Prof. Dr. E. Cramer, Prof. Dr. M. Kateri, RWTH Aachen, SS 2021
28^{th} September 2021

Applied Data Analysis

PT 2

Exercise 1

Consider the linear model

$$\boldsymbol{Y} = B\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{1}$$

with

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

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

(a) Suppose $\boldsymbol{\beta} = (1, -1)'$ is fixed and $\sigma^2 = 1$.

(i) Let the matrix $A \in \mathbb{R}^{2\times 3}$ be given by

$$A = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \end{pmatrix}.$$

Then, $\boldsymbol{Z} = A\boldsymbol{Y}$, say, is (bivariate) normally distributed, i.e. $\boldsymbol{Z} \sim \mathcal{N}_2(\boldsymbol{\eta}, \Sigma)$. Find $\boldsymbol{\eta} \in \mathbb{R}^2$ and the trace of Σ.

Solution: Since $\boldsymbol{Y} \sim \mathcal{N}_3(B\boldsymbol{\beta}, I_3)$, by properties of the normal distribution

$$\boldsymbol{\eta} = AB\boldsymbol{\beta} = A \begin{pmatrix} 2 & \frac{1}{2} \\ 1 & -1 \\ 0 & \frac{1}{2} \end{pmatrix} \begin{pmatrix} 1 \\ -1 \end{pmatrix} = A \begin{pmatrix} \frac{3}{2} \\ 2 \\ -\frac{1}{2} \end{pmatrix} = \begin{pmatrix} 1 \\ \frac{3}{2} \end{pmatrix}$$

and

$$\Sigma = AA' = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$$

so that trace$(\Sigma) = 4$.

(ii) Consider the matrix A from (i) and the row matrix

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

Find the (uniquely determined) constant $c \in \mathbb{R}$ so that $A\boldsymbol{Y}$ and $C\boldsymbol{Y}$ are independent random vectors.

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

$$AI_3C' = 0 \quad \Leftrightarrow \quad \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \end{pmatrix} \begin{pmatrix} 1 \\ c \\ -1 \end{pmatrix} = 0 \quad \Leftrightarrow \quad c = 1.$$

Remark: A typo in Task (ii) results in 0.5 points for each participant.

(b) Let $\gamma = 5\beta_1 - 3\beta_2$ and $\hat{\gamma} = 5\hat{\beta}_1 - 3\hat{\beta}_2$ be the respective LSE. An exact upper $(1-\alpha)$-confidence interval for $\hat{\gamma}$ is given by

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

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 $\alpha = 0.01$ and the vector of observations $\boldsymbol{y} = (-1, 1, -2)'$, 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

$$B'B = \begin{pmatrix} 5 & 0 \\ 0 & \frac{3}{2} \end{pmatrix} \Longrightarrow (B'B)^{-1} = \begin{pmatrix} \frac{1}{5} & 0 \\ 0 & \frac{2}{3} \end{pmatrix}$$

the estimates are given by

$$\hat{\boldsymbol{\beta}} = \begin{pmatrix} \frac{1}{5} & 0 \\ 0 & \frac{2}{3} \end{pmatrix} \begin{pmatrix} 2 & 1 & 0 \\ \frac{1}{2} & -1 & \frac{1}{2} \end{pmatrix} \begin{pmatrix} -1 \\ 1 \\ -2 \end{pmatrix} = \begin{pmatrix} \frac{2}{5} & \frac{1}{5} & 0 \\ \frac{1}{3} & -\frac{2}{3} & \frac{1}{3} \end{pmatrix} \begin{pmatrix} -1 \\ 1 \\ -2 \end{pmatrix} = \begin{pmatrix} -\frac{1}{5} \\ -\frac{5}{3} \end{pmatrix}$$

and in particular $\hat{\gamma} = -1 + 5 = 4$.

Furthermore, according to I.4.32 with the choice $\boldsymbol{c} = (5, -3)'$ and the quantile table for the t-distribution we get

$$q(0.01) = t_{0.99}(1) \approx 31{,}82$$

and

$$d^2 = \begin{pmatrix} 5 & -3 \end{pmatrix} \begin{pmatrix} \frac{1}{5} & 0 \\ 0 & \frac{2}{3} \end{pmatrix} \begin{pmatrix} 5 \\ -3 \end{pmatrix} = \begin{pmatrix} 1 & -2 \end{pmatrix} \begin{pmatrix} 5 \\ -3 \end{pmatrix} = 11.$$

Thus, $d = \sqrt{11}$.

(c) Consider the testing problem

$$H_0\colon \beta_2 = 0 \quad \longleftrightarrow \quad H_1\colon \beta_2 \neq 0.$$

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

$$\text{Reject } H_0 \text{ if } \frac{\boldsymbol{Y}'Q_0^*\boldsymbol{Y}}{\boldsymbol{Y}'Q^*\boldsymbol{Y}} > c(\alpha)$$

for some appropriate orthogonal projectors Q_0^*, Q^*, and an appropriately chosen critical value $c(\alpha)$, respectively. List the diagonal elements of $Q_0^* = (q_{ij}^{(0)})_{i,j}$ and $Q^* = (q_{ij})_{i,j}$, respectively, and find the critical value $c(\alpha)$ for $\alpha = 0.1$.

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

$$B_0 = \begin{pmatrix} 2 \\ 1 \\ 0 \end{pmatrix}$$

and proceeding according to Theorem I.4.39 we get

$$Q = B(B'B)^{-1}B' \overset{(b)}{=} \begin{pmatrix} 2 & \frac{1}{2} \\ 1 & -1 \\ 0 & \frac{1}{2} \end{pmatrix} \begin{pmatrix} \frac{2}{5} & \frac{1}{5} & 0 \\ \frac{1}{3} & -\frac{2}{3} & \frac{1}{3} \end{pmatrix} = \begin{pmatrix} \frac{29}{30} & \frac{1}{15} & \frac{1}{6} \\ \frac{1}{15} & \frac{13}{15} & -\frac{1}{3} \\ \frac{1}{6} & -\frac{1}{3} & \frac{1}{6} \end{pmatrix}$$

and

$$Q_0 = B_0(B_0'B_0)^{-1}B_0' = \frac{1}{5}B_0B_0' = \frac{1}{5}\begin{pmatrix} 4 & 2 & 0 \\ 2 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

Thus,

$$Q_0^* = Q - Q_0 = \begin{pmatrix} \frac{1}{6} & -\frac{1}{3} & \frac{1}{6} \\ -\frac{1}{3} & \frac{2}{3} & -\frac{1}{3} \\ \frac{1}{6} & -\frac{1}{3} & \frac{1}{6} \end{pmatrix}$$

and

$$Q^* = I_n - Q = \begin{pmatrix} \frac{1}{30} & -\frac{1}{15} & -\frac{1}{6} \\ -\frac{1}{15} & \frac{2}{15} & \frac{1}{3} \\ -\frac{1}{6} & \frac{1}{3} & \frac{5}{6} \end{pmatrix}$$

and

$$c(\alpha) = F_{0,9}(1,1) \approx 39{,}86.$$

(d) Let $\sigma^2 = 1$ and define

$$\boldsymbol{W} = \begin{pmatrix} \sqrt{5} & 0 \\ 0 & \sqrt{\frac{3}{2}} \end{pmatrix} \hat{\boldsymbol{\beta}} = \begin{pmatrix} \sqrt{5}\hat{\beta}_1 \\ \sqrt{\frac{3}{2}}\hat{\beta}_2 \end{pmatrix}.$$

Consider the matrix

$$V = v \cdot \begin{pmatrix} 1 & 2 \\ 2 & 4 \end{pmatrix}, \quad v \in \mathbb{R}\backslash\{0\}.$$

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

$$\boldsymbol{W}'V\boldsymbol{W} \sim \chi^2(p, \delta)$$

is (non-centrally) χ^2-distributed and give the degrees of freedom $p \in \mathbb{N}$.

Solution: According to I.4.25

$$\hat{\boldsymbol{\beta}} \sim \mathcal{N}_2(\boldsymbol{\beta}, \sigma^2(B'B)^{-1})$$

so that $\boldsymbol{W} \sim \mathcal{N}_2(\boldsymbol{\nu}, I_2)$ with

$$\boldsymbol{\nu} = \begin{pmatrix} \sqrt{5} & 0 \\ 0 & \sqrt{\frac{3}{2}} \end{pmatrix} \boldsymbol{\beta}.$$

For $v = \frac{1}{5}$, V is an orthogonal projector, that is, $V = V'$ and $V = V^2$ so that by I.3.5 $\boldsymbol{W}'V\boldsymbol{W}$ has a non-central χ^2-distribution. Since $\mathrm{trace}(V) = \mathrm{rank}(V) = 1$, we get $p = 1$.

Exercise 2

Let a family of distributions be given by their pdfs (probability density functions) defined for $\lambda > 0$ as

$$f(x;\lambda) = \lambda e^{-\lambda(x-3)}, \quad x > 3. \tag{2}$$

$f(\,\cdot\,;\lambda)$ defines a subfamily of the exponential dispersion family (EDF) of distributions with $a(\phi) = 1$.

(a) Let $X \sim f(\cdot;\lambda)$.

(i) Determine the value of the natural parameter θ when $\lambda = 4$.

Solution: We have

$$\ln(f(x;\lambda)) = -\,\lambda x + \ln(\lambda) + 3\lambda = -\lambda x - (-\ln(\lambda) - 3\lambda).$$

Thus, $\theta = -\lambda = -4$ and $b(\theta) = -\ln(-\theta) + 3\theta$ with $a(\phi) = 1$.

(ii) Calculate $\mathrm{E}(X)$, when $\lambda = 4$.

Solution: Since $b(\theta) = -\ln(-\theta) + 3\theta$ we have

$$\mathrm{E}(X) = b'(\theta) = -\frac{1}{\theta} + 3 = \frac{1}{\lambda} + 3 = 3.25.$$

(iii) Calculate $\mathrm{Var}(X)$, when $\lambda = 4$.

Solution:

$$\mathrm{Var}(X) = b''(\theta)a(\phi) = \frac{1}{\theta^2}\cdot 1 = \frac{1}{\lambda^2} = \frac{1}{16}.$$

(b) For modelling independent response variables Y_i with $Y_i \sim f(\,\cdot\,;\lambda_i)$, consider a GLM with *canonical* link function $g(\,\cdot\,)$ so that

$$g(\mu_i) = g(\mathrm{E}(Y_i)) = \beta_0 + \sum_{j=1}^{p}\beta_j x_{ij}, \quad i = 1,\dots,n,$$

with model parameters $\beta_0, \beta_1, \dots, \beta_p \in \mathbb{R}$, $p \in \mathbb{N}$.

Suppose we have observed the responses

$$y_1 = 6, \quad y_2 = 4, \quad y_3 = 5$$

with $n = 3$. Derive the maximum likelihood estimates $\hat{\mu}_1$ and $\hat{\beta}_0$ of μ_1 and β_0, respectively, under the null model.

Solution: Since $\mu = \mathrm{E}(X) = -\frac{1}{\theta}$ we have

$$\theta = g(\mu) = -\frac{1}{\mu}.$$

Thus, under the null model

$$-\mu_i^{-1} = \beta_0, \quad i = 1, 2, 3.$$

Then, according to II.2.19 and the assumption of the canonical link we get the likelihood equation

$$\sum_{i=1}^{3}(y_i - \mu_i) = 0 \quad \Leftrightarrow \quad -\beta_0^{-1} = \bar{y} \quad \Leftrightarrow \quad \beta_0 = -\bar{y}^{-1}$$

for β_0.

Thus, $\hat{\beta}_0 = -\frac{1}{5}$ and due to the invariance property of MLEs $\hat{\mu}_i = -\hat{\beta}_0^{-1} = 5$, independent of $i = 1, 2, 3$.

(c) Consider independent random measurements $Y_i \sim f(\cdot;\lambda_i)$, $i \in \{1, 2, 3\}$, 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

$$\boldsymbol{X} = (x_{ij})_{i=1,2,3;j=1,2} = \begin{pmatrix} 1 & 2 \\ 1 & 4 \\ 0 & -1 \end{pmatrix}$$

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 $\lambda_i = \frac{i}{i+1}$, $i \in \{1, 2, 3\}$.

Solution: According to (a) we have

$$\theta = -\lambda, \quad a(\phi) = 1, \quad b(\theta) = -\ln(-\theta) + 3\theta \quad \text{and} \quad b''(\theta) = \frac{1}{\theta^2} = \frac{1}{\lambda^2}.$$

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

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

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

$$\begin{aligned}\mathcal{I}_F &= \begin{pmatrix} 1 & 1 & 0 \\ 2 & 4 & -1 \end{pmatrix} \begin{pmatrix} 4 & 0 & 0 \\ 0 & \frac{9}{4} & 0 \\ 0 & 0 & \frac{16}{9} \end{pmatrix} \begin{pmatrix} 1 & 2 \\ 1 & 4 \\ 0 & -1 \end{pmatrix} \\ &= \begin{pmatrix} 4 & \frac{9}{4} & 0 \\ 8 & 9 & -\frac{16}{9} \end{pmatrix} \begin{pmatrix} 1 & 2 \\ 1 & 4 \\ 0 & -1 \end{pmatrix} = \begin{pmatrix} \frac{25}{4} & 17 \\ 17 & \frac{484}{9} \end{pmatrix}.\end{aligned}$$

Exercise 3

Suppose $Y_1, \ldots, Y_n$, $n \in \mathbb{N}$, is a sequence of independent Poisson distributed random counts corresponding to a random sample of n items. They are modeled by a GLM with link function g and based on $p \in \mathbb{N}$ explanatory variables $X_1, \ldots, X_p$, for which the fixed values for the i-th item are $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})$, $i \in \{1, \ldots, n\}$, so that

$$g(\mathrm{E}(Y_i)) = g(\mu_i) = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}, \quad i = 1, \ldots, n,$$

with model parameters $\beta_0, \beta_1, \ldots, \beta_p \in \mathbb{R}$, $p \in \mathbb{N}$.

(a) Let $p = 1$ and $\boldsymbol{\beta} = (\beta_0, \beta_1)'$ be fixed. Assume that the asymptotic distribution of the associated sequence of MLEs $\hat{\boldsymbol{\beta}}_n = (\hat{\beta}_{0n}, \hat{\beta}_{1n})'$, $n \in \mathbb{N}$, is given by

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \xrightarrow{d} \mathcal{N}_2(\mathbf{0}, \Sigma) \quad \text{as } n \to \infty \ , \quad \text{with} \quad \Sigma = \begin{pmatrix} 4 & 2 \\ 2 & 3 \end{pmatrix}.$$

Derive the asymptotic variance σ^2, say, of

$$\sqrt{n}(\hat{\beta}_{0n} + \hat{\beta}_{1n}^2) \quad \text{as } n \to \infty$$

by applying the Delta method assuming $\boldsymbol{\beta} = (1, 2)'$.

Solution: Applying the Delta method with

$$g \colon \mathbb{R}^2 \to \mathbb{R}, (\beta_0, \beta_1) \mapsto \beta_0 + \beta_1^2$$

and respective matrix of partial derivatives

$$D_g(\boldsymbol{\theta}) = \begin{pmatrix} 1 & 4 \end{pmatrix}$$

we get

$$\sigma^2 = \begin{pmatrix} 1 & 4 \end{pmatrix} \Sigma \begin{pmatrix} 1 \\ 4 \end{pmatrix} = \begin{pmatrix} 12 & 14 \end{pmatrix} \begin{pmatrix} 1 \\ 4 \end{pmatrix} = 68.$$

(b) Let $n = 3$ and suppose the link function g is (*non-canonical* and) given by

$$g(\mu_i) = \mu_i^2, \quad i = 1, 2, 3.$$

(i) Suppose we have observed the counts

$$y_1 = 3, \quad y_2 = 2, \quad y_3 = 4.$$

Calculate the maximum likelihood estimate $\hat{\beta}_0$ with respect to the parameter β_0 under the null model.

Solution: For the null model

$$\mu_i^2 = \beta_0, \quad i = 1, 2, 3,$$

and for the Poisson distribution $\mathrm{E}(Y_i) = \mathrm{Var}(Y_i) = \mu_i$.
Thus, following II.2.16 we get for the likelihood equation with $\eta_i = g(\mu_i)$ and $x_{i0} = 1$ for all $i = 1, 2, 3$

$$\begin{aligned}
& \sum_{i=1}^{3} \left(\frac{y_i - \mathrm{E}(Y_i)}{\mathrm{Var}(Y_i)} \cdot \frac{\partial g^{-1}(\eta_i)}{\partial \eta_i} \right) x_{i0} = 0 \\
\Leftrightarrow \quad & \sum_{i=1}^{3} \left(\frac{y_i - \mu_i}{\mu_i} \cdot \frac{1}{g'(\mu_i)} \right) = 0 \\
\Leftrightarrow \quad & \sum_{i=1}^{3} \left(\frac{y_i - \sqrt{\beta_0}}{\mu_i} \cdot \frac{1}{2\mu_i} \right) = 0 \\
\Leftrightarrow \quad & \sum_{i=1}^{3} \left(\frac{y_i - \sqrt{\beta_0}}{2\beta_0} \right) = 0 \qquad (*) \\
\Leftrightarrow \quad & \bar{y} = \sqrt{\beta_0} \\
\Leftrightarrow \quad & \beta_0 = \bar{y}^2.
\end{aligned}$$

Thus, the estimate is given by $\hat{\beta}_0 = 9$.

(ii) Calculate the expected Fisher information $\mathcal{I}(\beta_0)$ with respect to the parameter β_0 under the null model assuming $\beta_0 = 1$.

Solution: Following equation $(*)$ in (i) we get for the second derivative of the log-likelihood

$$\frac{\partial^2 l}{\partial \beta_0^2} = \frac{\partial}{\partial \beta_0}\left(\frac{3\bar{y}}{2\beta_0} - \frac{3}{2\sqrt{\beta_0}}\right) = -\frac{3\bar{y}}{2}\beta_0^{-2} + \frac{3}{4}\beta_0^{-3/2}.$$

Thus, since $\mathrm{E}(Y_i) = \mu_i = \sqrt{\beta_0}$ independent of i,

$$\mathcal{I}(\beta_0) = \frac{3}{2\beta_0^2}\mathrm{E}(\bar{Y}) - \frac{3}{4}\beta_0^{-3/2} \overset{\beta_0=1}{=} \frac{3}{4}.$$

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 at the beginning of every sub-task below, in which data are generated, i.e. at the beginning of task (a) and at the beginning of task (b).
Set $n = 150$. Let X_1 be a uniformly distributed random variable on $[0, 20]$. Let X_2 be a $\mathcal{N}_1(5, 4^2)$ distributed random variable.

(a) Sample a vector of n observations from X_1 denoted by $(x_{1,1}, ..., x_{1,n})$ and a vector of n observations from X_2 denoted by $(x_{2,1}, ..., x_{2,n})$, using the functions `runif()` and `rnorm()`, respectively. Compute the values of the sum $\sum_{i=1}^n x_{1,i}$ and the sum of squares $\sum_{i=1}^n x_{2,i}^2$. **(requested precision: whole numbers)**

(b) Consider a random variable $X = \dfrac{2 \cdot \varepsilon - 4}{3}$ that is exponentially distributed with expected value $\mathbb{E}(X) = \frac{1}{6}$. Based on this distribution (the distribution of X), and using the function `rexp()`, sample a vector of n values of ε, denoted by $(\varepsilon_1, ..., \varepsilon_n)$. What is the mean value of the sample $(\varepsilon_1, ..., \varepsilon_n)$? **(requested precision: 2 digits)**

(c) Compute the resulting values for the response variable $y_i = \mu_i + \varepsilon_i,\ i = 1, ..., n$ where

$$\mu_i = \beta_1 + \beta_2 \cdot x_{1,i} + \beta_3 \cdot x_{2,i} + \beta_4 \cdot x_{1,i} \cdot x_{2,i} \tag{3}$$

holds with $\beta_1 = 10$, $\beta_2 = -2, \beta_3 = 4$ and $\beta_4 = 0.2$. What is the proportion of values in $\boldsymbol{y} = (y_1, ..., y_n)$ satisfying $y_i < 2, i = 1, ..., n$? **(requested precision: 2 digits)**

(d) Use the least squares estimator $\hat{\boldsymbol{\beta}}$ to estimate $\boldsymbol{\beta} = (\beta_1, ..., \beta_4)$. What is the resulting estimate for the coefficient of the interaction term? **(requested precision: 4 digits)**

(e) Compute the resulting error $||\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}||_2$ of the coefficients using the true values of $\boldsymbol{\beta} = (\beta_1, ..., \beta_4)$ given in (c). **(requested precision: 2 digits)**

(f) Compute the residual sum of squares (RSS) for the model in (c) with the least squares estimates calculated in (d). **(requested precision: 2 digits)**

Solution

set.seed((A) 2021, (B) 123, (C) 456, (D) 789)
n=150
beta = c(10,-2,4,0.2)

(a) set.seed((A) 2021, (B) 123, (C) 456, (D) 789)
x1 = runif(n,0,20)
sum(x1)

(A) 1536.503 (Dynexite: 1537)
(B) 1512.508 (Dynexite: 1513)
(C) 1631.341 (Dynexite: 1631)
(D) 1446.366 (Dynexite: 1446)

x2 = rnorm(n,5,4)
sum(x2^2)

(A) 5832.93 (Dynexite: 5833)
(B) 5876.665(Dynexite: 5877)
(C) 5626.195 (Dynexite: 5626)
(D) 5860.729 (Dynexite: 5861)

(b) set.seed((A) 2021, (B) 123, (C) 456, (D) 789)
eps = (3*rexp(n,6)+4)*0.5
mean(eps)

(A) 2.262425 (Dynexite: 2.26)
(B) 2.249902 (Dynexite: 2.25)
(C) 2.216802 (Dynexite: 2.22)
(D)2.260102(Dynexite: 2.26)

(c) mu = beta[1]+beta[2]*x1+beta[3]*x2+ beta[4]*x1*x2
y = mu + eps
sum(y < 2)/length(y)

(A) 0.1866667 (Dynexite: 0.19)
(B) 0.18 (Dynexite: 0.18)
(C) 0.2333333(Dynexite: 0.23)
(D) 0.24 (Dynexite: 0.24)

(d) lm.fit = lm(y∼x1+x2+x1*x2)
beta.hat = lm.fit$coefficients
beta.hat[4]

(A) 0.1998057 (Dynexite: 0.1998)
(B) 0.2011993 (Dynexite: 0.2012)
(C) 0.2001496 (Dynexite: 0.2001)
(D)0.2004728 (Dynexite: 0.2005)

(e) sqrt(sum((beta-beta.hat)^2))

(A) 2.261367 (Dynexite: 2.26)
(B) 2.23251 (Dynexite: 2.23)
(C) 2.243323(Dynexite: 2.24)
(D) 2.352022 (Dynexite: 2.35)

(f) sum((lm.fit$fitted.values-y)^2)

(A)10.25201(Dynexite: 10.25)
(B) 8.387435 (Dynexite: 8.39)
(C) 6.3682 (Dynexite: 6.37)
(D) 10.27945 (Dynexite: 10.28)

Task 2

Clear your `R` workspace. Consider the following three-way contingency table presenting a sample of residents (aged over 30) of two countries, cross-classified by their gender (X_1), country of origin (X_2) and whether they have a college degree or not (X_3).

Gender	Country	College	
		Yes	No
Male	A	n_1	n_5
	B	n_2	n_6
Female	A	n_3	n_7
	B	n_4	n_8

Execute the following code to get the observed frequencies of the above contingency table and store it in the sequel as a data frame into your `R` workspace. Note that the observed cell frequencies $(n_1, ..., n_8)$ will be represented by the variable `freq` in the code below. These values are considered as realizations of the random cell frequencies $N_i, i = 1, ..., 8$, which are assumed to be independent and Poisson distributed, that is, $N_i \sim \mathcal{P}(m_i)$, $i = 1, \ldots, 8$.

```
set.seed((A) 2021, (B) 123, (C) 456, (D) 789)
m=5*c(12,11,9,8,14,13,11,7); freq=rpois(8,m)
row< −rep(c(1,2),each=2); lay< −rep(c(1,2),2); col< −c(rep(1,4),rep(2,4))
row.lb< −c(”Male”,”Female”); lay.lb< −c(”A”,”B”); col.lb< −c(”yes”,”no”)
gender< −factor(row,labels=row.lb) ; country< −factor(lay,labels=lay.lb)
college< −factor(col,labels=col.lb)
educ< −data.frame(gender,country,college,freq)
```

(a) The sample odds of having a college degree is .... times higher for country B than for A, independently of gender. **(requested precision: 2 digits)**

(b) Fit the saturated log-linear model on the data in `educ`. What is the AIC value of this model? **(requested precision: 2 digits)**

(c) Starting with the saturated model discussed in (b), use a backward selection algorithm to select the best nested hierarchical log-linear model based on AIC. What is the value of the null deviance for this model? **(requested precision: 2 digits)**

(d) Fit the hierarchical log-linear model (X_1, X_2X_3). What is the estimate of the gender main effect for the category "female"? **(requested precision: 4 digits)**

(e) Compute the Pearsonian residuals and deviance residuals for the model (X_1, X_2X_3) in (d). Compute the proportion of values where the deviance residuals are less than the corresponding Pearsonian residuals. **(requested precision: 1 digit)**

Solution

set.seed((A) 2021, (B) 123, (C) 456, (D) 789)
m=5*c(12,11,9,8,14,13,11,7); freq=rpois(8,m)
row< −rep(c(1,2),each=2); lay< −rep(c(1,2),2); col< −c(rep(1,4),rep(2,4))
row.lb< −c("Male","Female"); lay.lb< −c("A","B"); col.lb< −c("yes","no")
gender< −factor(row,labels=row.lb) ; country< −factor(lay,labels=lay.lb)
college< −factor(col,labels=col.lb)
educ< −data.frame(gender,country,college,freq)

(a) tab=xtabs(freq country+college, data=educ)
OR = (tab[1,1]*tab[2,2])/(tab[2,1]*tab[1,2])
1/OR

(A)1.58408 (Dynexite:1.58)
(B) 1.641181 (Dynexite: 1.64)
(C) 1.268969 (Dynexite: 1.27)
(D) 1.055973 (Dynexite: 1.06)

(b) sat.model< −glm(freq ∼ gender*country*college,poisson, data=educ)
sat.model$aic **(A)62.16796(Dynexite: 62.17)**
(B)61.70389(Dynexite: 61.70)
(C)62.15456(Dynexite: 62.15)
(D)61.66911 (Dynexite: 61.67)

(c) glm.select= step(sat.model, direction="backward")
glm.select$null.deviance **(A)25.18492 (Dynexite:25.18)**
(B)54.23455 (Dynexite: 54.23)
(C) 22.43731 (Dynexite: 22.44)
(D)26.91342 (Dynexite: 26.91)

(d) glm.x1.x2x3=glm(freq ∼ gender + country + college + country:college,poisson, data=educ)
glm.x1.x2x3$coefficients["genderFemale"] **(A) -0.3266842 (Dynexite: -0.3267)**
(B) -0.6425949(Dynexite: -0.6426)
(C) -0.3086472 (Dynexite: -0.3086)
(D) -0.212922 (Dynexite: -0.2129)

(e) res.p = residuals(glm.x1.x2x3, type="pearson")
res.d = residuals(glm.x1.x2x3, type="deviance")
sum(res.p>res.d)/length(res.p) **(A) 1 (Dynexite: 1)**
(B) 1 (Dynexite: 1)
(C) 1 (Dynexite: 1)
(D) 1 (Dynexite: 1)

Task 3

Clear your R workspace.
Consider independent binomial responses Y_i, $i = 1, \ldots, n$, with $Y_i \in \{0, 1\}$, where 1 denotes the event of success. Model these random responses by a GLM with canonical link and linear predictor

$$\eta_i = 3 + 2.5 \cdot x_{1,i} + 0.6 \cdot x_{2,i} + 0.5 \cdot x_{2,i} \cdot x_{1,i} \tag{4}$$

Consider a sample size of $n = 100$ and sample a vector of n observations from $X_1 \sim \mathcal{N}_1(-1, 1)$ denoted by $(x_{1,1}, ..., x_{1,n})$ and a vector of n observations from $X_2 \sim \mathcal{N}_1(2, 4^2)$ denoted by $(x_{2,1}, ..., x_{2,n})$ using the following R code

```
n=100
set.seed((A) 2021, (B) 123, (C) 456 (D) 789)
x1=rnorm(n,-1,1)
x2=rnorm(n,2,4)
lin.pred=3+2.5*x1+0.6*x2+0.5*x1*x2
mu = exp(lin.pred)/(1+exp(lin.pred))
y = rbinom(n,1,mu)
```

(a) What is the proportion of successes in $y = (y_1, ..., y_n)$? **(requested precision: 2 digits)**

(b) Based on the sampled data, fit a logistic regression model that predicts the response variable Y using X_1, X_2 and the interaction of these two variables as explanatory variables. Let β_3 denote the true coefficient of the interaction term. What is the standard error of the estimate $\hat{\beta}_3$? **(requested precision: 2 digits)**

(c) Based on the model fitted in (b), compute a 90 % profile likelihood confidence interval for $\hat{\beta}_3$. **(requested precision: 2 digits)**

(d) What is the percentage of correctly classified observations of the model fitted in (b) using $P(Y = 1) \geq 0.5$ as threshold? **(requested precision: 2 digits)**

(e) Based on the model fitted in (b), compute the p-value for testing whether there is evidence against the assumption that $\beta_1 = 0$, where β_1 denotes the parameter corresponding to X_1. If β_1 is statistically significant at significance level $\alpha = 0.05$ then type in "1", else type in "0" (without quotation marks)

(f) Fit a logistic regression model that predicts the response Y based on X_1 and X_2 ignoring the interaction term of X_1 and X_2. What is the mean value of the fitted values of this model? **(requested precision: 2 digits)**

(g) Compute the area under the curve (AUC) for the model fitted in (b) and the model fitted in (f). What are the respective values for AUC? **(requested precision: 4 digits)**

Solution

(a) sum(y==1)/length(y)

(A) 0.48 (Dynexite: 0.48)
(B) 0.6 (Dynexite: 0.60)
(C)0.61 (Dynexite: 0.61)
(D) 0.59 (Dynexite: 0.59)

(b) data2 = data.frame(y=y,x1=x1,x2=x2)
model.2=glm(y∼x1*x2,data=data2,family="binomial")
summary(model.2) # standard error for β_3 is

(A)0.1327 (Dynexite: 0.13)
(B)0.1500 (Dynexite: 0.15)
(C) 0.1655 (Dynexite: 0.17)
(D) 0.1313 (Dynexite: 0.13)

(c) CI=confint(model.1, level=0.9)
CI[4,] #CI for $\hat{\beta}_3$ given by

(A) [0.2511776 0.6920013] (Dynexite: [0.25, 0.69])
(B) [0.2002842, 0.6981671] (Dynexite: [0.20,0.70])
(C)[0.2857130, 0.8353383](Dynexite: [0.29,0.84])
(D) [0.2456832, 0.6840810] (Dynexite: [0.25,0.68])

(d) y.pred=ifelse(model.2$fitted.values > 0.5, 1, 0)
tab1=table(data2$y,y.pred)
sum(diag(tab1))/sum(tab1)

(A) 0.8 (Dynexite: 0.80)
(B) 0.85 (Dynexite: 0.85)
(C) 0.75 (Dynexite: 0.75)
(D) 0.85 (Dynexite: 0.85)

(e) summary(model.2) #p- value is 3.97e-05 < 0.05 so reject H_0 so solution is "1" for (A), same holds for (B) with p-value 5.28e-06 and (C) with p-value 6.33e-05 and (D) with p-value 5.67e-05

(f) model.3=glm(y∼x1+x2,data=data2,family="binomial") mean(model.3$fitted.values)

(A) 0.48 (Dynexite: 0.48)
(B) 0.6 (Dynexite: 0.60)
(C) 0.61 (Dynexite: 0.61)
(D) 0.59 (Dynexite: 0.59)

(g) roc.curve1=roc(y ∼ fitted(model.2), data=data2)
roc.curve2=roc(y ∼ fitted(model.3), data=data2)
auc(roc.curve1)

(A)0.9091 (Dynexite: 0.9091)
(B) 0.8933 (Dynexite: 0.8933)
(C) 0.8743 (Dynexite: 0.8743)
(D) 0.8995 (Dynexite: 0.8995)

auc(roc.curve2)

(A) 0.8658 (Dynexite: 0.8658)
(B) 0.8633 (Dynexite: 0.8633)
(C) 0.8331 (Dynexite: 0.8331)
(D) 0.8603 (Dynexite: 0.8603)