Prof. Dr. E. Cramer, Prof. Dr. M. Kateri
T. v. Bentum, M.Sc., L. Kaufmann, M.Sc.

RWTH Aachen, SS 2021

publication: June 11th — solutions: June 18th

Applied Data Analysis

Exercise Sheet 4

Exercise 14

Let a normal linear model $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ be given, with $\boldsymbol{X} \in \mathbb{R}^{n\times d}$ constant and full rank, $\boldsymbol{\beta} \in \mathbb{R}^d$ and $\boldsymbol{\varepsilon} \sim N_n(0, \sigma^2 I_n)$.

(a) Explain, how to test the following hypotheses

(i) $H_0\colon \beta_1 = \beta_2 \qquad \longleftrightarrow \qquad H_1\colon \beta_1 \neq \beta_2,$

(ii) $H_0\colon \beta_1 = \cdots = \beta_d \qquad \longleftrightarrow \qquad H_1\colon \beta_j \neq \beta_k,$ for at least one pair $j \neq k$,

by using general linear hypothesis testing and nested model comparison.

(b) Construct level $(1-\alpha)$-confidence regions, $\alpha \in (0,1)$, for $\beta_1 - \beta_2$.

Hint: Confidence regions can be constructed by inverting a hypothesis test in the sense of constructing a non-rejecting region.

(c) If H_0 is rejected in test problem (a)(ii), one might ask, which pairs (j,k) have significantly different effects. To analyze these, the multiple comparison problem

$$H_0^{(j,k)}\colon \beta_j = \beta_k \qquad \longleftrightarrow \qquad H_1^{(j,k)}\colon \beta_j \neq \beta_k,$$

has to be decided for all $m := \binom{d}{2}$ pairs $j \neq k \in \{1,\dots,d\}$.

Show that if each single hypothesis $H_0^{(j,k)}$ is tested on a level $\frac{\alpha}{m}$, $\alpha \in (0,1)$, the multiple comparison tests together hold a family-wise error rate of α (Bonferroni correction). That means, if $I_0(\beta)$ denotes the subset of all hypotheses, which are true for $\boldsymbol{\beta} \in \mathbb{R}^d$ and $\varphi_i = 1$ denotes that test i rejects the ith hypothesis, then it holds for all $\boldsymbol{\beta} \in \mathbb{R}^d$ that

$$\mathsf{P}_{\boldsymbol{\beta}}\Bigg(\bigcup_{i \in I_0(\boldsymbol{\beta})} \{\varphi_i = 1\} \Bigg) \leq \alpha.$$

Exercise 15

(a) Let an iid sequence $\{\varepsilon_i\}_{i\in\mathbb{N}}$ with $\varepsilon_1 \sim \mathcal{N}(0,\sigma^2)$ and another sequence $\{\boldsymbol{x}_i\}_{i\in\mathbb{N}} \subset \mathbb{R}^{1\times d}$ be given. Assume that the linear model $Y_i = \boldsymbol{x}_i\boldsymbol{\beta} + \varepsilon_i$ holds, with $\boldsymbol{\beta} \in \mathbb{R}^d$. Assume that Y_i can only be observed, if $Y_i < L$ for a known constant $L > 0$. This means, after observing n data points, it is unknown how many data points are not observed. Construct a likelihood function for n observed realizations of Y_i, which allows for a consistent estimation of $\boldsymbol{\beta}$ and σ^2.

(b) Let a normal linear model be given like in (a), where the random variables

$$Y_i^* = \begin{cases} Y_i, & Y_i < L, \\ L, & Y_i \geq L, \end{cases}$$

$i \in \mathbb{N}$, will be observed. Construct a likelihood function which allows for an estimation of $\boldsymbol{\beta}$ and σ^2. Here, we observe all data points, but for values larger than the threshold L, we only observe L.

(c) Let iid sequences $\{\boldsymbol{x}_i\}_{i\in\mathbb{N}}$ and $\{\varepsilon_i\}_{i\in\mathbb{N}}$ be given, where $\boldsymbol{x}_1$ is a $(1\times d)$-random vector with $\mathrm{E}(\boldsymbol{x}_1'\boldsymbol{x}_1) = Q \in \mathbb{R}^{d\times d}$ regular and symmetric, $\mathrm{E}(\boldsymbol{x}_1'\varepsilon_1) = \mathbf{0}$ and $\mathrm{cov}(\boldsymbol{x}_1'\varepsilon_1) = V \in \mathbb{R}^{d\times d}$ positive definite. Further, let $\boldsymbol{\beta}$ be a d-dimensional model parameter vector. For each $n \in \mathbb{N}$, let a linear model

$$\begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} \boldsymbol{x}_1 \\ \vdots \\ \boldsymbol{x}_n \end{pmatrix} \boldsymbol{\beta} + \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

be given. Assume that there is an $n' \in \mathbb{N}$ such that $\sum_{i=1}^n \boldsymbol{x}_i'\boldsymbol{x}_i$ is almost surely regular for all $n > n'$. Show that the least squares estimator for $\boldsymbol{\beta}$ is consistent and asymptotically normally distributed with covariance matrix $Q^{-1}VQ^{-1}$, i.e., $\hat{\boldsymbol{\beta}}_n \xrightarrow{\mathsf{P}} \boldsymbol{\beta}$ and

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \xrightarrow{D} Z \sim \mathcal{N}_d(0, Q^{-1}VQ^{-1}).$$

Exercise 16

(a) Let Y be a random variable with density f. Confirm that the

(i) Gaussian (or normal) distribution, that is $Y \sim \mathcal{N}(\mu,\sigma^2)$ with

$$f(y;\, \mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(y-\mu)^2}{2\sigma^2}\right\}, \quad y \in \mathbb{R}, \mu \in \mathbb{R}, \sigma^2 > 0,$$

(ii) Poisson distribution, that is $Y \sim \mathcal{P}(\mu)$ with

$$f(y;\, \mu) = \frac{\mu^y}{y!} e^{-\mu}\, \mathbb{1}_{\mathbb{N}_0}(y), \quad \mu > 0,$$

is a member of the (univariate) Exponential Dispersion Family (EDF) by finding a suitable representation of the density in the form given in Definition II.2.3. In particular, confirm the representations for the natural parameter θ and the functions $b(\theta), a(\phi), c(y,\phi)$ given in Example II.2.5.

(b) Using the results of (a), confirm the expressions for the mean and variance given in Example II.2.5 for

   (i) the Gaussian distribution.

   (ii) the Poisson distribution.

Exercise 17

Using a gamma GLM with a log-link function gives similar results as applying a normal linear model to $\log(Y_i)$, where Y_i, $i = 1, \ldots, n$, denote random variables modeling the response data.

(a) Show that the gamma distribution is a member of the exponential dispersion family of distributions and determine the canonical link function.

(b) Describe a practical situation (in the sense of properties of a data set), where modeling data with a gamma distribution could be appropriate.

(c) Use a Taylor approximation to show that, when Y_i has standard deviation σ_i proportional to $\mathrm{E}(Y_i) = \mu_i$, $\log(Y_i)$ has approximately constant variance for small σ_i.

(d) The gamma GLM with log-link refers to $\log(\mathrm{E}(Y_i))$, whereas the ordinary linear model for the transformed response refers to $\mathrm{E}(\log(Y_i))$. Show that if $\log(Y_i) \sim \mathcal{N}(\mu_i, \sigma^2)$, then $\log(\mathrm{E}(Y_i)) = \mathrm{E}(\log(Y_i)) + \frac{\sigma^2}{2}$.

Exercise 18

(a) Let a GLM with $g(\mu_i) = \beta_0 + \beta_1 x_i$, $i = 1, \ldots, n$, be given, where $x_i \in \{0, 1\}$ and g is an appropriate link function, such that the GLM is well defined. Show that the fitted means equal the sample mean for the two groups with $x_i = 0$, $x_i = 1$ respectively.

(b) Let the family of distributions with density

$$p_\vartheta(x) = \begin{cases} \exp(\vartheta - x), & x \geq \vartheta, \\ 0, & x < \vartheta, \end{cases}$$

be given. Why is this family not in the exponential dispersion family?

(c) Let a GLM with design matrix $\boldsymbol{X}$ and canonical link function be given. Show that the residual vector $\boldsymbol{e} = \hat{\boldsymbol{\varepsilon}} := \boldsymbol{Y} - \hat{\boldsymbol{\mu}}$ is an element of the orthogonal complement of the column space of $\boldsymbol{X}$. Why does this not hold in general for non-canonical link functions?