1}^{n} Y_i x_i = \sum_{i=1}^{n} p_{\boldsymbol{\beta}}(x_i) x_i. \tag{2}$$

For the second derivative of $l_{\mathcal{D}}(\boldsymbol{\beta})$ with respect to β_1 we get

$$\begin{aligned}
\frac{\partial^2 l_{\mathcal{D}}(\boldsymbol{\beta})}{\partial \beta_1^2} &= \sum_{i=1}^{n} -Y_i x_i \frac{\partial p_{\boldsymbol{\beta}}(x_i)}{\partial \beta_1} - (1 - Y_i) x_i \frac{\partial p_{\boldsymbol{\beta}}(x_i)}{\partial \beta_1} \\
&= \sum_{i=1}^{n} -Y_i x_i^2 p_{\boldsymbol{\beta}}(x_i)(1 - p_{\boldsymbol{\beta}}(x_i)) - (1 - Y_i) x_i^2 p_{\boldsymbol{\beta}}(x_i)(1 - p_{\boldsymbol{\beta}}(x_i)) < 0.
\end{aligned}$$

We conclude that the maximum likelihood estimator $\widehat{\boldsymbol{\beta}}$ for $\boldsymbol{\beta}$ is the solution to equations (1) and (2), which have to be solved numerically.

(e) The in-sample loss is the loss obtained on the training sample that is used to estimate the parameters. Minimizing this error can lead to overfitting especially if our considered parametric model is too flexible. Therefore, the quality of the model should be evaluated on a test data set (out-of-sample loss) that has not been used for model estimation.

In our case we prefer the continuous feature because the corresponding model has the smallest out-of-sample loss (0.2180). Moreover, the continuous feature model has less parameters than the categorical one, therefore we also expect a lower parameter estimation uncertainty in the former.

Question 2 (10 points)

Assume we have n observations given by

$$\mathcal{D} = \{(Y_1, \boldsymbol{x}_1), \ldots, (Y_n, \boldsymbol{x}_n)\}.$$

Assume that Y_i are independent and Bernoulli distributed for $i = 1, \ldots, n$ with

$$Y_i = \begin{cases} 1 & \text{with probability } p, \\ 0 & \text{with probability } 1 - p, \end{cases}$$

for a given (but unknown) parameter $p \in (0, 1)$.

(a) Define a Bayesian Bernoulli model for the estimation of the unknown parameter $p \in (0, 1)$ using a non-degenerate prior distribution.
Hint: The Beta distribution has density supported on $(0, 1)$ given by

$$\pi(y) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} y^{\alpha-1}(1 - y)^{\beta-1}, \qquad \text{for } y \in (0, 1),$$

and given parameters $\alpha, \beta > 0$. The corresponding mean and variance are given by $\alpha/(\alpha + \beta)$ and $\alpha\beta/((\alpha + \beta)^2(\alpha + \beta + 1))$, respectively.

(b) Calculate the posterior estimator $\widehat{p}^{\text{post}}$ for p, given data $\mathcal{D}$, under the Bayesian model assumptions made in item (a) (using π as prior density).

(c) Give a credibility theory interpretation of the posterior estimator $\widehat{p}^{\text{post}}$ derived in the previous item. Can the posterior estimator lead to a degenerate probability model under the above model assumptions (give an argument for your answer)?

(d) Derive the (conditional) mean square error of prediction of $\widehat{p}^{\text{post}}$ derived under item (b). What happens with this error if $n \to \infty$?

(e) Explain why this Bayesian Bernoulli model can be useful in regression tree constructions.

Solution 2

(a) In a Bayesian Bernoulli model, we assume that the parameter p is a random variable whose density π is supported on a subset of $(0,1)$. Additionally, we assume that

$$Y_i \,|\, p \sim \text{Bernoulli}(p)$$

for all $i = 1, \dots, n$ and that, conditionally on p, the random variables Y_i and Y_j are independent for all $i \neq j$.

Using the definition of conditional density (Bayes' theorem) and our assumptions, the joint distribution of the data $\boldsymbol{Y} = (Y_1, \dots, Y_n)'$ and the parameter p is given by the density

$$f(\boldsymbol{Y}, p) = f(\boldsymbol{Y} \,|\, p)\pi(p) = \left(\prod_{i=1}^n f(Y_i \,|\, p) \right) \pi(p),$$

where $f(\boldsymbol{Y} \,|\, p)$ denotes the conditional probability mass function of $\boldsymbol{Y}$ given p and, analogously, $f(Y_i \,|\, p)$ denotes the conditional probability mass function of Y_i given p. The posterior distribution of p is then the distribution of p given the data $\boldsymbol{Y}$ and is given by the density

$$f(p \,|\, \boldsymbol{Y}) = \frac{f(\boldsymbol{Y}, p)}{f(\boldsymbol{Y})} \propto f(\boldsymbol{Y} \,|\, p)\pi(p).$$

(b) In order to identify the posterior distribution, we select for the prior distribution of p the Beta distribution given in the hint of (a). In that case, the joint distribution of the data $\boldsymbol{Y}$ and the parameter p is given by

$$\begin{aligned} f(p \,|\, \boldsymbol{Y}) &\propto \left(\prod_{i=1}^n p^{Y_i}(1-p)^{1-Y_i} \right) \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1} \\ &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha+\sum_{i=1}^n Y_i - 1}(1-p)^{\beta+n-\sum_{i=1}^n Y_i - 1}. \end{aligned}$$

So,

$$f(p \,|\, \boldsymbol{Y}) \propto p^{\alpha+\sum_{i=1}^n Y_i - 1}(1-p)^{\beta+n-\sum_{i=1}^n Y_i - 1},$$

which is the unnormalized density of the Beta distribution with parameters

$$\hat{\alpha}^{\text{post}} = \alpha + \sum_{i=1}^n Y_i \quad \text{and} \quad \hat{\beta}^{\text{post}} = \beta + n - \sum_{i=1}^n Y_i.$$

Since we have $\hat{p}^{\text{post}} = \mathbb{E}[p \,|\, \boldsymbol{Y}]$, using again the hint from (a), we obtain

$$\hat{p}^{\text{post}} = \frac{\hat{\alpha}^{\text{post}}}{\hat{\alpha}^{\text{post}} + \hat{\beta}^{\text{post}}} = \frac{\alpha + \sum_{i=1}^n Y_i}{\alpha+\beta+n}.$$

(c) We can write

$$\hat{p}^{\text{post}} = \frac{\alpha + \sum_{i=1}^n Y_i}{\alpha+\beta+n} = \frac{\alpha+\beta}{\alpha+\beta+n}\,\frac{\alpha}{\alpha+\beta} + \frac{n}{\alpha+\beta+n}\,\frac{\sum_{i=1}^n Y_i}{n} = (1-w)\,p_0 + w\,\hat{p},$$

where p_0 is the mean of the prior distribution π, $\hat{p}$ is the MLE from Question 1 (a) and w is the credibility weight given by

$$w = \frac{n}{\alpha+\beta+n} \in (0,1).$$

Using the above fact that $\hat{p}^{\text{post}}$ can be expressed as a weighted average of the prior and sample mean (MLE) and that $p_0 \in (0,1)$ for all $\alpha, \beta > 0$, we see that the posterior estimator $\hat{p}^{\text{post}}$ never leads to a degenerate model for any $n \in \mathbb{N}$.

(d) We have that

$$\mathrm{MSE}\left(\hat{p}^{\text{post}} \,|\, \boldsymbol{Y}\right) = \mathbb{E}\left[(\hat{p}^{\text{post}} - p)^2 \,\middle|\, \boldsymbol{Y}\right] = \mathbb{E}\left[(\mathbb{E}\left[p \,|\, \boldsymbol{Y}\right] - p)^2 \,\middle|\, \boldsymbol{Y}\right] = \mathrm{Var}\left(p \,|\, \boldsymbol{Y}\right).$$

Using again the hint from (a), we obtain

$$\begin{aligned}\mathrm{MSE}\left(\hat{p}^{\text{post}} \,|\, \boldsymbol{Y}\right) &= \frac{\hat{\alpha}^{\text{post}}\hat{\beta}^{\text{post}}}{(\hat{\alpha}^{\text{post}} + \hat{\beta}^{\text{post}})^2(\hat{\alpha}^{\text{post}} + \hat{\beta}^{\text{post}} + 1)} = \frac{\hat{\alpha}^{\text{post}}\hat{\beta}^{\text{post}}}{(\hat{\alpha}^{\text{post}} + \hat{\beta}^{\text{post}})^2} \frac{1}{\alpha + \beta + n + 1} \\ &= \frac{\hat{\alpha}^{\text{post}}}{(\hat{\alpha}^{\text{post}} + \hat{\beta}^{\text{post}})} \frac{\hat{\beta}^{\text{post}}}{(\hat{\alpha}^{\text{post}} + \hat{\beta}^{\text{post}})} \frac{1}{\alpha + \beta + n + 1} \\ &= \frac{1}{\alpha + \beta + n + 1} \hat{p}^{\text{post}}(1 - \hat{p}^{\text{post}}).\end{aligned}$$

Now, since $\hat{p}^{\text{post}} \in (0,1)$, we clearly have that $\mathrm{MSE}\left(\hat{p}^{\text{post}} \,|\, \boldsymbol{Y}\right) \to 0$ as $n \to \infty$.

(e) A general issue that might occur in insurance claims frequency modeling is that in a certain node $\mathcal{X}_t$ of a regression tree we only have observations $\boldsymbol{x}_i$ with $Y_i = 0$ (or with $Y_i = 1$). In such a case we would obtain a degenerate model on $\mathcal{X}_t$ with maximum likelihood estimator $\widehat{p} = 0$ (or $\hat{p} = 1$, respectively). With the Bayesian Bernoulli model we get an estimator which is never degenerate, as shown in the solution to (c).

Question 3 (10 points)

Assume we have n large claims given by

$$\mathcal{D} = \{(Y_1, \boldsymbol{x}_1), \dots, (Y_n, \boldsymbol{x}_n)\}.$$

Assume that $\boldsymbol{x}_i \in \mathcal{X} = \mathbb{R}$, and that Y_i are independent and Pareto distributed for $i = 1, \dots, n$ with density supported in $[M, \infty)$ and given by

$$Y_i \sim f(y|\boldsymbol{x}_i) = \frac{\alpha(\boldsymbol{x}_i)}{M} \left(\frac{y}{M}\right)^{-\alpha(\boldsymbol{x}_i)-1}, \qquad \text{for } y \geq M,$$

for a given (known) large claims threshold $M > 0$ and a given (but unknown) regression function $\alpha : \mathcal{X} \to \mathbb{R}_+$.

(a) Calculate the deviance statistics for this problem.

(b) Set up a single hidden layer neural network with more than two hidden neurons for this regression problem using the sigmoid activation function. How many parameters does the model have?

(c) Calculate one step of the gradient descent optimization algorithm explicitly for the deviance statistics loss function derived in item (a) and the single hidden layer neural network defined in item (b). Explain why the gradient descent method is of interest in neural network calibrations.

(d) Assume we have a large number of hidden neurons (say more than 100). Why are we in this situation in general not interested in finding the maximum likelihood estimator? What alternative solution do you propose?

(e) Assume we have feature space $\mathcal{X} = [-1, 1]^2$. Compare a single hidden layer neural network with 3 hidden neurons and step function activation to a gradient boosting machine, where for the latter we use single split regression trees for totally 3 boosting steps. Which of the two models has the smaller optimal in-sample loss (give an argument for your answer)? Which of the two models has the smaller out-of-sample loss (give an argument for your answer)?

Solution 3

(a) The likelihood function $L_{\mathcal{D}}(\alpha(\cdot))$ of the data $\mathcal{D}$ is given by

$$L_{\mathcal{D}}(\alpha(\cdot)) \;=\; \prod_{i=1}^{n} \frac{\alpha(x_i)}{M} \left(\frac{Y_i}{M}\right)^{-\alpha(x_i)-1}.$$

Thus, for the log-likelihood we get

$$l_{\mathcal{D}}(\alpha(\cdot)) \;=\; \log(L_{\mathcal{D}}(\alpha(\cdot))) \;=\; \sum_{i=1}^{n} \log(\alpha(x_i)) - \log(M) - (\alpha(x_i)+1)\log\left(\frac{Y_i}{M}\right).$$

In the saturated model we have one parameter (α_i) per observation (i). That is, we have to maximize

$$g(\alpha_i) \overset{\text{def}}{=} \log(\alpha_i) - \log(M) - (\alpha_i+1)\log\left(\frac{Y_i}{M}\right),$$

with respect to α_i, for all $i = 1, \ldots, n$. If we take the derivative with respect to α_i, we get

$$\frac{\partial g(\alpha_i)}{\partial \alpha_i} = \frac{1}{\alpha_i} - \log\left(\frac{Y_i}{M}\right),$$

for all $i = 1, \ldots, n$. This is equal to 0 if and only if

$$\alpha_i = \frac{1}{\log\left(\frac{Y_i}{M}\right)}, \tag{3}$$

for all $i = 1, \ldots, n$. For the second derivative of $g(\alpha_i)$ with respect to α_i we get

$$\frac{\partial^2 g(\alpha_i)}{\partial \alpha_i^2} = -\frac{1}{\alpha_i^2} < 0,$$

for all $i = 1, \ldots, n$. That is, in the saturated model we have the parameter $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_n)$ with α_i given as in (3), for all $i = 1, \ldots, n$. For the log-likelihood of the saturated model we then have

$$\begin{aligned} l_{\mathcal{D}}(\boldsymbol{\alpha}) \;&=\; \sum_{i=1}^{n} \log\left(\frac{1}{\log\left(\frac{Y_i}{M}\right)}\right) - \log(M) - \left(\frac{1}{\log\left(\frac{Y_i}{M}\right)} + 1\right)\log\left(\frac{Y_i}{M}\right) \\ &=\; \sum_{i=1}^{n} -\log\left(\log\left(\frac{Y_i}{M}\right)\right) - \log(M) - 1 - \log\left(\frac{Y_i}{M}\right). \end{aligned}$$

Finally, the (scaled) deviance statistics is given by

$$\begin{aligned} D^*(\boldsymbol{\alpha}, \alpha(\cdot)) &= 2(l_{\mathcal{D}}(\boldsymbol{\alpha}) - l_{\mathcal{D}}(\alpha(\cdot))) \\ &= 2\sum_{i=1}^{n} -\log\left(\log\left(\frac{Y_i}{M}\right)\right) - 1 - \log(\alpha(x_i)) + \alpha(x_i)\log\left(\frac{Y_i}{M}\right). \end{aligned}$$

(b) We choose a single hidden layer neural network with q hidden neurons. As our feature space is $\mathcal{X} = \mathbb{R}$, we have only one neuron in the input layer. The sigmoid activation function (on $\mathbb{R}$) is given by

$$\phi(x) = \frac{1}{1+e^{-x}}.$$

We then have the activations, for all $j = 1, \ldots, q$,

$$z_j(x) = \phi(w_{j,0} + w_{j,1}x),$$

with unknown parameters $w_{j,0}, w_{j,1} \in \mathbb{R}$, for the q neurons in the hidden layer. Since the codomain of $\alpha(\cdot)$ has to be the positive real line, we define a log-linear regression approach as follows

$$\alpha(x) = e^{\beta_0 + \sum_{j=1}^q \beta_j z_j(x)},$$

with unknown parameters $\beta_0, \beta_1, \ldots, \beta_q \in \mathbb{R}$. Overall, we have

$$(1+1)q + (q+1) = 3q+1$$

parameters in the model.

(c) We note that for the derivative of the sigmoid activation function ϕ we have

$$\frac{\partial \phi(x)}{\partial x} = \frac{e^{-x}}{(1+e^{-x})^2} = \phi(x)(1-\phi(x)).$$

We write

$$\boldsymbol{\theta} = (w_{1,0}, w_{1,1}, \ldots, w_{q,0}, w_{q,1}, \beta_0, \beta_1, \ldots, \beta_q) \in \mathbb{R}^{3q+1}$$

for the vector of the unknown model parameters. Thus, the regression function $\alpha_{\boldsymbol{\theta}}(\cdot)$ depends on $\boldsymbol{\theta}$. In the gradient descent optimization algorithm the goal is to decrease a given loss function by iteratively updating the model parameters. In our case we would like to decrease the deviance statistics

$$D^*(\boldsymbol{\alpha}, \alpha_{\boldsymbol{\theta}}(\cdot)) = 2\sum_{i=1}^n -\log\left(\log\left(\frac{Y_i}{M}\right)\right) - 1 - \log(\alpha_{\boldsymbol{\theta}}(x_i)) + \alpha_{\boldsymbol{\theta}}(x_i)\log\left(\frac{Y_i}{M}\right).$$

To this end, for a given $\boldsymbol{\theta}$, we move in the direction of the maximal local decrease of the deviance statistics, i.e. in the direction of the negative gradient $\nabla_{\boldsymbol{\theta}} D^*(\boldsymbol{\alpha}, \alpha_{\boldsymbol{\theta}}(\cdot))$ of the deviance statistics. We calculate

$$\nabla_{\boldsymbol{\theta}} D^*(\boldsymbol{\alpha}, \alpha_{\boldsymbol{\theta}}(\cdot)) = \frac{\partial D^*(\boldsymbol{\alpha}, \alpha_{\boldsymbol{\theta}}(\cdot))}{\partial \boldsymbol{\theta}} = 2\sum_{i=1}^n \left(-\frac{1}{\alpha_{\boldsymbol{\theta}}(x_i)} + \log\left(\frac{Y_i}{M}\right)\right)\frac{\partial \alpha_{\boldsymbol{\theta}}(x_i)}{\partial \boldsymbol{\theta}},$$

where we have

$$\begin{aligned}
\frac{\partial \alpha_{\boldsymbol{\theta}}(x_i)}{\partial w_{j,0}} &= \alpha_{\boldsymbol{\theta}}(x_i)\beta_j z_j(x_i)(1-z_j(x_i)), \\
\frac{\partial \alpha_{\boldsymbol{\theta}}(x_i)}{\partial w_{j,1}} &= \alpha_{\boldsymbol{\theta}}(x_i)\beta_j z_j(x_i)(1-z_j(x_i))x_i, \\
\frac{\partial \alpha_{\boldsymbol{\theta}}(x_i)}{\partial \beta_0} &= \alpha_{\boldsymbol{\theta}}(x_i), \\
\frac{\partial \alpha_{\boldsymbol{\theta}}(x_i)}{\partial \beta_j} &= \alpha_{\boldsymbol{\theta}}(x_i)z_j(x_i),
\end{aligned}$$

for all $i = 1, \ldots, n$ and $j = 1, \ldots, q$. In one single step of the gradient descent optimization algorithm we have the update

$$\boldsymbol{\theta} \longrightarrow \boldsymbol{\theta} - \rho\nabla_{\boldsymbol{\theta}} D^*(\boldsymbol{\alpha}, \alpha_{\boldsymbol{\theta}}(\cdot)),$$

where $\rho > 0$ is the so-called learning rate. The gradient descent method is of interest in neural network calibrations because it allows us to reduce the deviance statistics and in this way to get a better fit of the model to the data. Note that one should carefully choose an appropriate stopping time of the algorithm in order to prevent from overfitting; and one should also carefully choose $\rho > 0$ because the gradient descent steps lead to a decrease locally.

(d) A neural network model with a large number of hidden neurons is heavily over-parametrized. Therefore, a maximum likelihood estimator would lead to overfitting of the model to the training dataset. Thus, we are only interested in finding a sufficiently good approximation which has also a good out-of-sample performance. We believe that such a 'good' parametrization can be reached for example by the gradient descent method.

(e) With a single hidden layer neural network with 3 hidden neurons and step function activation we can split the feature space $\mathcal{X} = [-1, 1]^2$ using three hyperplanes. This allows us to assign a different value to at most 7 (disjoint) subsets of $\mathcal{X}$. Using a boosting machine consisting of single split regression trees for totally 3 boosting steps, we can split the feature space $\mathcal{X}$ in at most 6 (disjoint) subsets of $\mathcal{X}$. We conclude that with the neural network we have more degrees of freedom in this example, which implies a lower in-sample loss for the neural network. (However, we note that in a neural network we have to estimate the network parameter using an optimization technique like gradient descent. These techniques heavily depend on the stopping time and the starting value. Thus, in reality, one might get a neural network model with a higher in-sample loss.)

With regard to the out-of-sample loss we cannot make any statement. There isn't any general rule because out-of-sample everything is possible.

Question 4 (10 points)

Assume we have n independently distributed claims count observations given by the data

$$\mathcal{D} = \{(N_1, \boldsymbol{x}_1), \dots, (N_n, \boldsymbol{x}_n)\}.$$

An actuary wants to have your opinion based on the following output. Take your decisions on a test level of $\alpha = 5\%$.

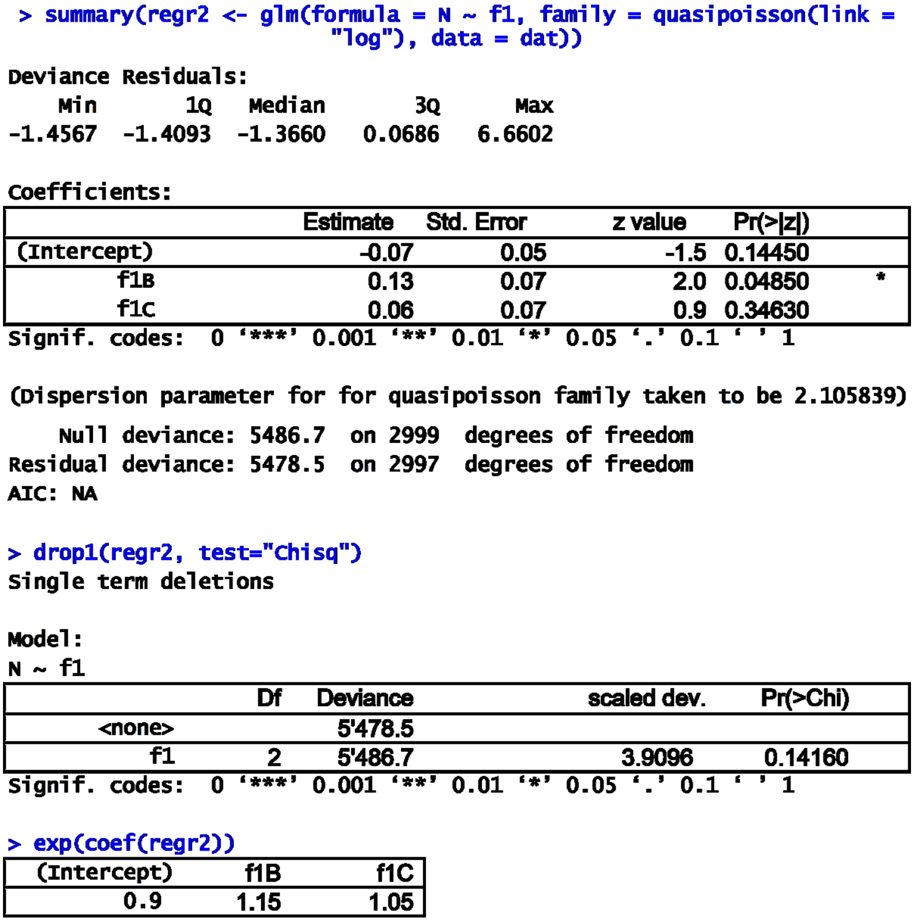
(a) Define an appropriate generalized linear model for claims frequency modeling based on the given data $\mathcal{D}$. What conditions need to be fulfilled so that the model can be applied?

(b) The actuary gives you the following R output of his analysis. Answer the following questions based on his output:

(i) How many observations do we have?

(ii) How many explanatory variables are available and what structure do they have?

(iii) Based on the output below: which variables have a significant relationship with the observed claims frequency? Give statistical arguments for your statements.



```
 > summary(regr2 <- glm(formula = N ~ f1, family = quasipoisson(link =
                         "log"), data = dat))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.4567  -1.4093  -1.3660   0.0686   6.6602

Coefficients:
```

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-0.07	0.05	-1.5	0.14450	
f1B	0.13	0.07	2.0	0.04850	*
f1C	0.06	0.07	0.9	0.34630	

```
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for for quasipoisson family taken to be 2.105839)

    Null deviance: 5486.7  on 2999  degrees of freedom
Residual deviance: 5478.5  on 2997  degrees of freedom
AIC: NA

> drop1(regr2, test="Chisq")
Single term deletions

Model:
N ~ f1
```

	Df	Deviance	scaled dev.	Pr(>Chi)
<none>		5'478.5		
f1	2	5'486.7	3.9096	0.14160

```
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> exp(coef(regr2))
```

(Intercept)	f1B	f1C
0.9	1.15	1.05

(c) Assume that you have decided to keep variable f1 in the model. Give the resulting prediction for the claims frequencies of the different policies.

(d) You intend to improve your existing generalized linear model and want to keep the interpretability at the same time.

(i) What could you do to improve the prediction of your existing model?

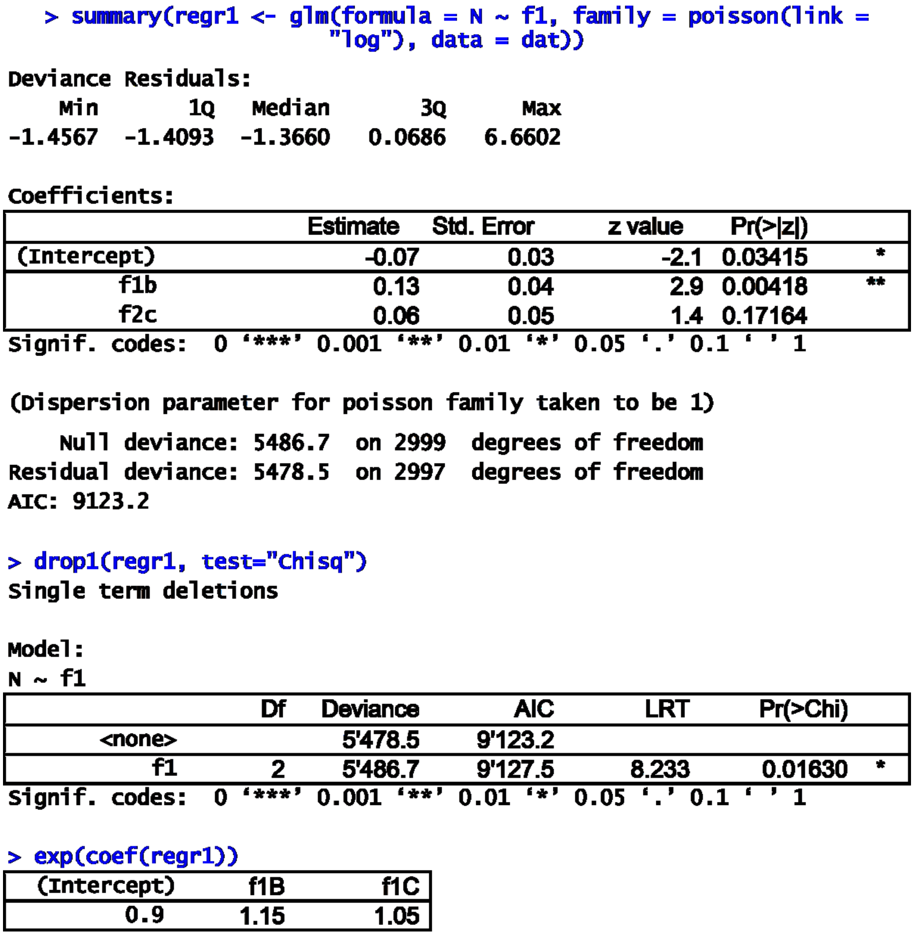
(ii) How would you compare different models to check which model performs better?

(e) Consider the following output below and compare it to the output from item (b).

(i) What are the differences between the two models?

(ii) Would you revise one or more statements that you have taken in item (b) based on this new output? Give arguments for your statements.

(iii) Which model fits better to the data? Give arguments for your statements.



```
   > summary(regr1 <- glm(formula = N ~ f1, family = poisson(link =
                         "log"), data = dat))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.4567  -1.4093  -1.3660   0.0686    6.6602

Coefficients:
                         Estimate   Std. Error     z value    Pr(>|z|)
(Intercept)                 -0.07         0.03        -2.1     0.03415     *
        f1b                  0.13         0.04         2.9     0.00418    **
        f2c                  0.06         0.05         1.4     0.17164
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 5486.7  on 2999  degrees of freedom
Residual deviance: 5478.5  on 2997  degrees of freedom
AIC: 9123.2

> drop1(regr1, test="Chisq")
Single term deletions

Model:
N ~ f1
                   Df   Deviance        AIC        LRT       Pr(>Chi)
       <none>           5'478.5     9'123.2
           f1       2   5'486.7     9'127.5      8.233        0.01630    *
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> exp(coef(regr1))
  (Intercept)      f1B        f1C
          0.9     1.15       1.05
```

Solution 4

(a) An example of an appropriate GLM would be a Poisson GLM with a log link given by

$$N_i \sim \mathrm{Poi}(\lambda(\boldsymbol{x}_i)),$$

where $\lambda : \mathcal{X} \to \mathbb{R}$ is given by $\boldsymbol{x} \mapsto \exp(\boldsymbol{\beta}'\boldsymbol{x})$. In order for the model to be applicable, we need $N_1, \dots, N_n$ to be independent.

(b) (i) There are 3000 observations. This can be read off from the number of degrees of freedom of the null model in the first output.

(ii) There is a single variable. This can be read off from the second output since there is only a single one-variable model (the other one is the null model). The fact that this variable is categorical can be seen from the first output — there are three coefficients, each corresponding to one category/label.

(iii) There is only a single variable. The last column of the table in the first output shows the p-value from a z-test testing equality of the given coefficient to 0. While we see that the coefficient for `f1B` is statistically significantly different from 0 on the 5% test level, this is not the case for the coefficients corresponding to `f1A` (intercept) and `f1C`. This means that we cannot reject that the effect on N differs among `f1A` and `f1C`. The p-value from the second output also says that we cannot reject the null model in favor of the single-variable model `regr2`.

(c) The third output can be used to carry out predictions using the single-variable model stored in `regr2`. Since our model gives us that

$$\widehat{N} = \widehat{\mathbb{E}[N]} = \exp\left(\hat{\beta}_0 + \hat{\beta}_1 1_{\{\mathtt{f1} \in \mathtt{f1B}\}} + \hat{\beta}_2 1_{\{\mathtt{f1} \in \mathtt{f1C}\}}\right),$$

we can predict N by its expectation depending on the class as

$$\hat{N} = \begin{cases} \exp(\hat{\beta}_0) = 0.9 & \text{if } \mathtt{f1} \in \mathtt{f1A}, \\ \exp(\hat{\beta}_0 + \hat{\beta}_1) = \exp \hat{\beta}_0 \exp \hat{\beta}_1 = 1.035 & \text{if } \mathtt{f1} \in \mathtt{f1B}, \\ \exp(\hat{\beta}_0 + \hat{\beta}_2) = \exp \hat{\beta}_0 \exp \hat{\beta}_2 = 0.945 & \text{if } \mathtt{f1} \in \mathtt{f1C}. \end{cases}$$

(d) (i) In order to improve the GLM, we could look for other features that have a justifiable relationship with the claim frequency variable N. In particular, the dispersion estimate of 2.10 might indicate that we are missing an important feature in the model. One could also opt for other features whose relationship with N cannot be easily justified, but it could then be argued that the interpretability of the resulting model would be worse. In a similar fashion, one could try to include transformations of the newly introduced non-categorical features or even try generalizations of the GLM such as GAMs, but these would all be arguably much less interpretable than the original `regr2` model. Alternatively, we could also try to use a different link function or to compare with a negative binomial or a Poisson GLM.

We do not know whether the categorical variable `f1` is nominal (such as a variable encoding three different colors) or is obtained by partitioning the range of a continuous variable similarly to the way we have done it in Question 1 (c). In case of the latter, one could also try to partition the range of the original continuous variable in a different way.

(ii) Which model performs better depends on what our goal is. If we only care about prediction, the best way to compare the models is to compare their out-of-sample loss under an appropriate loss function, such as the MSE. As long as we also care about interpretability of the selected model, we might want to accept a model which is not the best in terms of out-of-sample loss but the compensates it by being more parsimonious.

(e) (i) The only difference between the two models is that the family used in the second model is Poisson instead of quasi Poisson. The quasi Poisson family allows for the dispersion parameter being different from 1, which effectively breaks the equality of mean and variance associated with the Poisson distribution.

(ii) What changes in comparison to (b) is that the intercept now becomes statistically significantly different from 0 on the 5% test level. Additionally, based on the second output, the null model is now rejected in favor of the single-variable model `regr1` on the 5% test level.

(iii) Which model fits better to the data better depends on how the fit is measured. Out-of-sample statistics are not available, but if we predict using expectation, both models would perform the same with the same set of features since the expectations in the two models are the same. The difference only matters when it comes to assessing uncertainty.

As far as the Poisson GLM model goes, we can decide between the single-variable model and the null model based on the AIC, in which case we would opt for the null model since its AIC is lower.

In the quasi Poisson model, the AIC is not available, but since the null model cannot be rejected on the a priori chosen 5% test level, we might also opt for the null model.

Choosing between the null models with Poisson and quasi Poisson family is more complicated, since more detailed output is only available for the single-variable models. In the case of the single-variable models, one could argue that since we have 3000 observations the dispersion parameter is greater than 2, its difference from 1 would likely be statistically significant, which would suggest that the equality of the mean and the variance in the Poisson GLM does not hold, and we would thus opt for the quasi Poisson single-variable model. Alternatively, one could argue that the large dispersion is only a syndrome of omitted features and that one would therefore opt for the Poisson GLM and look for missing features.