Prof. Dr. E. Cramer, Prof. Dr. M. Kateri,

RWTH Aachen, SS 2021
23^{rd} August 2021

Applied Data Analysis

PT 1

Part 1: R Tasks

The solutions of the tasks have to be given in the required precisions. For example: if the output in R is given by 1.23456 and it should be given in a precision of 4 digits, this means that the solution is 1.2346. Thus, you have to **round the result with a precision of 4 digits**. If the output is given by 0.999 (or 0.901), the answer in a **precision of 2 digits** would be 1.00 (or 0.90) which is simplified in Dynexite to 1 (or 0.9). Note that numbers given in the wrong precision are evaluated as wrong!

Task 1

Clear your `R` workspace. Set the seed to (A) 2021, (B) 123, (C) 456, (D) 789. Please execute the function set.seed() with the requested seed every time you generate random numbers etc.
Set $n = 150$.

(a) Let X_1 be a uniformly distributed random variable on $[-40, 70]$ and let X_2 be an exponentially distributed random variable with mean equal to $\frac{1}{2}$. Sample n observations from X_1 and n observations from X_2. For $X = (X_1, X_2)$, consider the linear model $(Y|X = x) \sim \mathcal{N}(\mu(x), \sigma^2)$ with

$$\mathbb{E}(Y \mid X = x) = \mu(x) = 10 - 37 \cdot x_1 + 2 \cdot 10^{-8} \cdot x_1^3 - 10^{-7} \cdot x_1^4 + 3 \cdot 10^{-9} \cdot x_1^6 - 0.5 \cdot x_2$$

and $\sigma = 5$. Calculate values for the response variable Y based on the sampled x-data and the model above. What is the mean value for the realizations of the response? **(requested precision: 2 digits)**

(b) Fit the model $\mu^{(1)} = \beta_0 + \beta_1 x_1 + \beta_2 x_1^3 + \beta_3 x_1^4 + \beta_4 x_1^6 + \beta_5 x_2$. Calculate the p-value for testing whether there is evidence against the assumption that $\beta_1 = 0$. If β_1 is statistically significant at significance level $\alpha = 0.05$ then type in "1", else type in "0" (without quotation marks).

(c) Fit the model $\mu^{(2)} = \beta_0 + \beta_1 x_1$. Calculate the p-value for testing whether there is evidence against the assumption that the residuals of this model are normally distributed. Furthermore, if there is evidence against the normality assumption for the residuals at significance level $\alpha = 0.05$, then type in "1" and type in "0" else. (without quotation marks)

(d) Calculate the predicted residual sum of squares (PRESS) for the model fitted in (b) **(requested precision: 2 digits)** and for the model fitted in (c) **(requested precision: 2 digits)**. Which of these two models do you recommend based on the previous

analyses? Type in "1" for the model fitted in (b) and "2" for the model fitted in (c) (without quotation marks)

Solution

(a) set.seed((A) 2021, (B) 123, (C) 456, (D) 789)
n = 150
x = runif(n,-40,70) #x1
z = rexp(n,2) #x2
$mu = 10 - 37 * x + 2e^{-8} * x^3 - 1e^{-7} * x^4 + 3e^{-9} * x^6 - 0.5 * z$
y = mu + rnorm(n,sd=5)
mean(y)

(A) -554.3593 (Dynexite: -554.36)
(B) -528.8502 (Dynexite -528.85)
(C) -680.9756 (Dynexite: -680.98)
(D)-448.3529 (Dynexite: -448.35)

Alternative solution
set.seed((A) 2021, (B) 123, (C) 456, (D) 789)
n = 150
x = runif(n,-40,70) #x1
set.seed((A) 2021, (B) 123, (C) 456, (D) 789)
z = rexp(n,2) #x2
$mu = 10 - 37 * x + 2e^{-8} * x^3 - 1e^{-7} * x^4 + 3e^{-9} * x^6 - 0.5 * z$
set.seed((A) 2021, (B) 123, (C) 456, (D) 789)
y = mu + rnorm(n,sd=5)
mean(y)

(A) -554.8881 (Dynexite: -554.89)
(B) -528.7067 (Dynexite -528.71)
(C) -681.3351 (Dynexite: -681.34)
(D)-448.844 (Dynexite: -448.84)

(b) set.seed((A) 2021, (B) 123, (C) 456, (D) 789)
model.correct=lm($y \sim x + I(x^3) + I(x^4) + I(x^6) + z$)
summary(model.correct)
$\Rightarrow$ p-value for testing H_0: $\beta_1 = 0$ against H_1: $\beta_1 \neq 0$ is $< 2e^{-16}$ for (A), (B), (C) and (D) so less than 0.05 so reject H_0 so there is evidence against the assumption $\beta_1 = 0$ and the answer is **"1"**.

Alternative solution
The answer is **"1"** (same `R` code as above, but with the alternatively generated dataset)

(c) set.seed((A) 2021, (B) 123, (C) 456, (D) 789)
model.simple=lm($y \sim x$)
# shapiro test: H_0: normal distribution, H_1: no normal distribution
shapiro.test(model.simple$residuals) # p-value is 2.141e-07 for (A), 4.1 e-10 for (B), 1.346e-08 for (C) and 3.61e-15 for (D) so reject H_0 in all cases of different seeds so there is evidence against the assumption that the residuals are normally distributed

and the answer is **"1"**.

Alternative solution
The answer is **"1"** (same `R` code as above, but with the alternatively generated dataset)

(d) **(1) Initial solution with a typo in the code (`mu` should be `y`)**
set.seed((A) 2021, (B) 123, (C) 456, (D) 789)
sum((model.correct$fitted.values-mu)^2) **(A) 229.9803 (Dynexite: 229.98)**
(B) 53.96609 (Dynexite 53.97)
(C) 93.13733 (Dynexite 93.14)
(D) 118.8648 (Dynexite 118.86)

sum((model.simple$fitted.values-mu)^2) **(A) 460298.6 (Dynexite: 460298.6)**
(B) 426683.4 (Dynexite 426683.4)
(C) 511065 (Dynexite: 511065)
(D) 340150.8 (Dynexite: 340150.8)
$\Rightarrow$ we prefer the model of (b) so the answer is **"1"**.

Alternative solution of (1)
set.seed((A) 2021, (B) 123, (C) 456, (D) 789)
sum((model.correct$fitted.values-mu)^2) **(A) 226.5954 (Dynexite: 226.60)**
(B) 73.31219 (Dynexite 73.31)
(C) 116.4634 (Dynexite 116.46)
(D) 50.76128 (Dynexite 50.76)

sum((model.simple$fitted.values-mu)^2) **(A) 460554.8 (Dynexite: 460554.8)**
(B) 426378.8 (Dynexite 426378.8)
(C) 510568 (Dynexite: 510568)
(D) 339337.4 (Dynexite: 339337.4)
$\Rightarrow$ we prefer the model of (b) so the answer is **"1"**.

(2) Corrected solution
set.seed((A) 2021, (B) 123, (C) 456, (D) 789)
sum((model.correct$fitted.values-y)^2) **(A)3300.709 (Dynexite 3300.71)**
(B) 3690.321 (Dynexite 3690.32)
(C)3207.737(Dynexite 3207.74)
(D) 3742.568 (Dynexite 3742.57)

sum((model.simple$fitted.values-y)^2) **(A)473172 (Dynexite 473172)**
(B) 429122.1 (Dynexite 429122.1)
(C) 518879.2 (Dynexite 518879.2)
(D)340727.4 (Dynexite 340727.4)
$\Rightarrow$ we prefer the model of (b) so the answer is **"1"**.

Alternative solution of (2)
set.seed((A) 2021, (B) 123, (C) 456, (D) 789)

sum((model.correct$fitted.values-y)^2) **(A)4098.634 (Dynexite 4098.63)**
(B)3288.961(Dynexite 3288.96)
(C) 3535.901 (Dynexite 3535.9)
(D)3626.164(Dynexite 3626.16)

sum((model.simple$fitted.values-y)^2) **(A)471590.8 (Dynexite 471590.8)**
(B)433679.1 (Dynexite 433679.1)
(C) 511237.6 (Dynexite 511237.6)
(D) 344115 (Dynexite 344115)

$\Rightarrow$ we prefer the model of (b) so the answer is **"1"**.

Task 2

Clear your `R` workspace. Set the seed to (A) 2021, (B) 123, (C) 456, (D) 789. Please execute the function set.seed() with the requested seed every time you generate random numbers etc.
Set $n = 50$. Consider a GLM for independent Poisson responses Y_i, $i = 1, \ldots, n$, with canonical link and linear predictor

$$\eta_i = 3 - 2 \cdot x_{1,i} + 1.8 \cdot x_{2,i} + x_{1,i} \cdot x_{2,i} - x_{3,i} - x_{3,i} \cdot x_{2,i} \tag{1}$$

Sample n independent observations $x_{1,i}$, $i = 1, ..., n$, from a random variable $X_1 \sim \mathcal{N}(0, 1)$, n independent observations $x_{2,i}$, $i = 1, ..., n$, from a random variable $X_2 \sim \mathcal{N}(0, 1)$ and n independent observations $x_{3,i}$, $i = 1, ..., n$, from an exponentially distributed random variable X_3 with mean equal to $\frac{1}{2}$.

(a) What is the mean value of the linear predictor η for the sampled observations? **(requested precision: 2 digits)**

(b) Calculate the values y_i for the Poisson response variable Y_i of the GLM above. Fit a Poisson GLM that predicts the response variable Y of (b) using x_1, x_2, x_3 and all two-way interactions of x_i and x_j for $i, j \in \{1, 2, 3\}$, $i \neq j$. What is the deviance of the resulting model ? **(requested precision: 2 digits)**

(c) Based on the model fitted in (b), calculate the p-value for testing the statistical significance of the coefficient of the interaction term $x_1 \cdot x_3$ at level $\alpha = 0.05$. Furthermore, if the interaction term is statistically significant at significance level $\alpha = 0.05$, then type in "1" and type in "0" else (without quotation marks)

(d) Execute a stepwise selection algorithm based on BIC to select the best model nested in the model fitted in (b). What is the deviance of the resulting model? **(requested precision: 2 digits)**

Solution

(a) set.seed((A) 2021, (B) 123, (C) 456, (D) 789)
n=50
x1=rnorm(n,0,1)

x2=rnorm(n,0,1)
x3=rexp(n,2)
lin.pred=3-2 $\cdot x1 + 1.8 \cdot x2 + x1 \cdot x2 - 1 \cdot x3 - 1 \cdot x3 \cdot x2$ # true model
mean(lin.pred)

(A) 2.051457 (Dynexite: 2.05)
(B) 2.679366 (Dynexite 2.68)
(C) 2.136726 (Dynexite: 2.14)
(D) 2.804195 (Dynexite: 2.8)

Alternative solution
set.seed((A) 2021, (B) 123, (C) 456, (D) 789)
n=50
x1=rnorm(n,0,1)
set.seed((A) 2021, (B) 123, (C) 456, (D) 789)
x2=rnorm(n,0,1)
set.seed((A) 2021, (B) 123, (C) 456, (D) 789)
x3=rexp(n,2)
lin.pred=3-2 $\cdot x1 + 1.8 \cdot x2 + x1 \cdot x2 - 1 \cdot x3 - 1 \cdot x3 \cdot x2$ # true model
mean(lin.pred)

(A) 3.806267 (Dynexite: 3.81)
(B) 3.282305 (Dynexite 3.28)
(C) 3.646956 (Dynexite: 3.65)
(D) 3.489525 (Dynexite: 3.49)

(b) set.seed((A) 2021, (B) 123, (C) 456, (D) 789)
y1=rpois(n, exp(lin.pred)) # response variable
data1=data.frame(y=y1,Int=rep(1,n),x1=x1,x2=x2,x3=x3)
model.1=glm(y ~ x1+x2+x3+x1*x2+x1*x3+x2*x3,family=poisson(link='log'),data=data1)
model.1$deviance

(A) 40.55574 (Dynexite: 40.56)
(B) 36.99027 (Dynexite 36.99)
(C) 38.73619 (Dynexite: 38.74)
(D) 34.36591 (Dynexite: 34.37)

Alternative solution
set.seed((A) 2021, (B) 123, (C) 456, (D) 789)
y1=rpois(n, exp(lin.pred)) # response variable
data1=data.frame(y=y1,Int=rep(1,n),x1=x1,x2=x2,x3=x3)
model.1=glm(y ~ x1+x2+x3+x1*x2+x1*x3+x2*x3,family=poisson(link='log'),data=data1)
model.1$deviance

(A)61.40064(Dynexite: 61.4)
(B) 34.52523 (Dynexite 34.53)
(C)34.93963 (Dynexite: 34.94)
(D)48.86952(Dynexite: 48.87)

(c) summary(model.1) #p-value interaction x1 $\cdot$ x3 is (A) 0.258, (B) 0.636 (C) 0.62 and (D) 0.346 , so in all cases we do not reject the null hypothesis that x1*x3 is non influential and the answer is **"0"** (coefficient not significant)

Alternative solution
summary(model.1) #p-value interaction x1 $\cdot$ x3 is (A) <2e-16, (B)<2e-16 (C) <2e-16

and (D) <2e-16 , so in all cases wereject the null hypothesis that x1*x3 is non influential and the answer is **"1"**

(d) set.seed((A) 2021, (B) 123, (C) 456, (D) 789)
model.BIC =step(model.1, direction="backward",k=log(n))
model.BIC$deviance

(A) 41.8148 (Dynexite: 41.81)
(B) 37.21597 (Dynexite 37.22)
(C) 38.98014(Dynexite: 38.98)
(D) 35.25819 (Dynexite: 35.26)

Alternative solution
set.seed((A) 2021, (B) 123, (C) 456, (D) 789)
model.BIC =step(model.1, direction="backward",k=log(n))
model.BIC$deviance

(A)61.40064(Dynexite: 61.4)
(B)34.52523(Dynexite 34.53)
(C) 34.93963 (Dynexite: 34.94)
(D) 48.86952 (Dynexite: 48.87)

Task 3

Clear your `R` workspace. Set the seed to (A) 2021, (B) 123, (C) 456, (D) 789. Please execute the function set.seed() with the requested seed every time you generate random numbers etc.
Set $n = 100$. Consider a GLM for independent binomial responses Y_i, $i = 1, \ldots, n$, with canonical link and linear predictor

$$\eta_i = 3 + 2 \cdot x_i \tag{2}$$

Sample n independent observations x_i, $i = 1, \ldots, n$ from a random variable $X \sim \mathcal{N}(0, 1)$.

(a) Calculate the n corresponding values y_i for the binomial response variable Y_i of the GLM considered above. What is the mean value of the resulting values for y_i ($i = 1, ..., n$)? **(requested precision: 2 digits)**

(b) Fit a logistic regression model that predicts the response variable Y using X as explanatory variable and based on the above simulated x-values. Let β_1 denote the true coefficient of the explanatory variable X. What is the standard error of the estimate $\hat{\beta}_1$? **(requested precision: 2 digits)**

(c) Based on the model fitted in (b), compute a 90 % profile likelihood confidence interval for $\hat{\beta}_1$. **(requested precision: 2 digits)**

(d) What is the percentage of correct classified observations of the model fitted in (b) using as threshold $P(Y = 1) \geq 0.5$? **(requested precision: 1 digit)**

(e) Fit the logistic regression null model. What is the mean value of the fitted values of this null model? **(requested precision: 2 digits)**

(f) Compute the area under the curve (AUC) for the model fitted in (b) and the null model fitted in (e). What are the values for AUC? **(requested precision: 1 digit)** Which model do you prefer between these two in terms of AUC? Type in "1" for the model fitted in (b) and "2" for the model fitted in (e) (without quotation marks).

Solution

(a) set.seed((A) 2021, (B) 123, (C) 456, (D) 789)
x1=rnorm(n,0,1)
lin.pred=3+2*x1 # true model
mu = exp(lin.pred)/(1+exp(lin.pred)) y2 = rbinom(n,1,mu) # response variable
mean(y2)

(A) 0.83 (Dynexite: 0.83)
(B) 0.84 (Dynexite 0.84)
(C) 0.92 (Dynexite: 0.92)
(D) 0.88 (Dynexite: 0.88)

Alternative Solution
set.seed((A) 2021, (B) 123, (C) 456, (D) 789)
x1=rnorm(n,0,1)
lin.pred=3+2*x1 # true model
mu = exp(lin.pred)/(1+exp(lin.pred))
set.seed((A) 2021, (B) 123, (C) 456, (D) 789)
y2 = rbinom(n,1,mu) # response variable
mean(y2)

(A) 0.88 (Dynexite: 0.88)
(B) 0.92 (Dynexite 0.92)
(C) 0.86 (Dynexite: 0.86)
(D) 0.92 (Dynexite: 0.92)

(b) set.seed((A) 2021, (B) 123, (C) 456, (D) 789)
data2 = data.frame(y=y2,x1=x1)
model.2=glm(y ∼ x1,data=data2,family="binomial")
summary(model.2) # standard error for β_1 is

(A) 0.5864 (Dynexite: 0.59)
(B) 0.6602 (Dynexite 0.66)
(C) 0.5526 (Dynexite: 0.55)
(D) 0.4998 (Dynexite: 0.50)

Alternative Solution
set.seed((A) 2021, (B) 123, (C) 456, (D) 789)
data2 = data.frame(y=y2,x1=x1)
model.2=glm(y ∼ x1,data=data2,family="binomial")
summary(model.2) # standard error for β_1 is

(A) 0.4776 (Dynexite: 0.48)
(B) 0.987 (Dynexite 0.99)
(C) 0.4432 (Dynexite: 0.44)
(D) 0.7692 (Dynexite: 0.77)

(c) set.seed((A) 2021, (B) 123, (C) 456, (D) 789)
CI=confint(model.1, level=0.9)

CI[2,] #profile Likelihood confidence interval for $\hat{\beta}_1$ given by

(A) [1.621358, 3.577764] (Dynexite: [1.62,3.58])
(B)[1.672394, 3.865569] (Dynexite [1.67, 3.87])
(C) [0.8810776, 2.7277505] (Dynexite: [0.88,2.73])
(D) [0.9950024, 2.6661921] (Dynexite: [1,2.67])

Alternative Solution

set.seed((A) 2021, (B) 123, (C) 456, (D) 789)
CI=confint(model.1, level=0.9)
CI[2,] #profile Likelihood confidence interval for $\hat{\beta}_1$ given by

(A) [0.9349578, 2.5264154] (Dynexite: [0.93,2.53])
(B)[1.883119, 5.215146] (Dynexite [1.88,5.22])
(C) [0.9559483, 2.4286185] (Dynexite: [1.96,2.43])
(D) [1.390721, 3.982154] (Dynexite: [1.4,3.98])

(d) set.seed((A) 2021, (B) 123, (C) 456, (D) 789)
y2.pred=ifelse(model.2$fitted.values > 0.5, 1, 0)
tab1=table(data2$y,y2.pred)
sum(diag(tab1))/sum(tab1)

(A) 0.9 (Dynexite: 0.9)
(B) 0.87 (Dynexite 0.9)
(C) 0.92 (Dynexite: 0.9)
(D) 0.91 (Dynexite: 0.9)

Alternative Solution

set.seed((A) 2021, (B) 123, (C) 456, (D) 789)
y2.pred=ifelse(model.2$fitted.values > 0.5, 1, 0)
tab1=table(data2$y,y2.pred)
sum(diag(tab1))/sum(tab1)

(A) 0.89 (Dynexite: 0.89)
(B) 0.94 (Dynexite 0.94)
(C) 0.87(Dynexite: 0.87)
(D) 0.95 (Dynexite: 0.95)

(e) set.seed((A) 2021, (B) 123, (C) 456, (D) 789)
null.model=glm(y $\sim$ 1,data=data2,family="binomial")
mean(null.model $ fitted.values)

(A) 0.83 (Dynexite: 0.83)
(B) 0.84 (Dynexite 0.84)
(C) 0.92 (Dynexite: 0.92)
(D) 0.88 (Dynexite: 0.88)

Alternative Solution

set.seed((A) 2021, (B) 123, (C) 456, (D) 789)

null.model=glm(y ∼ 1,data=data2,family="binomial")
mean(null.model $ fitted.values)

(A) 0.88 (Dynexite: 0.88)
(B) 0.92 (Dynexite 0.92)
(C) 0.86 (Dynexite: 0.86)
(D) 0.92 (Dynexite: 0.92)

(f) set.seed((A) 2021, (B) 123, (C) 456, (D) 789)
roc.curve1=roc(y ∼ fitted(model.2), data=data2)
roc.curve2=roc(y ∼ fitted(null.model), data=data2)
auc(roc.curve1)

(A) 0.9093 (Dynexite: 0.9)
(B) 0.8899 (Dynexite 0.9)
(C) 0.837 (Dynexite: 0.8)
(D) 0.8352 (Dynexite: 0.8)

auc(roc.curve2)

(A) 0.5 (Dynexite: 0.5)
(B) 0.5 (Dynexite 0.5)
(C) 0.5 (Dynexite: 0.5)
(D) 0.5 (Dynexite: 0.5)

⇒ in all cases we prefer the model fitted in (b) so the answer is **"1"**.

Alternative Solution
set.seed((A) 2021, (B) 123, (C) 456, (D) 789)
roc.curve1=roc(y ∼ fitted(model.2), data=data2)
roc.curve2=roc(y ∼ fitted(null.model), data=data2)
auc(roc.curve1)

(A) 0.8333(Dynexite: 0.83)
(B)0.9457(Dynexite 0.95)
(C) 0.8472 (Dynexite: 0.85)
(D)0.8818(Dynexite: 0.88)

auc(roc.curve2)

(A)0.5 (Dynexite: 0.5)
(B) 0.5 (Dynexite 0.5)
(C) 0.5 (Dynexite: 0.5)
(D) 0.5 (Dynexite: 0.5)

⇒ in all cases we prefer the model fitted in (b) so the answer is **"1"**.

Part II: Theory Tasks

Task 1 (Theory Task 1)

Let $\boldsymbol{X} \sim \mathcal{N}_2(\boldsymbol{\mu}, \Sigma)$ be a normally distributed random vector with

$$\boldsymbol{\mu} = \begin{pmatrix} 1 \\ -1 \end{pmatrix} \quad \text{and} \quad \Sigma = \frac{1}{2} \cdot \begin{pmatrix} 17 & 15 \\ 15 & 17 \end{pmatrix} = \begin{pmatrix} \frac{17}{2} & \frac{15}{2} \\ \frac{15}{2} & \frac{17}{2} \end{pmatrix}.$$

The singular value decomposition (SVD) of Σ is given by $\Sigma = V\Lambda V'$, with

$$V = \begin{pmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix} \quad \text{and} \quad \Lambda = \begin{pmatrix} 16 & 0 \\ 0 & 1 \end{pmatrix}$$

(no proof required).

(a) Let the row matrix $A \in \mathbb{R}^{1\times 2}$ be given by

$$A = \begin{pmatrix} -1 & 2 \end{pmatrix}$$

and define $Y = 3 + A\boldsymbol{X}$. Then, the random variable Y is univariate normally distributed, i.e. $Y \sim \mathcal{N}_1(\nu, a^2)$. Find $\nu \in \mathbb{R}$ and $a^2 > 0$.

Solution: By properties of normal distribution

$$\nu = 3 + A\boldsymbol{\mu} = 3 + \begin{pmatrix} -1 & 2 \end{pmatrix} \begin{pmatrix} 1 \\ -1 \end{pmatrix} = 3 - 3 = 0$$

and

$$a^2 = A\Sigma A' = \begin{pmatrix} -1 & 2 \end{pmatrix} \frac{1}{2} \cdot \begin{pmatrix} 17 & 15 \\ 15 & 17 \end{pmatrix} \begin{pmatrix} -1 \\ 2 \end{pmatrix} = \frac{1}{2} \begin{pmatrix} 13 & 19 \end{pmatrix} \begin{pmatrix} -1 \\ 2 \end{pmatrix} = \frac{25}{2} = 12.5.$$

(b) Consider the matrix A defined in (a) and the row matrix B given by

$$B = \begin{pmatrix} 19 & b \end{pmatrix}$$

with some $b \in \mathbb{R}$. Find $b \in \mathbb{R}$ so that $A\boldsymbol{X}$ and $B\boldsymbol{X}$ are independent random vectors.

Solution: According to I.2.14 independence holds if and only if

$$A\Sigma B' = 0 \quad \overset{(a)}{\iff} \quad \begin{pmatrix} 13 & 19 \end{pmatrix} \begin{pmatrix} 19 \\ b \end{pmatrix} = 0 \quad \iff 19 \cdot 13 + 19b = 0$$

such that $b = -13$.

(c) Consider the random vector $\boldsymbol{Z} = \Sigma^{-1/2}(\boldsymbol{X} - \boldsymbol{\mu}) + \boldsymbol{\mu}$ and the matrix

$$C = \begin{pmatrix} 3 & -3 \\ -3 & 3 \end{pmatrix}.$$

Find the (uniquely determined) constant $c \in \mathbb{R}\backslash\{0\}$ so that

$$c \cdot \boldsymbol{Z}'C\boldsymbol{Z} \sim \chi^2(p, \delta)$$

is χ^2-distributed. In particular, compute the degrees of freedom $p \in \mathbb{N}$ and the value of the non-centrality parameter $\delta \geq 0$.

Solution: According to Theorem I.3.5 $Q = c \cdot C$ needs to be an orthogonal projector, which is true for $c = \frac{1}{6}$, since then Q is both symmetric and idempotent. Furthermore $p = 1$ since $\operatorname{rank}(C) = 1$ and

$$\delta = \frac{1}{2}\boldsymbol{\mu}'Q\boldsymbol{\mu} = \frac{1}{12} \begin{pmatrix} 1 & -1 \end{pmatrix} \begin{pmatrix} 3 & -3 \\ -3 & 3 \end{pmatrix} \begin{pmatrix} 1 \\ -1 \end{pmatrix} = \frac{1}{12} \begin{pmatrix} 1 & -1 \end{pmatrix} \begin{pmatrix} 6 \\ -6 \end{pmatrix} = 1.$$

Task 2 (Theory Task 2)

Consider the linear model

$$\boldsymbol{Y} = B\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{3}$$

with

$$B = (x_{ij})_{i=1,2,3;j=1,2} = \begin{pmatrix} 1 & 1 \\ 1 & -2 \\ 1 & 1 \end{pmatrix}, \qquad \boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} \in \mathbb{R}^2, \qquad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \end{pmatrix} \sim \mathcal{N}_3(\mathbf{0}, \sigma^2 I_3),\ \sigma^2 > 0.$$

Denote by $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \hat{\beta}_2)$ the least squares estimator (LSE) for $\boldsymbol{\beta}$.

(a) Let $\gamma = 3\beta_1 - 6\beta_2$ and $\hat{\gamma} = 3\hat{\beta}_1 - 6\hat{\beta}_2$ be the respective LSE. An exact lower $(1-\alpha)$-confidence interval for $\hat{\gamma} = 3\hat{\beta}_1 - 6\hat{\beta}_2$ is given by

$$I_{\beta_1} = \left(-\infty, \hat{\gamma} + q(\alpha) \cdot ||\boldsymbol{Y} - B\hat{\boldsymbol{\beta}}|| \cdot d\right]$$

with appropriate choice of $q(\alpha)$ and d; $q(\alpha)$ denotes a quantile of an appropriate distribution; $||\mathbf{z}|| = \sqrt{\mathbf{z}'\mathbf{z}}$.

For $\alpha = 0.1$ and the vector of observations $\boldsymbol{y} = (1, -1, 3)'$, determine the values of $\hat{\gamma}$, $q(\alpha)$ and d.

Solution: According to I.4.12 we have

$$\hat{\boldsymbol{\beta}} = (B'B)^{-1}B'\boldsymbol{y}.$$

Since

$$B'B = \begin{pmatrix} 3 & 0 \\ 0 & 6 \end{pmatrix} \Longrightarrow (B'B)^{-1} = \begin{pmatrix} \frac{1}{3} & 0 \\ 0 & \frac{1}{6} \end{pmatrix}$$

the estimates are given by

$$\hat{\boldsymbol{\beta}} = \begin{pmatrix} \frac{1}{3} & 0 \\ 0 & \frac{1}{6} \end{pmatrix} \begin{pmatrix} 1 & 1 & 1 \\ 1 & -2 & 1 \end{pmatrix} \begin{pmatrix} 1 \\ -1 \\ 3 \end{pmatrix} = \begin{pmatrix} \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{6} & -\frac{1}{3} & \frac{1}{6} \end{pmatrix} \begin{pmatrix} 1 \\ -1 \\ 3 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

and in particular $\hat{\gamma} = 3 - 6 = -3$.

Furthermore, according to I.4.32 with the choice $\boldsymbol{c} = (3, -6)'$ and by symmetry of the t-distribution we get

$$q(\alpha) = t_{0,9}(1) \approx 3{,}078$$

and

$$d^2 = \begin{pmatrix} 3 & -6 \end{pmatrix} \begin{pmatrix} \frac{1}{3} & 0 \\ 0 & \frac{1}{6} \end{pmatrix} \begin{pmatrix} 3 \\ -6 \end{pmatrix} = \begin{pmatrix} 3 & -6 \end{pmatrix} \begin{pmatrix} 1 \\ -1 \end{pmatrix} = 9.$$

Thus, $d = 3$.

(b) Consider the testing problem

$$H_0 \colon \beta_1 = 0 \quad \longleftrightarrow \quad H_1 \colon \beta_1 \neq 0.$$

Then, there exists an α-level statistical test for H_0 whose decision rule can be formulated as

$$\text{Reject } H_0 \text{ if } \frac{\boldsymbol{Y}' A_0 \boldsymbol{Y}}{\boldsymbol{Y}' A \boldsymbol{Y}} > c(\alpha)$$

for some appropriate orthogonal projectors A_0, A, and an appropriately chosen critical value $c(\alpha)$, respectively. List the diagonal elements of $A_0 = (a_{ij}^{(0)})_{i,j}$ and $A = (a_{ij})_{i,j}$, respectively, and find the critical value $c(\alpha)$ for $\alpha = 0.05$.

Solution: Testing the null hypothesis is equivalent to testing the reduced model with design matrix

$$B_0 = \begin{pmatrix} 1 \\ -2 \\ 1 \end{pmatrix}$$

and proceeding according to Theorem I.4.39 we get

$$Q = B(B'B)^{-1}B' \overset{(a)}{=} \begin{pmatrix} 1 & 1 \\ 1 & -2 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{6} & -\frac{1}{3} & \frac{1}{6} \end{pmatrix} = \begin{pmatrix} \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & 1 & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} \end{pmatrix}$$

and

$$Q_0 = B_0(B_0'B_0)^{-1}B_0' = \frac{1}{6}B_0B_0' = \frac{1}{6}\begin{pmatrix} 1 & -2 & 1 \\ -2 & 4 & -2 \\ 1 & -2 & 1 \end{pmatrix}$$

Thus,

$$A_0 = Q - Q_0 = \begin{pmatrix} \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{pmatrix} = \frac{1}{3}\mathbb{1}_{3\times 3}$$

and

$$A = I_n - Q = \begin{pmatrix} \frac{1}{2} & 0 & -\frac{1}{2} \\ 0 & 0 & 0 \\ -\frac{1}{2} & 0 & \frac{1}{2} \end{pmatrix}$$

and

$$c(\alpha) = F_{0,95}(1,1) \approx 161{,}45.$$

(c) Consider, instead, the model

$$Y_i = x_{i1}\beta_1 + x_{i2}^2\beta_2 + \varepsilon_i, \quad i = 1,2,3. \tag{4}$$

with x_{ij}, $i = 1,2,3$, $j = 1,2$ as above. Then, an α-level statistical test for the hypotheses

$$H_0 \colon \beta_1 = 0 \quad \longleftrightarrow \quad H_1 \colon \beta_1 \neq 0$$

can be formulated in terms of quantiles of the t-distribution by the decision rule

$$\text{Reject } H_0 \text{ if } \left| \frac{\hat{\beta}_1}{d_* \cdot ||\boldsymbol{Y} - B\hat{\boldsymbol{\beta}}||} \right| > t_{1-\alpha/2}(n),$$

where $t_\beta(n)$, $\beta \in (0,1)$, denotes the β-quantile of the t-distribution with $n \in \mathbb{N}$ degrees of freedom. Find n and calculate d_* for model (2).

Solution: We get the transformed design matrix

$$B_* = \begin{pmatrix} 1 & 1 \\ 1 & 4 \\ 1 & 1 \end{pmatrix}.$$

Following I.5.18 we get $n = 1$ and d_* is the $(1,1)$-element of the matrix $(B_*'B_*)^{-1}$, i.e.

$$d_*^2 = ((B_*'B_*)^{-1})_{11}.$$

Since

$$B_*'B_* = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 4 & 1 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 1 & 4 \\ 1 & 1 \end{pmatrix} = \begin{pmatrix} 3 & 6 \\ 6 & 18 \end{pmatrix} \Longrightarrow (B_*'B_*)^{-1} = \frac{1}{18} \begin{pmatrix} 18 & -6 \\ -6 & 3 \end{pmatrix}$$

we have $d_* = 1$.

Task 3 (Theory Task 3)

Let a family of distributions be given by their pdfs (probability density functions) defined for $\lambda > 0$, $k \in \mathbb{N}$ as

$$f(x; \lambda, k) = \frac{\lambda^k}{(k-1)!} x^{k-1} \exp\{-\lambda x\}, \quad x > 0. \tag{5}$$

For fixed (known) $k \in \mathbb{N}$, $f_k(\,\cdot\,; \lambda) = f(\,\cdot\,; \lambda, k)$ defines a subfamily of the exponential dispersion family (EDF) of distributions.

(a) Let $k = 4$ and $X \sim f_4(\cdot; \lambda)$

(i) Determine the value of the natural parameter θ and the dispersion $a(\phi)$, when $\lambda = 2$.

Solution: We have

$$\begin{aligned} \ln(f(x; \lambda, k)) =& k \ln(\lambda) - \ln((k-1)!) + (k-1)\ln(x) - \lambda x \\ =& \frac{-\lambda x + k \ln(\lambda)}{1} + (k-1)\ln(x) - \ln((k-1)!) \end{aligned}$$

Thus, $\theta = -\lambda = -2$ and $a(\phi) = 1$.

Alternative: $\theta = -\frac{\lambda}{k} = -0.5$ and $a(\phi) = \frac{1}{k} = 0.25$.
Remark: In general, solutions of the type $\theta = -2c$ and $a(\phi) = c$, $c > 0$, with $a(\phi)$ independent of λ are also correct. These alternative solutions lead to the same results in (ii) and (iii) (c.f. below).

(ii) Calculate $\mathrm{E}(X)$, when $\lambda = 2$.

Solution: Since $b(\theta) = -k \ln(-\theta)$ we have

$$\mathrm{E}(X) = b'(\theta) = -\frac{k}{\theta} = \frac{k}{\lambda} = 2.$$

Alternative: For $\theta = -\frac{\lambda}{k}$, $b(\theta) = -\ln(-\theta) - \ln(k)$ and

$$\mathrm{E}(X) = b'(\theta) = -\frac{1}{\theta} = \frac{k}{\lambda} = 2.$$

(iii) Calculate $\mathrm{Var}(X)$, when $\lambda = 2$.

Solution:

$$\mathrm{Var}(X) = b''(\theta)a(\phi) = \frac{k}{\theta^2} \cdot 1 = \frac{k}{\lambda^2} = 1.$$

Alternative: For $\theta = -\frac{\lambda}{k}$, $b(\theta) = -\ln(-\theta) - \ln(k)$ and $a(\phi) = \frac{1}{k}$, thus

$$\mathrm{Var}(X) = b''(\theta)a(\phi) = \frac{1}{\theta^2} \cdot \frac{1}{k} = \frac{k}{\lambda^2} = 1.$$

(b) For modelling a response variable Y with $Y \sim f_k(\,\cdot\,;\lambda)$ for some fixed (known) $k \in \mathbb{N}$, consider a GLM with *canonical* link function $g(\cdot\,;k)$. Further suppose $\mu_1 = k$, $\mu_2 = 2k$. Calculate

$$g(\mu_1;k) - g(\mu_2;k).$$

Solution: Since $\theta = -\lambda$ and $\mu = \mathrm{E}(X) = -\frac{k}{\theta}$ we have

$$\theta = g(\mu;k) = -\frac{k}{\mu}$$

and thus

$$g(\mu_1;k) - g(\mu_2;k) = -\frac{k}{k} + \frac{k}{2k} = -\frac{1}{2}.$$

Alternative: In that case $g(\mu) = -\mu^{-1}$ and

$$g(\mu_1;k) - g(\mu_2;k) = -\frac{1}{k} + \frac{1}{2k} = -\frac{1}{2k} = -0.125.$$

Remark: In general, the alternative solution has to be consistent with the choice of parametrization in (a)(i).

(c) Consider independent random measurements $Y_i \sim f_k(\cdot;\lambda_i)$, $i \in \{1,2,3\}$, for some fixed $k \in \mathbb{N}$ satisfying a GLM with *canonical* link g such that

$$g(\mathrm{E}(Y_i)) = \beta_1 x_{i1} + \beta_2 x_{i2}, \quad i = 1,2,3$$

with model parameters $\beta_1, \beta_2 \in \mathbb{R}$ and design matrix

$$\boldsymbol{X} = (x_{ij})_{i=1,2,3;j=1,2} = \begin{pmatrix} 1 & 1 \\ 0 & -3 \\ 1 & 1 \end{pmatrix}$$

Calculate the Fisher information matrix

$$\mathcal{I}_F = \begin{pmatrix} v_{11} & v_{12} \\ v_{21} & v_{22} \end{pmatrix}$$

with respect to the parameter vector $\boldsymbol{\beta} = (\beta_1, \beta_2)'$ when $k = 10$ and $\lambda_i = i^2 + 1$, $i \in \{1,2,3\}$.

Solution: According to (a) we have

$$\theta = -\lambda, \quad a(\phi) = 1, \quad b(\theta) = -k\ln(-\theta) \quad \text{and} \quad b''(\theta) = \frac{k}{\theta^2} = \frac{k}{\lambda^2}.$$

Then, by assumption of the canonical link and Theorem II.2.24

$$\mathcal{I}_F = \boldsymbol{X}'\boldsymbol{W}_c\boldsymbol{X}$$

with $\boldsymbol{W}_c = \operatorname{diag}(b''(\theta_1), b''(\theta_2), b''(\theta_3))$. Accordingly

$$\begin{aligned}\mathcal{I}_F &= \begin{pmatrix}1 & 0 & 1\\ 1 & -3 & 1\end{pmatrix}\begin{pmatrix}5/2 & 0 & 0\\ 0 & 2/5 & 0\\ 0 & 0 & 1/10\end{pmatrix}\begin{pmatrix}1 & 1\\ 0 & -3\\ 1 & 1\end{pmatrix}\\ &= \begin{pmatrix}5/2 & 0 & 1/10\\ 5/2 & -6/5 & 1/10\end{pmatrix}\begin{pmatrix}1 & 1\\ 0 & -3\\ 1 & 1\end{pmatrix} = \begin{pmatrix}2.6 & 2.6\\ 2.6 & 6.2\end{pmatrix}\end{aligned}$$

. **Alternative**: According to (a) we have

$$\theta = -\frac{\lambda}{k}, \quad a(\phi) = \frac{1}{k}, \quad b(\theta) = -\ln(-\theta) - \ln(k) \quad \text{and} \quad b''(\theta) = \frac{1}{\theta^2} = \frac{k^2}{\lambda^2}.$$

So

$$b''(\theta)/a(\phi) = \frac{1}{\theta^2}\cdot k = \frac{k}{\theta^2} = \frac{k^3}{\lambda^2}.$$

Accordingly

$$\begin{aligned}\mathcal{I}_F &= \begin{pmatrix}1 & 0 & 1\\ 1 & -3 & 1\end{pmatrix}\begin{pmatrix}250 & 0 & 0\\ 0 & 40 & 0\\ 0 & 0 & 10\end{pmatrix}\begin{pmatrix}1 & 1\\ 0 & -3\\ 1 & 1\end{pmatrix}\\ &= \begin{pmatrix}250 & 0 & 10\\ 250 & -120 & 10\end{pmatrix}\begin{pmatrix}1 & 1\\ 0 & -3\\ 1 & 1\end{pmatrix} = \begin{pmatrix}260 & 260\\ 260 & 620\end{pmatrix}.\end{aligned}$$

Remark: In general, the FI matrix has to be consistent with the choice of parametrization in (a)(i) and be proportional to the one(s) given above, (regardless of the chosen parametrization).

(d) For $n \in \mathbb{N}$ and some fixed $\boldsymbol{\beta} \in \mathbb{R}^2$ let independent random observations $Y_1, \ldots, Y_n$ with $Y_i \sim f_k(\cdot; \lambda_i), i \in \{1, \ldots, n\}$, be given. Suppose the asymptotic distribution of the associated sequence of MLEs $\hat{\boldsymbol{\beta}}_n = (\hat{\beta}_{1n}, \hat{\beta}_{2n})'$ is given by

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \xrightarrow{d} \mathcal{N}_2(\mathbf{0}, \Sigma) \quad \text{as } n \to \infty, \quad \text{with} \quad \Sigma = \begin{pmatrix}2 & 1\\ 1 & 4\end{pmatrix}.$$

Derive the asymptotic variance σ^2, say, of

$$\sqrt{n}\left(\hat{\beta}_{1n}^2 + \frac{1}{2}\hat{\beta}_{2n}\right) \quad \text{as } n \to \infty$$

when $\boldsymbol{\beta} = (1, -1)'$ by applying the Delta method.

Solution: Applying the Delta method with

$$g\colon \mathbb{R}^2 \to \mathbb{R}, (\beta_1, \beta_2) \mapsto \beta_1^2 + \frac{1}{2}\beta_2$$

and respective matrix of partial derivatives

$$D_g(\boldsymbol{\beta}) = (2\beta_1, \frac{1}{2})$$

we get for $\boldsymbol{\beta} = (1, -1)'$

$$\sigma^2 = \begin{pmatrix} 2 & \frac{1}{2} \end{pmatrix} \Sigma \begin{pmatrix} 2 \\ \frac{1}{2} \end{pmatrix} = \begin{pmatrix} \frac{9}{2} & 4 \end{pmatrix} \begin{pmatrix} 2 \\ \frac{1}{2} \end{pmatrix} = 11.$$

Task 4 (Theory Task 4)

Consider the following seven statements. Which of them are in general true?

(1) If the model matrix $\boldsymbol{X}$ of a GLM $g(\mathrm{E}(\boldsymbol{Y})) = \boldsymbol{X}\boldsymbol{\beta}$ is of full rank, then the MLE $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ exists uniquely.

(2) A step-wise GLM selection procedure among nested models may lead to a different model when model comparison is based on AIC instead of the deviance measure.

(3) The error terms of a binary logistic regression model are binomially distributed.

(4) Consider a binary response Y with $\pi_i = P(Y = 1 \mid X = i)$, where X is an ordinal explanatory variable of I levels. Supose that π_i satisfy the linear logit model

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 x_i, \quad i = 1, \ldots, I \qquad (*)$$

with $x_1 \leq \ldots \leq x_I$ ($x_1 < x_I$) known scores assigned to the categories of X.

Then, the probabilities π_i of model $(*)$ are invariant under a linear transformation of the scores x_i to $\tilde{x}_i$, $i = 1, \ldots, I$, with $\tilde{x}_1 \neq \tilde{x}_I$.

(5) The log-link is the canonical link for a negative binomial GLM.

(6) For the population having binary responses Y, suppose the conditional distribution of X given $Y = y$ is $\mathcal{N}_1(\mu_y, \sigma^2)$, $y = 0, 1$. Then $P(Y = 1 \mid X = x)$ satisfies the logistic regression model with $\beta_1 = (\mu_1 - \mu_0)/\sigma^2$.

(7) Consider $n \in \mathbb{N}$ independent Poisson distributed responses $Y_i \sim \mathcal{P}(\mu_i)$, $i = 1, \ldots, n$, $\mu_i > 0$, satisfying the Poisson regression model

$$\log(\mu_i) = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}.$$

For the null model, the MLE of β_0, based on an observed sample $y_1, \ldots, y_n$, is $\hat{\beta}_0 = \frac{1}{n}\log(\bar{y})$, where $\bar{y}$ is the sample mean.

Provide the item-numbers of correct statements in increasing order.

For example, if statements 1,5,6,7 are correct, then type 1567.

6 points will be given if and only if all correct statements are identified and none false statement is given as correct. 3 points will be given if at least 50% of the correct statements are identified and none false. Hence in the example above, the response 1567 gets 6 points; 16 or 167 get 3 points while 12567, 146, 1367 get 0 points.

Solution:

(1) In general existence and uniqueness not only depend on $\boldsymbol{X}$ but also on the distribution from the EDF. Thus, while regularity of $\boldsymbol{X}$ is a necessary condition, it is not sufficient (c.f. Remark II.2.25).

(2) Following II.2.35/36/41 and II.2.39 we can see that the deviance decreases for an increasing number of predictors, while the AIC penalizes model complexity.

Furthermore, according to II.2.42 we can see for the particular case $\phi = 1$, that comparing the difference of deviances or AICs, respectively, for two nested models $\mathcal{M}_0 \subset \mathcal{M}_1$, yields a difference of $2(p_0 - p_1)$, where p_j is the number of parameters of model $j \in \{0, 1\}$.

(3) According to II.2.32 and II.3.3 we get for the residuals

$$\hat{\boldsymbol{\varepsilon}} = (\hat{\varepsilon}_1, \ldots, \hat{\varepsilon}_n)' = (Y_1 - \pi_1(\boldsymbol{x}_1\hat{\boldsymbol{\beta}}), \ldots, Y_n - \pi_n(\boldsymbol{x}_n\hat{\boldsymbol{\beta}}))' = \boldsymbol{Y} - \hat{\boldsymbol{\pi}}(\boldsymbol{X})$$

where $\boldsymbol{x}_i$ is the i-th row of the predictor $\boldsymbol{X}$, $i = 1, \ldots, n$, and $\hat{\boldsymbol{\pi}}(\boldsymbol{X}) = g^{-1}\left(\boldsymbol{X}\hat{\boldsymbol{\beta}}\right)$ with canonical (logit) link g (c.f. II.2.32).

In general $\hat{\varepsilon}_i \in [0, 1]$, $i = 1, \ldots, n$, and not in $\{0, 1\}$. Hence the residuals cannot be binomially distributed.

(4) For the linear logit model we have

$$Y_i \mid X_i = x_i \sim \mathcal{B}(n_i, \pi_i)$$

with

$$\pi_i = \text{logit}^{-1}(\beta_0 + \beta_1 x_i).$$

Considering the likelihood

$$l(\boldsymbol{\beta} \mid \boldsymbol{y}) \propto \prod_{i=1}^{I} \left(\frac{\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)}\right)^{y_i} \left(\frac{1}{1 + \exp(\beta_0 + \beta_1 x_i)}\right)^{n_i - y_i}$$

any linear transformation of the x_i implies an inverse transformation for β_1 to get the same maximum of the likelihood as before. Due to the invariance property of the MLEs, the result follows.

Alternative Solution: Consider the model

$$\text{logit}(\tilde{\pi}_i) = \tilde{\beta}_0 + \tilde{\beta}_1 \tilde{x}_i$$

with

$$\tilde{x}_i = a + bx_i, \quad n \neq 0$$

Then for

$$\beta_0 = \tilde{\beta}_0 + a\tilde{\beta}_1$$

and

$$\beta_1 = b\tilde{\beta}_1$$

we have $\pi_i = \tilde{\pi}_i$.

(5) According to II.6.10 with $Y \sim \mathcal{NB}(\lambda, k)$

$$\log(P(Y = y)) = y \log(1 - \lambda) + k \log(\lambda) + \log(\Gamma(k + y)) - \log(\Gamma(k)) - \log(y!)$$

with

$$\mu = k\frac{1 - \lambda}{\lambda}$$

so that

$$\theta = \log(1 - \lambda) \quad \Longrightarrow \theta = g(\mu) = \log\left(\frac{\mu}{k}\right) - \log\left(1 + \frac{\mu}{k}\right) = \log\left(\frac{\mu}{\mu + k}\right).$$

See also Remark II.6.13.

(6) By Bayes' Theorem and definition of the probability density function (pdf) of the normal distribution f

$$\begin{aligned}
\operatorname{logit}(P(Y = 1 \mid X = x)) =& \log\left(\frac{f(x \mid y = 1)}{f(x \mid y = 0)}\right) + \operatorname{logit}(P(Y = 1)) \\
=& \operatorname{logit}(P(Y = 1)) - \frac{(x - \mu_1)^2}{2\sigma^2} + \frac{(x - \mu_0)^2}{2\sigma^2} \\
=& \underbrace{\operatorname{logit}(\mathrm{P(Y{=}1)}) + \frac{\mu_0^2 - \mu_1^2}{2\sigma^2}}_{=\beta_0} + \underbrace{\left(\frac{\mu_1 - \mu_0}{\sigma^2}\right)}_{=\beta_1} x
\end{aligned}$$

and in particular $\beta_1 = \frac{\mu_1 - \mu_0}{\sigma^2}$.

(7) For the null model

$$\log(\mu_i) = \beta_0, \quad i = 1, \ldots, n.$$

and according to Remark II.2.20 we get the likelihood equation

$$\sum_{i=1}^{n}(y_i - \mu_i) = 0 \quad \Leftrightarrow \quad \bar{y} = \exp(\beta_0) \quad \Leftrightarrow \quad \beta_0 = \log(\bar{y}).$$

Thus, the correct input is given by "246".