Prof. Dr. E. Cramer, Prof. Dr. M. Kateri
T. v. Bentum, M.Sc., L. Kornely, M.Sc.
RWTH Aachen, SS 2021
publication: June 11th solutions available online: June 18th

Applied Data Analysis

Exercise Sheet 4 - Solutions

Exercise 14

Consider Theorems I.4.32, I.4.39 and I.4.40, I.5.21.

(a) **General linear hypothesis testing**:

(i) Choose $\boldsymbol{c} = (1, -1, 0, \ldots, 0)$ and test $\boldsymbol{c}\boldsymbol{\beta} = 0$. Since $\beta_1 = \beta_2 \Leftrightarrow \beta_1 - \beta_2 = 0$ this is clearly an equivalent formulation of H_0.

(ii) Choose

$$K = \begin{pmatrix} 1 & -1 & 0 & & \cdots & 0 \\ 0 & 1 & -1 & 0 & \cdots & 0 \\ \vdots & & \ddots & & & \vdots \\ 0 & & \cdots & 0 & 1 & -1 \end{pmatrix}$$

to test $K\boldsymbol{\beta} = \boldsymbol{0}$, since $\beta_1 = \ldots = \beta_d$ is equivalent to $\{\beta_1 = \beta_2, \beta_2 = \beta_3, \ldots, \beta_{d-1} = \beta_d\}$.

Note that the given matrix is not unique. Another possible matrix to test the same null hypothesis is e.g. given by

$$K_* = \begin{pmatrix} 1 & -1 & 0 & & \cdots & 0 \\ 1 & 0 & -1 & 0 & \cdots & 0 \\ \vdots & & \ddots & & & \vdots \\ 1 & & \cdots & 0 & 0 & -1 \end{pmatrix}$$

by choosing the formulation $\{\beta_1 = \beta_2, \ldots, \beta_1 = \beta_d\}$.

Remark: In an ANOVA, the global null hypothesis is usually stated as *no factor has any effect*, which can be formulated as (ii). Notice that for reduction of complexity in the formulation, here we would have used a model without intercept. The analogous hypothesis for models with intercept (respectively, the effect of a factor compared to the overall mean) can be stated completely analogous. Only the intercept term should not be tested. After rejecting this global hypothesis, one conducts an analysis using so called "post-hoc" tests, to determine which effect caused the rejection of the hypothesis (see (c)).

Nested model testing:

(i) We add the first and the second column of the design matrix $\boldsymbol{X}$ and remove one parameter from the parameter vector $\boldsymbol{\beta}$ to form model $\mathcal{M}_1$. The more complex model $\mathcal{M}_2$ is the model with the unchanged design matrix $\boldsymbol{X}$. Then we can test $\mathcal{M}_1$ against

$\mathcal{M}_2$.

$\mathcal{M}_2$ is given by a linear model

$$Y_i = \sum_{j=1}^{d} \beta_j x_{ij}, \quad i = 1, \ldots, n,$$

and the nested model $\mathcal{M}_1$ can be formulated as

$$Y_i = \beta_1(x_{i1} + x_{i2}) + \sum_{j=3}^{d} \beta_j x_{ij}, \quad i = 1, \ldots, n.$$

(ii) We add all columns and proceed like in (i).

(b) The hypothesis $K\boldsymbol{\beta} = \boldsymbol{\delta}$ for some $\boldsymbol{\delta} \in \mathbb{R}^d$ would not be rejected using an F-test (cf. lecture), if for the corresponding value of the statistic F it holds that $F < F_{q,n-d}(\alpha)$. For the confidence region, we trust all values $K\boldsymbol{\beta} = \boldsymbol{\delta}$ on a confidence level $1-\alpha$, for which $F < F_{q,n-d}(\alpha)$ and therefore the (unfortunately, in general, only abstractly given) $1-\alpha$ confidence region $\mathcal{K}$ for $K\boldsymbol{\delta}$ is given by

$$\mathcal{K} = \Big\{ \boldsymbol{\delta} \in \mathbb{R}^q \Big| (K\hat{\boldsymbol{\beta}} - \boldsymbol{\delta})' K(\boldsymbol{X}'\boldsymbol{X})^{-1}K'(K\hat{\boldsymbol{\beta}} - \boldsymbol{\delta}) \le q \cdot SSE^2 F_{q,n-d}(\alpha) \Big\}.$$

It is possible to construct an explicitly given confidence interval for $\beta_1 - \beta_2$. We know that (see Theorems I.4.32 and I.5.21)

$$\begin{aligned} \hat{\boldsymbol{\beta}} &\sim \mathcal{N}_d(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}), \\ K\hat{\boldsymbol{\beta}} &\sim \mathcal{N}(K\boldsymbol{\beta}, \sigma^2 K(\boldsymbol{X}'\boldsymbol{X})^{-1}K'), \\ \frac{SSE}{\sigma^2} &\sim \chi^2(n-d) \end{aligned}$$

Here, we have $K = (1, -1, 0, \ldots, 0)$ and therefore we receive

$$\frac{(\hat{\beta}_1 - \hat{\beta}_2) - (\beta_1 - \beta_2)}{\sqrt{SSE/(n-d)}\sqrt{v}} \sim t_{n-d},$$

where $v = (\boldsymbol{X}'\boldsymbol{X})^{-1}_{11} + (\boldsymbol{X}'\boldsymbol{X})^{-1}_{22} - 2(\boldsymbol{X}'\boldsymbol{X})^{-1}_{12}$ and $(\boldsymbol{X}'\boldsymbol{X})^{-1}_{ij}$ denotes the element ij of the matrix $(\boldsymbol{X}'\boldsymbol{X})^{-1}$. Hence, we get the $1-\alpha$ confidence interval of $\beta_1 - \beta_2$ as

$$\mathcal{K} = \big[\hat{\beta}_1 - \hat{\beta}_2 - t_{1-\frac{\alpha}{2},n-d}\sqrt{SSE/(n-d)}\sqrt{v},\ \hat{\beta}_1 - \hat{\beta}_2 + t_{1-\frac{\alpha}{2},n-d}\sqrt{SSE/(n-d)}\sqrt{v}\big].$$

Remark: This corresponds to hypothesis (a)(i) and we reject it, if $0 \notin \mathcal{K}$.

(c) Since φ_i is a random variable for a random vector $\boldsymbol{\varepsilon}$, $B_i = \{\varphi_i = 1\}$ defines an event, $i = 1, \ldots, m$. Therefore, we get from the sub-additivity of measures and the Bonferroni correction for each $\boldsymbol{\beta}$

$$\mathsf{P}_{\boldsymbol{\beta}}\Big(\bigcup_{i \in I_0(\beta)} \{\varphi_i = 1\} \Big) \le \sum_{i \in I_0(\beta)} \mathsf{P}_{\boldsymbol{\beta}}(\{\varphi_i = 1\}) = \frac{m_0 \alpha}{m} \le \alpha,$$

where $m_0 := |I_0(\boldsymbol{\beta})|$ and it always holds that $m_0 \leq m$.

Remember: We get the sub-additivity by defining the events

$$E_1 := B_1, \quad E_i := B_i \cap \left(\bigcap_{j=1}^{i-1} E_j^{\mathsf{c}} \right), \quad i = 2, \ldots, m.$$

Then we have $\bigcup_{i=1}^m E_i = \bigcup_{i=1}^m B_i$, $E_i \subset B_i$, $i = 1, \ldots, m$, and $E_1, \ldots, E_m$ are disjoint.

Exercise 15

(a) The response variable has a truncated distribution, in the sense that the distribution of a response variable is given under the condition that the value is smaller than L. Therefore, the cdf $F(\cdot)$ of the truncated distribution of Y_1 is given by the conditional probability

$$F(t) = \mathsf{P}(Y_1 \leq t \mid Y_1 < L) = \frac{\mathsf{P}(Y_1 \leq \min(t, L))}{\mathsf{P}(Y_1 \leq L)} = \frac{\Phi\left(\frac{\min(t,L) - \boldsymbol{x}_1\boldsymbol{\beta}}{\sigma}\right)}{\Phi\left(\frac{L - \boldsymbol{x}_1\boldsymbol{\beta}}{\sigma}\right)},$$

for $t \in \mathbb{R}$. Taking the derivative with respect to t to derive the density, the likelihood for $\boldsymbol{\beta}$ and σ^2 is given by

$$\mathcal{L}(\boldsymbol{\beta}, \sigma^2; \boldsymbol{y}) = \prod_{i=1}^n \frac{\phi\left(\frac{y_i - \boldsymbol{x}_1\boldsymbol{\beta}}{\sigma}\right)}{\sigma\Phi\left(\frac{L - \boldsymbol{x}_i\boldsymbol{\beta}}{\sigma}\right)}.$$

(b) The random variable Y_i^* has positive mass on the value L and is continuous for lower values. Therefore, its distribution is a mixture of a continuous and a discrete distribution. We have

$$\begin{aligned}
\mathsf{P}(Y_i^* = L) &= \mathsf{P}(Y_i \geq L) = \mathsf{P}(Y_i > L) = 1 - \mathsf{P}(Y_i \leq L) = 1 - \mathsf{P}(\varepsilon_i \leq L - \boldsymbol{x}_i\boldsymbol{\beta}) \\
&= 1 - \Phi\left(\frac{L - \boldsymbol{x}_i\boldsymbol{\beta}}{\sigma}\right),
\end{aligned}$$

for $i = 1, \ldots, n$. Further, for values smaller than L, Y^* has a normal density. Denoting the indicator function by $I(\cdot)$, the likelihood is given by

$$\mathcal{L}(\boldsymbol{\beta}, \sigma^2; \boldsymbol{y}^*) = \prod_{i=1}^n \left(\frac{1}{\sigma}\phi\left(\frac{y_i^* - \boldsymbol{x}_i\boldsymbol{\beta}}{\sigma}\right)\right)^{I(y_i^* < L)} \cdot \left(1 - \Phi\left(\frac{L - \boldsymbol{x}_i\boldsymbol{\beta}}{\sigma}\right)\right)^{I(y_i^* = L)}.$$

(c) **Consistency**:
In the notation of the Exercise, the least squares estimator $\hat{\boldsymbol{\beta}}_n$ of $\boldsymbol{\beta}$ is for sufficiently large n almost surely given by

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_n &= \left(\sum_{i=1}^n \boldsymbol{x}_i'\boldsymbol{x}_i \right)^{-1} \sum_{i=1}^n \boldsymbol{x}_i'\boldsymbol{y}_i = \left(\sum_{i=1}^n \boldsymbol{x}_i'\boldsymbol{x}_i \right)^{-1} \sum_{i=1}^n \boldsymbol{x}_i'(\boldsymbol{x}_i\boldsymbol{\beta} + \varepsilon_i) \\
&= \left(\sum_{i=1}^n \boldsymbol{x}_i'\boldsymbol{x}_i \right)^{-1} \left(\sum_{i=1}^n \boldsymbol{x}_i'\boldsymbol{x}_i \right)\boldsymbol{\beta} + \left(\sum_{i=1}^n \boldsymbol{x}_i'\boldsymbol{x}_i \right)^{-1} \sum_{i=1}^n \boldsymbol{x}_i'\varepsilon_i \\
&= \boldsymbol{\beta} + \left(\frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i'\boldsymbol{x}_i \right)^{-1} \left(\frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i'\varepsilon_i \right).
\end{aligned}$$

Since $\{\boldsymbol{x}_i\}_{i\in\mathbb{N}}$ is an iid sequence, $\{\boldsymbol{x}_i'\boldsymbol{x}_i\}_{i\in\mathbb{N}}$ is iid as well and it has finite first moments according to the prerequisites. It follows by the weak law of large numbers that $\frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i'\boldsymbol{x}_i \xrightarrow{\mathsf{P}} \boldsymbol{Q}$. Analogous, we have $\frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i'\varepsilon_i \xrightarrow{\mathsf{P}} \mathrm{E}(\boldsymbol{x}_1'\varepsilon_1) = 0$ and so we get from the continuous mapping theorem $\hat{\boldsymbol{\beta}}_n \xrightarrow{\mathsf{P}} \boldsymbol{\beta}$ and hence the consistency.

Asymptotic normality:
Since $\mathrm{E}(\boldsymbol{x}_1'\varepsilon_1) = 0$, we have

$$\begin{aligned}\sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) &= \sqrt{n}\Bigg(\boldsymbol{\beta} + \Big(\frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i'\boldsymbol{x}_i\Big)^{-1}\frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i'\varepsilon_i - \boldsymbol{\beta}\Bigg)\\ &= \Big(\frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i'\boldsymbol{x}_i\Big)^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^n \big(\boldsymbol{x}_i'\varepsilon_i - \mathrm{E}(\boldsymbol{x}_1'\varepsilon_1)\big).\end{aligned}$$

Since $\{\boldsymbol{x}_i'\varepsilon_i\}_{i\in\mathbb{N}}$ is an iid sequence with $\mathrm{cov}(\boldsymbol{x}_1'\varepsilon_1) = \boldsymbol{V} \in \mathbb{R}^{d\times d}$ positive definite and especially finite, we can apply the multivariate central limit theorem to get

$$\frac{1}{\sqrt{n}}\sum_{i=1}^n \big(\boldsymbol{x}_i'\varepsilon_i - \mathrm{E}(\boldsymbol{x}_1'\varepsilon_1)\big) \xrightarrow{\mathcal{D}} \boldsymbol{Z}' \sim \mathcal{N}_d(0, \boldsymbol{V}).$$

With $(\frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i'\boldsymbol{x}_i)^{-1} \xrightarrow{\mathsf{P}} \boldsymbol{Q}^{-1}$ (see last part) and Slutsky's theorem follows the result.

Exercise 16

(a) (i) We can write

$$\begin{aligned}\ln(f(y;\mu,\sigma^2)) &= -\frac{1}{2}\ln(2\pi\sigma^2) - \frac{(y-\mu)^2}{2\sigma^2}\\ &= \frac{y\mu - \frac{\mu^2}{2}}{\sigma^2} - \frac{1}{2}\ln(2\pi\sigma^2) - \frac{y^2}{2\sigma^2}, \quad y \in \mathbb{R},\end{aligned}$$

confirming the given identities in Example II.2.5.

(ii) For $y \in \mathbb{N}_0$ we can write

$$\begin{aligned}\ln(f(y;\mu)) &= y\ln(\mu) - \ln(y!) - \mu\\ &= y\theta - e^\theta - \ln(y!)\end{aligned}$$

for $\theta = \ln(\mu)$, confirming the given identities in Example II.2.5.

(b) (i) By (a) we have

$$b(\theta) = \frac{\theta^2}{2}$$

and therefore by Proposition II.2.4

$$\mu = \mathrm{E}(Y) = b'(\theta) = \theta.$$

Furthermore, by (a) we have

$$a(\phi) = \sigma^2$$

and therefore by Proposition II.2.4

$$\mathrm{Var}(Y) = b''(\theta)a(\phi) = 1\cdot\sigma^2 = \sigma^2.$$

(ii) By (a) we have

$$b(\theta) = \frac{\theta^2}{2}$$

and therefore by Proposition II.2.4

$$\mu = \mathrm{E}(Y) = b'(\theta) = e^{\theta}.$$

Furthermore, by (a) we have

$$a(\phi) = 1$$

and therefore by Proposition II.2.4

$$\mathrm{Var}(Y) = b''(\theta)a(\phi) = e^{\theta} \cdot 1 = e^{\theta}.$$

Exercise 17

(a) We use the gamma distribution in the parameterization

$$f(y; k, \mu) = \begin{cases} \left(\frac{k}{\mu}\right)^k \frac{y^{k-1}}{\Gamma(k)} \exp\left(\frac{-ky}{\mu}\right), & y \geq 0, \\ 0, & y < 0. \end{cases}$$

Then we have for $y \geq 0$

$$\begin{aligned} f(y; k, \mu) &= \exp\left(-\frac{k}{\mu} y + k \log\left(\frac{k}{\mu}\right) + (k-1)\log(y) - \log\big(\Gamma(k)\big) \right) \\ &= \exp\left(\frac{-\frac{1}{\mu} y - \big(-\log\left(\frac{1}{\mu}\right)\big)}{\frac{1}{k}} + k\log(k) + (k-1)\log(y) - \log\big(\Gamma(k)\big) \right) \\ &= \exp\left(\frac{\theta y - b(\theta)}{\frac{\psi}{\omega}} + c(y, \psi, \omega) \right), \end{aligned}$$

where the natural parameter is $\theta = -\frac{1}{\mu}$ and therefore the natural link is $g(\mu) = -\mu^{-1}$, the dispersion parameter is $\psi = \frac{1}{k}$, and for the other parameters it holds that

$$b(\theta) = -\log(-\theta),\ c(y, \psi, \omega) = \frac{-\log(\psi)}{\psi} + \left(\frac{1}{\psi} - 1\right)\log(y) - \log\left(\Gamma\left(\frac{1}{\psi}\right)\right), \text{ and } \omega = 1.$$

Remark: If the shape k of the gamma distribution is also treated as a parameter, it is a member of the two-parametric exponential family, whose probability density or mass functions are given by (cf. slide 86, Def. 3.4)

$$f(y; \theta) = h(y) \exp\big(\theta' T(y) - B(\theta)\big)$$

as we can see by setting

$$\begin{aligned} f(y; k, \mu) &= \left(\frac{k}{\mu}\right)^k \frac{y^{k-1}}{\Gamma(k)} \exp\left(\frac{-ky}{\mu}\right) \mathbb{1}_{y \geq 0} \\ &= \exp\left(k \log\left(\frac{k}{\mu}\right) + (k-1)\log(y) - \log\big(\Gamma(k)\big) - \frac{k}{\mu} y \right) \mathbb{1}_{y \geq 0} \\ &= \exp\left[\underbrace{\begin{pmatrix} k-1 \\ -\frac{k}{\mu} \end{pmatrix}'}_{\theta} \underbrace{\begin{pmatrix} \log(y) \\ y \end{pmatrix}}_{T(y)} - \underbrace{\left(\log\big(\Gamma(k)\big) - k\log\left(\frac{k}{\mu}\right) \right)}_{B(\theta)} \right] \underbrace{\mathbb{1}_{y \geq 0}}_{h(y)}. \end{aligned}$$

(b) The gamma distribution is skewed to the right with decreasing skewness, if k increases. A usual approach to target skewness is applying a log-transformation and a normal linear model. A gamma GLM could be an alternative. For the gamma distribution, it holds that $\mathrm{Var}(Y) = \frac{\mathrm{E}(Y)^2}{k}$, so the property given in (c) is interesting. If we observe positive numbers with standard deviation proportional to the mean (for example this appears often for time measures), modeling with a gamma GLM could be a proper approach.

(c) With $X_i = \log(Y_i)$ we get from the Taylor expansion

$$X_i = \log\big(\mathrm{E}(Y_i)\big) + \frac{1}{\mathrm{E}(Y_i)}\big(Y_i - \mathrm{E}(Y_i)\big) + \mathcal{O}\big(|Y_i - \mathrm{E}(Y_i)|^2\big)$$

and therefore for Y_i close to $\mathrm{E}(Y_i)$

$$\log(Y_i) \approx \log\big(\mathrm{E}(Y_i)\big) + \frac{Y_i}{\mathrm{E}(Y_i)} - 1.$$

Using this to approximate the expected value and the variance, we get

$$\begin{aligned}
\mathrm{E}\big(\log(Y_i)\big) &\approx \mathrm{E}\Big(\log\big(\mathrm{E}(Y_i)\big) + \frac{Y_i}{\mathrm{E}(Y_i)} - 1\Big) = \log\big(\mathrm{E}(Y_i)\big),\\
\mathrm{Var}\big(\log(Y_i)\big) &\approx \mathrm{Var}\Big(\log\big(\mathrm{E}(Y_i)\big) + \frac{Y_i}{\mathrm{E}(Y_i)} - 1\Big) = \frac{\mathrm{Var}(Y_i)}{\mathrm{E}(Y_i)^2}.
\end{aligned}$$

Since the standard deviation $\sigma = \sqrt{\mathrm{Var}(Y_i)}$ is proportional to the expected value $\mu = \mathrm{E}(Y_i)$, a real number k exists such that $\sigma = k\mu$ and we receive $\mathrm{Var}(Y_i) = \sigma^2 = k^2\mu^2$. If especially σ is small, we expect $|Y_i - \mathrm{E}(Y_i)|^2$ to be small as well, and we get from the Taylor approximation

$$\mathrm{Var}(\log(Y_i)) \approx \frac{k^2\mu^2}{\mu^2} = k^2,$$

i.e., the variance is approximately constant.

(d) Since $\log(Y_i) \sim \mathcal{N}(\mu_i, \sigma^2)$, we get with $Z \sim \mathcal{N}(0,1)$

$$\begin{aligned}
\mathrm{E}(Y_i) = \mathrm{E}\big(\exp(\sigma Z + \mu_i)\big) &= \int_{-\infty}^{\infty} \exp(\sigma z + \mu_i)\frac{1}{\sqrt{2\pi}}\exp\Big(-\frac{z^2}{2}\Big)\,\mathrm{d}z\\
&= \int_{-\infty}^{\infty} \exp(\mu_i)\frac{1}{\sqrt{2\pi}}\exp\Big(-\frac{z^2}{2} + \sigma z - \frac{\sigma^2}{2} + \frac{\sigma^2}{2}\Big)\,\mathrm{d}z\\
&= \exp\Big(\mu_i + \frac{\sigma^2}{2}\Big)\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\exp\Big(-\frac{1}{2}(z-\sigma)^2\Big)\,\mathrm{d}z = \exp\Big(\mathrm{E}\big(\log(Y_i)\big) + \frac{\sigma^2}{2}\Big).
\end{aligned}$$

Exercise 18

(a) Denote the random sample by $Y_1, \ldots, Y_n$. The explanatory variable defines two groups, which we split into index sets $I_j = \{i \in \mathbb{N} \mid 1 \le i \le n, x_i = j\}, j \in \{0,1\}$. Without loss of generality, we assume that $1 \in I_0$ and $n \in I_1$. Further, the GLM is written as $g(\mu_i) = \beta_0 x_{i0} + \beta_1 x_{i1} = \eta_i$, with $x_{i1} = x_i$ and for the intercept $x_{i0} = 1$, $i = 1, \ldots, n$.

Then it holds that $g(\mu_i) = \beta_0$ for $i \in I_0$ and $g(\mu_i) = \beta_0 + \beta_1$ for $i \in I_1$ such that

$$\begin{aligned} \frac{\partial g^{-1}(\eta_i)}{\partial \eta_i} &= \frac{\partial g^{-1}(\eta_1)}{\partial \eta_1} \qquad \text{for } i \in I_0 \\ \frac{\partial g^{-1}(\eta_i)}{\partial \eta_i} &= \frac{\partial g^{-1}(\eta_n)}{\partial \eta_n} \qquad \text{for } i \in I_1. \end{aligned}$$

The likelihood equations for the GLM become (cf. slide 90):

$$\sum_{i=1}^{n} \left(\frac{y_i - E(Y_i)}{\mathrm{Var}(Y_i)} \cdot \frac{\partial g^{-1}(\eta_i)}{\partial \eta_i} x_{ik} \right) = 0 \qquad \text{for } k \in \{0, 1\}.$$

The random sample $Y_1, \ldots, Y_n$ is assumed to follow the GLM. By construction of the GLM, it follows that $\mathrm{Var}(Y_i) = \mathrm{Var}(Y_\ell)$, if $x_{i1} = x_{\ell 1}$ (variance depends on the overall equal dispersion parameter, the weight (which is equal to one, since we take a look on each single observation in the group) and the natural parameter (which is equated by the GLM)).

Therefore we can set $E(Y_i) = \mu_j$, $\mathrm{Var}(Y_i) = \sigma_j^2$, $i \in I_j$, $j \in \{0, 1\}$, and conclude for $k = 1$

$$\begin{aligned} & \sum_{i=1}^{n} \left(\frac{y_i - E(Y_i)}{\mathrm{Var}(Y_i)} \cdot \frac{\partial g^{-1}(\eta_i)}{\partial \eta_i} x_{ik} \right) = \sum_{i \in I_1} \left(\frac{y_i - E(Y_i)}{\mathrm{Var}(Y_i)} \cdot \frac{\partial g^{-1}(\eta_i)}{\partial \eta_i} \right) = 0 \\ \iff \quad & \frac{1}{\sigma_1^2} \frac{\partial g^{-1}(\eta_n)}{\partial \eta_n} \sum_{i \in I_1} (y_i - \mu_1) = 0 \\ \iff \quad & \sum_{i \in I_1} (y_i - \mu_1) = 0 \\ \iff \quad & \mu_1 = \frac{1}{|I_1|} \sum_{i \in I_1} y_i =: \overline{\mu}_1, \end{aligned}$$

and for $k = 0$

$$\begin{aligned} & \sum_{i=1}^{n} \left(\frac{y_i - E(Y_i)}{\mathrm{Var}(Y_i)} \cdot \frac{\partial g^{-1}(\eta_i)}{\partial \eta_i} x_{ik} \right) \\ & \qquad = \underbrace{\sum_{i \in I_1} \left(\frac{y_i - E(Y_i)}{\mathrm{Var}(Y_i)} \cdot \frac{\partial g^{-1}(\eta_i)}{\partial \eta_i} \right)}_{=0} + \sum_{i \in I_0} \left(\frac{y_i - E(Y_i)}{\mathrm{Var}(Y_i)} \cdot \frac{\partial g^{-1}(\eta_i)}{\partial \eta_i} \right) = 0 \\ \iff \quad & \sum_{i \in I_0} (y_i - \mu_0) = 0 \\ \iff \quad & \mu_0 = \frac{1}{|I_0|} \sum_{i \in I_0} y_i =: \overline{\mu}_0. \end{aligned}$$

(b) The distributions of the exponential dispersion family are defined with respect to a single measure. Therefore, we have to represent the function

$$p_\vartheta(x) = \mathbb{1}_{x \geq \vartheta} \exp(\vartheta - x), \qquad x \in \mathbb{R},$$

in the form given in Definition 3.1 of the lecture (cf. slide 84). This is impossible, since $\exp(x) > 0$ for all $x \in \mathbb{R}$ and therefore there is no function $c(x, \vartheta)$ such that $\exp(c(x, \vartheta)) = \mathbb{1}_{x \geq \vartheta}$ for all $x \in \mathbb{R}$.

Remark: Moreover, since this argument depends only on the support of the distribution, no family of distributions, whose support is dependent on the parameters, is in the exponential family.

(c) In matrix notation, the likelihood equations for GLM with canonical link are given by

$$\boldsymbol{X}'(\boldsymbol{Y} - \boldsymbol{\mu}) = 0.$$

Denote the number of parameters by $q \in \mathbb{N}$. For $\boldsymbol{v} \in \text{Im}(\boldsymbol{X})$ a vector $\boldsymbol{\beta} \in \mathbb{R}^q$ exists, such that $\boldsymbol{v} = \boldsymbol{X}\boldsymbol{\beta}$. Then for the fitted values $\hat{\boldsymbol{\mu}}$ it follows from the likelihood equations $\boldsymbol{X}'\boldsymbol{e} = \boldsymbol{0}$ and especially

$$\boldsymbol{e}'\boldsymbol{v} = \boldsymbol{e}'\boldsymbol{X}\boldsymbol{\beta} = (\boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{e})' = (\boldsymbol{\beta}'\boldsymbol{0})' = 0.$$

Using a non-canonical link, the general likelihood equations (Remark 3.7, cf. slide 90)

$$\boldsymbol{X}\boldsymbol{D}\boldsymbol{V}^{-1}(\boldsymbol{Y} - \boldsymbol{\mu}) = \boldsymbol{0}$$

imply $\boldsymbol{X}'\boldsymbol{D}\boldsymbol{V}^{-1}\boldsymbol{e} = \boldsymbol{0}$. However, if $\boldsymbol{X}'\boldsymbol{D}\boldsymbol{V}^{-1}\boldsymbol{e} = \boldsymbol{0}$, then $\boldsymbol{X}'\boldsymbol{e} = \boldsymbol{0}$ does not necessarily hold.

We illustrate this with a short example:
We observed a binary sample $y = (0, 0, 1, 1, 0)'$ with the single explanatory variable $x = (3, 1, 2, 4, 2)'$, whose true relationship is $\pi_i = \frac{x}{8}$, $i = 1, \dots, 5$. If we consider the (correct) non-canonical link model $g(\pi_i) = \pi_i = \beta x_i$, $i = 1, \dots, 5$, we get the maximum likelihood estimate $\hat{\beta} \approx 0.18212$ and with $\alpha \in \mathbb{R}$ it follows that

$$(y - \hat{\pi})' x \alpha \approx -0.1920336 \cdot \alpha \neq 0$$

for $\alpha \neq 0$ (not even numerically) and therefore $(y - \hat{\pi}) \notin \text{Im}(\boldsymbol{X})^{\perp}$.

If we use the canonical link, we have $(y - \hat{\pi})'x \approx 0$ and therefore $(y - \hat{\pi})$ is in the orthogonal complement of the column space of $\boldsymbol{X}$.