E-test 4 15.07.22

Freitag, 15. Juli 2022 13:51

Dokument

Aufgabe 1

8 Punkte

Consider a linear model $Y = B\beta + \varepsilon$ according to Definition I.4.2 with $n \geq p$ and design matrix $B \in \mathbb{R}^{n\times p}$ having orthonormal columns, i.e. $B'B = I_p$. Further, let $y \in \mathbb{R}^n$ be a realization of Y.

Then, the usual objective function for that linear model $\psi : \mathbb{R}^p \longrightarrow [0, \infty)$ defined by

$$\psi(\beta) := \|y - B\beta\|_2^2, \ \beta \in \mathbb{R}^p,$$

with $\|\cdot\|_2$ denoting the corresponding Euclidean norm, leads to the (uniquely determined) least-squares-estimate $\hat{\beta}^{LS}$ of β.

As another objective function for that linear model, consider $\tilde{\psi} : \mathbb{R}^p \longrightarrow [0, \infty)$ defined by

$$\tilde{\psi}(\beta) := \|y - B\beta\|_2^2 + \lambda \|\beta\|_2^2, \ \beta \in \mathbb{R}^p,$$

with some $\lambda > 0$. Let the (also uniquely determined) solution of the corresponding minimization problem

$$\tilde{\psi}(\beta) \longrightarrow \min_{\beta \in \mathbb{R}^p},$$

known as ridge-estimate of β, be denoted by $\hat{\beta}^{\text{Ridge}}$. Here, we use the same notations for the estimates as for the corresponding estimators, i.e. the corresponding random variables.

Hint: The objective function $\tilde{\psi}$ can be represented as follows:

$$\tilde{\psi}(\beta) = \|\tilde{y} - \tilde{B}\beta\|_2^2, \ \beta \in \mathbb{R}^p,$$

1 2 3 4

with $\tilde{B} := \begin{pmatrix} B \\ \sqrt{\lambda} I_p \end{pmatrix} \in \mathbb{R}^{(n+p)\times p}$ and $\tilde{y} := \begin{pmatrix} y \\ 0_p \end{pmatrix} \in \mathbb{R}^{n+p}$.

(a) Under the given assumptions, show that the following equation holds:

$$\hat{\beta}^{\text{Ridge}} = c(\lambda)\,\hat{\beta}^{LS}$$

with some factor $c(\lambda) \in \mathbb{R}$, which (only) depends on λ.

5 Punkte

Calculate the factor $c(\lambda)$ for $\lambda = \frac{1}{4}$.

5 Punkte

Give the value of $c(\lambda)$ with a precision of **one decimal place.** $4/5$

(b) In this part, assume $p = 2$ and that the variance parameter of the considered linear model is given by $\sigma^2 = 9$. Further, assume that the two estimators fulfil the following equation:

$$\hat{\beta}^{\text{Ridge}} = c(\lambda)\,\hat{\beta}^{LS}$$

with $c(\lambda) = \frac{2}{3}$ for some suitably chosen $\lambda > 0$.

3 Punkte

Calculate the entries of the covariance matrix $\text{Cov}(\hat{\beta}^{\text{Ridge}}) = \begin{pmatrix} c_{11} & c_{12} \\ c_{12} & c_{22} \end{pmatrix}$.

1 Punkt

Give the value of c_{11}. 6

1 Punkt

Give the value of c_{12}. 0

1 2 3 4

3 Punkte

Calculate the entries of the covariance matrix $\text{Cov}(\hat{\beta}^{\text{Ridge}}) = \begin{pmatrix} c_{11} & c_{12} \\ c_{12} & c_{22} \end{pmatrix}$.

1 Punkt

Give the value of c_{11}. 6

1 Punkt

Give the value of c_{12}. 0

1 Punkt

Give the value of c_{22}. 6

1 2 3 4

1

a) $\psi(\beta) := \|y - B\beta\|_2^2 \ , \beta \in \mathbb{R}^p,$

$\tilde{\psi}(\beta) = \|y - B\beta\|_2^2 + \lambda\|\beta\|_2^2 \ , \beta \in \mathbb{R}^p, \lambda \geq 0$

$= \|\tilde{y} - \tilde{B}\beta\|_2^2 \ , \beta \in \mathbb{R}^p$

$\arg\max(\psi(\beta)) = \hat{\beta}^{LS} = (\overbrace{B'B}^{I_p})^{-1} B'y = B'y \quad (*1)$

$\Leftrightarrow \arg\max(\tilde{\psi}(\beta)) = (\tilde{B}'\tilde{B})^{-1}\tilde{B}'\tilde{y}$

- $\tilde{B}'\tilde{B} = (B' \ \sqrt{\lambda} I_p)\begin{pmatrix} B \\ \sqrt{\lambda} I_p \end{pmatrix} = B'B + \lambda I_p = I_p + \lambda I_p = \begin{pmatrix} 1+\lambda & & 0 \\ & \ddots & \\ 0 & & 1+\lambda \end{pmatrix}$
- $(\tilde{B}'\tilde{B})^{-1} = \frac{1}{1+\lambda}\begin{pmatrix} 1 & & 0 \\ & \ddots & \\ 0 & & 1 \end{pmatrix} = \frac{1}{1+\lambda} I_p \quad (*2)$
- $(\tilde{B}'\tilde{B})^{-1}\tilde{B}' = \frac{1}{1+\lambda} I_p \underbrace{(B' \ \sqrt{\lambda} I_p)}_{} = \frac{1}{1+\lambda}(B' \ \sqrt{\lambda} I_p)$

Side note: Matrix of matrices
⇝ same as multiplying a coeff to a matrix
$(I_p B' \ \sqrt{\lambda} I_p)$

- $(\tilde{B}'\tilde{B})^{-1}\tilde{B}'\tilde{y} = \frac{1}{1+\lambda}(B' \ \sqrt{\lambda} I_p)\begin{pmatrix} y \\ 0_p \end{pmatrix} = \frac{1}{1+\lambda} B'y$

with $(*1)$: $\hat{\beta}^{LS} = B'y$

$\arg\max(\tilde{\psi}(\beta)) = \frac{1}{1+\lambda}\hat{\beta}^{LS}$

$\rightarrow c(\lambda) = \frac{1}{1+\lambda}$, for $\lambda = \frac{1}{4} \leadsto \frac{4}{5} = c(\lambda)$

b)

$c(\lambda) = \frac{1}{1+\lambda} = \frac{2}{3}$, for $\lambda = \frac{1}{2}$

$\text{cov}(\hat{\beta}^{\text{Ridge}}) = \sigma^2 (\tilde{B}'\tilde{B})^{-1} \overset{(*2)}{=} \sigma^2 c(\lambda) I_p$

$= 9 \cdot \frac{2}{3} I_2 = \begin{pmatrix} 6 & 0 \\ 0 & 6 \end{pmatrix}$

I.4.8 Definition

Let $Y = B\beta + \varepsilon$ be a LM and y be a realisation of Y. Th

- A solution $\beta^* = \beta^*(y)$ of the minimization problem

$$\psi(\beta) = \|y - B\beta\|^2$$

is called **least-squares-estimate**.
- $\beta^*(Y)$ is called **least-squares-estimator (LSE)** for
- For $A \in \mathbb{R}^{k\times p}$, a LSE of $A\beta$ is defined as $A\beta^*(Y)$.

I.4.9 Theorem (LSE in LM)

Given a LM with design matrix $B \in \mathbb{R}^{n\times p}$, a LSE $\hat{\beta} = \hat{\beta}$

$$B'y = B'B\hat{\beta}.$$

If B^+ and $(B'B)^+$ denote the Moore-Penrose inverse of by

$$\hat{\beta}^+ = B^+y = (B'B)^+B'y \quad \text{with } \psi(\hat{\beta}^+)$$

The set of all LSEs is given by $\{\hat{\beta} = \hat{\beta}^+ + (I_p - B^+B)z \mid$

I.4.12 Corollary (LSE in LM)

Given a LM with design matrix B satisfying de

$$\hat{\beta} =$$

I.4.13 Remark

In case the LSE is unique, we write for short $\hat{\beta}$ inverse matrix.

Properties of the LSE $\hat{\beta}^+$

I.4.25 Theorem

Given a LM with design matrix B, the LSE $\hat{\beta}^+$ has the foll

1. $E\hat{\beta}^+ = B^+B\beta$, $\text{Cov}(\hat{\beta}^+) = \sigma^2(B'B)^+$
2. Under a NoLM, $\hat{\beta}^+ \sim N_p(B^+B\beta, \sigma^2(B'B)^+)$

If the design matrix B has $\text{rank}(B) = p$, then the (unique) that is,

$$E\hat{\beta}^+ = \beta \text{ and the Moore-Penrose inverse } (B'B)^+ \text{ e}$$

In particular, we get under a NoLM, $\hat{\beta} \sim N_p(\beta, \sigma^2(B'B)^{-1})$

Aufgabe 2

12 Punkte

Let $X_1, \ldots, X_n$ be stochastically independent and identically distributed continuous random variables, each having a density function $f(\bullet; \alpha, \beta) : \mathbb{R} \longrightarrow [0, \infty)$ defined by

$$(*) \quad f(x; \alpha, \beta) := \begin{cases} \frac{1}{\alpha}\exp\left(-\frac{x-\beta}{\alpha}\right), & x \geq \beta, \\ 0, & x < \beta, \end{cases}$$

with parameters $\alpha > 0$ and $\beta > 0$.

In the following tasks, round your derived numerical solutions to three decimal places, if necessary.

(a) For a given realization $x = (x_1, \ldots, x_n)' \in (0, \infty)^n$ of $X = (X_1, \ldots, X_n)'$, the corresponding (full) likelihood function $L(\bullet \mid x) : (0, \infty)^2 \longrightarrow [0, \infty)$ is given by

$$(**) \quad L(\alpha, \beta \mid x) = \begin{cases} \alpha^{-n}\exp\left(-\frac{n}{\alpha}(\bar{x} - \beta)\right), & \alpha \in (0, \infty), \beta \in (0, x_{(1)}], \\ 0, & \alpha \in (0, \infty), \beta \in (x_{(1)}, \infty), \end{cases}$$

1 2 3 4

where $x_{(1)} := \min\{x_1, \ldots, x_n\}$ denotes the minimum and $\bar{x} := \frac{1}{n}\sum_{i=1}^n x_i$ denotes the mean of the given realization $x = (x_1, \ldots, x_n)'$.

In this part, consider the profile likelihood approach for estimating the parameter α. Let $\hat{\beta}(\alpha)$ denote the maximum likelihood estimate of the nuisance parameter β for fixed $\alpha \in (0, \infty)$.

Suppose that the following sample of size $n = 4$ has been observed:

$$x_1 = 4, \ x_2 = 8, \ x_3 = 2, \ x_4 = 6.$$

2)

Note:
The maximum Likelyhood Estimator of the full likelyhood function is β for which the full likelyhood function becomes maximum.

$\leadsto \max(\beta) = X_{(1)} = \min(X_1, \ldots, X_n)$

$= 2 = \beta(\alpha)$

$\alpha^{\text{Profile}} := \text{MLE of } \alpha$

$\rightarrow \max(L(\alpha, 2))$

$\leadsto L(\alpha, 2) = \alpha^{-4}\exp\{-\frac{4}{\alpha}(5-2)\}$

$= \alpha^{-4}\exp\{-\frac{20}{\alpha} + \frac{8}{\alpha}\} = \alpha^{-4}\exp\{-\frac{12}{\alpha}\}$

$\max(L(\alpha, 2)) \rightarrow \nabla L(\alpha, 2) \overset{!}{=} 0$

$\rightarrow \frac{d}{d\alpha}\alpha^{-4}e^{-\frac{12}{\alpha}} = -4\alpha^{-5}e^{-12/\alpha} + \alpha^{-4}\,12\,\alpha^{-2}e^{-12/\alpha}$

$= (-4\alpha^{-5} + 12\alpha^{-6})e^{-12/\alpha} \overset{!}{=} 0$

note:
loglikelyhood will have the same result, since logarithm is a monotonic function $\leadsto \max(\ln(L(\alpha))) = \max(L(\alpha))$

BLUEs and MLEs

I.4.29 Definition

In the situation of Theorem I.4.28, the LSE $\hat{\beta}$ is called **best linear unbiased estimator (BLUE)** of β. For $c \in \mathbb{R}^p$, $c'\hat{\beta}$ is called BLUE of $c'\beta$.

I.4.30 Remark

Notice that the derivation of LSEs and BLUEs as well as their means and variances do not depend on the particular distributional assumption. However, their distributions depend on the distribution of the error term ε.

I.4.31 Theorem

Given a NoLM with a regular matrix $B'B$ and unknown variance parameter $\sigma^2 > 0$, $\hat{\beta}$ is also the **Maximum-Likelihood-Estimator (MLE)** of β.

With $P = I_n - B(B'B)^{-1}B'$, the MLE of σ^2 is given by $\widehat{\sigma^2} = \frac{1}{n}Y'PY$.

en:

$\longrightarrow \min_{\beta \in \mathbb{R}^p}$ (1.1)

β.

$y)$ has to satisfy the normal equations

B and $B'B$, respectively, a solution is given

$= y'(I_n - B(B'B)^+B')y$.

$z \in \mathbb{R}^p\}$.

et$(B'B) > 0$, the unique LSE is given by

$(B'B)^{-1}B'Y$.

instead of $\widehat{\beta}^+$ for the LSE based on the Moore-Penrose

owing properties:

LSE $\widehat{\beta}^+$ is an unbiased estimator for β,

quals the inverse matrix $(B'B)^{-1}$.

.

... the profile likelihood approach for estimating the parameter α. Let $\hat{\beta}(\alpha)$ denote the maximum likelihood estimate of the nuisance parameter β for fixed $\alpha \in (0, \infty)$.

Suppose that the following sample of size $n = 4$ has been observed.

$$x_1 = 4\,,\; x_2 = 8\,,\; x_3 = 2\,,\; x_4 = 6\,.$$

(i) Based on the observed sample, calculate the corresponding estimate $\hat{\beta}(\alpha)$ of β for $\alpha = 5$ — 4 Punkte

Give the value of $\hat{\beta}(\alpha)$. — 4 Punkte

2

(ii) Based on the observed sample, calculate the corresponding profile maximum likelihood estimate $\hat{\alpha}^{\text{profile}}$ of α. — 4 Punkte

1 2 3 4

01:28

(ii) Based on the observed sample, calculate the corresponding profile maximum likelihood estimate $\hat{\alpha}^{\text{profile}}$ of α. — 4 Punkte

Give the value of $\hat{\alpha}^{\text{profile}}$. — 4 Punkte

3

(b) In this part, let X be a continuous random variable with corresponding density function $f(\bullet;\alpha,\beta)$ given by (*) for some $\alpha, \beta > 0$. Further, let Y be a discrete random variable taking values in $\mathbb{N}_0$ with the conditional distribution of Y under the condition $X = x$, $x \in (0, \infty)$, being a Poisson distribution with parameter x, i.e.

$$P^{Y|X=x} = \mathcal{P}(x)\,,\; x \in (0, \infty)\,,$$

with P denoting the underlying probability distribution.

Then, by conditioning on X, the expectation $E(Y)$ of Y can be determined as follows:

$$E(Y) = E(E(Y|X)) = \int_0^\infty E(Y \mid X = x)\, f(x;\alpha,\beta)\, dx$$

Here, $E(Y \mid X = x)$ denotes the expectation with respect to the distribution $P^{Y|X=x}$ for $x \in (0, \infty)$.

Here, $E(Y \mid X = x)$ denotes the expectation with respect to the distribution $P^{Y|X=x}$ for $x \in (0, \infty)$.

For $\alpha = 5$ and $\beta = 1$, calculate the expectation $E(Y)$. — 4 Punkte

Give the value of $E(Y)$. — 4 Punkte

6

Nächste Aufgabe

← Vorherige Aufgabe

Alle Antworten wurden gespeichert!

1 2 3 4

R Task 1 — 01:28:38

Please provide numbers in the requested precision within each question. The use of different precision is evaluated as wrong.

Consider the following contingency table describing the attitude of Germans to genetically modified (GM) foods according to their income (low/middle/high). Let X denote the attitude (for GM foods/against GM foods) and Y the income (low/middle/high). This contingency table is a realization of a random contingency tables with independent cell entries $N_{ij} \sim \mathcal{P}(\mu_{ij})$ for $i = 1, 2$ and $j = 1, 2, 3$.

	High income	Middle income	Low income	Total
For GM foods	320	280	258	858
Against GM foods	150	166	112	428
Total	470	446	370	1286

Read the contingency table above in R in a `data.frame` format, appropriate for fitting log-linear models with the `glm` function.

(a) Provide the marginal probabilities π_{+1} and π_{2+}. **(requested precision: 2 digits)** — 1 Punkt

1 2 3 4

01:28:33

(a) Provide the marginal probabilities π_{+1} and π_{2+}. **(requested precision: 2 digits)** — 1 Punkt

π_{+1} — 0.5 Punkte

Zahl

π_{2+} — 0.5 Punkte

Zahl

(b) Fit the log-linear model of independence that estimates the expected cell frequencies μ_{ij}, $i = 1, 2, j = 1, 2, 3$, under the hypothesis of independence of the classification variables X and Y. Provide the corresponding Pearsonian residual for the case of responders having medium income and being against GM foods. Provide the value of the Pearsons χ^2 goodness of fit statistic. **(requested precision: 2 digits)** — 2 Punkte

Pearsonian residual for the case of responders having medium income and being against GM foods — 1 Punkt

1 2 3 4

$$\to \frac{d}{d\alpha}\,\alpha^{-4} e^{-\frac{12}{\alpha}} = -4\alpha^{-5} e^{-12/\alpha} + \alpha^{-4}\,12\,\alpha^{-2} e^{-12/\alpha}$$

$$= \left(-4\alpha^{-5} + 12\alpha^{-6}\right) e^{-12/\alpha} \overset{!}{=} 0$$

same result, since logarithm is a monotonic function $\rightsquigarrow \max(\ln(L(\alpha)) = \max(L(\alpha))$

$$\rightsquigarrow -\alpha^{-5} + 3\alpha^{-6} = 0$$

$$\alpha^{-5}\left(3\alpha^{-1} - 1\right) = 0$$

$$\rightsquigarrow \alpha = 0 \text{ or } 3\alpha^{-1} - 1 = 0$$

$$3\alpha^{-1} = 1$$

$$\to \alpha = 3$$

ii)

$$f(x;\alpha,\beta) := \begin{cases} 1/\alpha \exp\left(-\frac{x-\beta}{\alpha}\right), & x \geq \beta \\ 0, & x < \beta \end{cases}$$

$$E(Y) = \int_1^\infty \underbrace{E(Y \mid X = x)}_{= x} f(x;\alpha,\beta)\, dx$$

$$\alpha = 5,\; \beta = 1$$

$$\rightsquigarrow E(Y) = \int_1^\infty \frac{x}{5} \exp\left(-\frac{x-1}{5}\right) dx$$

$$= \frac{1}{5}\left(\int x \cdot \exp\left(\frac{1}{5} - \frac{x}{5}\right) dx\right)$$

$$= \frac{1}{5}\left(\left[\underbrace{-5x\exp\left(\frac{1}{5} - \frac{x}{5}\right)}_{\to 0}\right]_1^\infty + \int_{-\infty}^\infty 5 \exp\left(-\frac{x-1}{5}\right) dx\right)$$

$$= \frac{1}{5}\left(5 + \left[-25\exp\left(\frac{1}{5} - \frac{x}{5}\right)\right]_1^\infty\right)$$

$$= 30/5 = 6 \;✓$$

Pearsonian residual for the case of responders having medium income and being against GM foods

1 Punkt

Zahl

Pearsons χ^2 statstic

1 Punkt

Zahl

(c) For the model fitted in (b), provide the value of the likelihood ratio statistic G^2. Further, provide the corresponding asymptotic p-value. Based on G^2 and at significance level $\alpha = 0.05$, does the independence model adequately describe this dataset? **(requested precision: 2 digits)**

2 Punkte

1 Punkt

value of likelihood ratio statistic G^2

Zahl

0.5 Punkte

asymptotic p-value

0.5 Punkte

Zahl

Based on G^2 and at significance level $\alpha = 0.05$, does the independence model adequately describe this dataset?

0.5 Punkte

Bitte auswählen

(d) Fit the saturated log-linear model. Give the resulting estimate for λ. **(requested precision: 2 digits)**

2 Punkte

2 Punkte

estimate for λ

Zahl

(e) Starting with the model in (d), select an appropriate nested model based on AIC using backward selection. Give the resulting value for AIC of this selected model. **(requested precision: 2 digits)**

1 Punkt

(e) Starting with the model in (d), select an appropriate nested model based on AIC using backward selection. Give the resulting value for AIC of this selected model. **(requested precision: 2 digits)**

1 Punkt

AIC

1 Punkt

Zahl

(f) For the model selected in (e), give the 95% asymptotic (Wald) confidence interval for the intercept **(requested precision: 3 digits)**

2 Punkte

1 Punkt

lower bound

Zahl

1 Punkt

upper bound

R Task 2

Please provide numbers in the requested precision within each question. The use of different precision is evaluated as wrong.

The file `Houses.txt`, which can be downloaded by clicking on the button "Houses", contains data with information of 100 home sales in Gainesville, Florida. Find the details of the dataset below.

attribute	description & properties
case	index
price	whether the selling price of the house (in thousands of dollars) is ≥ 200 or not ($1 =$ yes, $0 =$ no)
size	size of the house (square feet)
new	whether the house is new ($1 =$ yes, $0 =$ no)
taxes	tax bill (dollars)
bedrooms	number of bedrooms
baths	number of bathrooms

Houses

(a) Load the data file into your workspace and transform the attribute `new` into a factor variable. Obtain the number of houses in the dataset for which the tax bills are greater than 3000 dollars and the number of bedrooms are at least equal to 2.

1 Punkt

1 Punkt

value

Zahl

(b) Fit a GLM using the canonical link from predicting the value of `price` using the variables `size, new, taxes, bedrooms, baths` as explanatory variables. Provide the resulting value for the estimated coefficient of the attribute `bedrooms`. **(requested precision: 3 digits)**

1 Punkt

1 Punkt

value for estimated coefficient of attribute `bedrooms`

Zahl

(c) Let β_4 denote the coefficient of the attribute `taxes` in the model fitted in (b) above. Test at significance level $\alpha = 0.01$ the hypotheses $H_0 : \beta_4 = 0$ versus $H_1 : \beta_4 \neq 0$. Base your test on

(c) Let β_4 denote the coefficient of the attribute `taxes` in the model fitted in (b) above. Test at significance level $\alpha = 0.01$ the hypotheses $H_0 : \beta_4 = 0$ versus $H_1 : \beta_4 \neq 0$. Base your test on the Wald's test statistic and provide the p-values and your decision. **(requested precision: 3 digits)**

1 Punkt

0.5 Punkte

p-value

Zahl

0.5 Punkte

decision

Bitte auswählen

(d) Based on the model fitted in (b), predict the probability of having a price ≥ 200 thousands dollars for a house that is new, has a size of 1000 square feets, a tax bill of 1200 dollars, 2 bedrooms and 2 baths. **(requested precision: 3 digits)**

3 Punkte

(d) Based on the model fitted in (b), predict the probability of having a price ≥ 200 thousands dollars for a house that is new, has a size of 1000 square feets, a tax bill of 1200 dollars, 2 bedrooms and 2 baths. **(requested precision: 3 digits)**

3 Punkte

3 Punkte

predicted probability

Zahl

(e) For the model fitted in (b), compute 95% profile likelihood confidence interval for β_6 (coefficient of `baths`). **(requested precision: 3 digits)***

2 Punkte

1 Punkt

lower bound

Zahl

1 Punkt

upper bound

Zahl

01:27:42

Zahl

(d) Obtain the area under the curve (AUC) for the ROC curve of the model in (b). **(requested precision: 2 digits)**

2 Punkte

2 Punkte

AUC

Zahl

← Vorherige Aufgabe

Alle Antworten wurden gespeichert!