Nächste Aufgabe

Gesamtpunktzahl **13 von 13 Punkte**

For $k \in \mathbb{N}$ and $\alpha \in (1, \infty)$, let Y be a discrete random variable with values in $M = \frac{1}{k}\,\mathbb{N}_0 = \left\{ \frac{i}{k} \,\middle|\, i \in \mathbb{N}_0 \right\}$ and corresponding probability mass function $p(\bullet; k, \alpha) : M \to [0, 1]$, given by

$$(*) \quad p(y; k, \alpha) := \binom{k\,y + k - 1}{k - 1} \frac{(\alpha - 1)^{k\,y}}{\alpha^{k\,(y+1)}} \;,\; y \in M \,.$$

It can be shown that the given probability mass function is a member of an Exponential Dispersion Family, i.e. it can be expressed in the following form:

$$(**) \quad p(y; k, \alpha) = \exp\left(\frac{y\,\theta - b(\theta)}{a(\phi)} + c(y, \phi) \right) \;,\; y \in M \,,$$

with $\phi := k$, $a(\phi) = a(k) := \frac{1}{k}$, $c(y, \phi) = c(y, k) := \ln\left(\binom{k\,y+k-1}{k-1} \right)$ for $y \in M$, an appropriately chosen natural parameter θ (depending on α) and an appropriately chosen function b of θ.

(a) For the following tasks, you need to derive a representation of the form $(**)$ for the probability mass function given by $(*)$.

(i) For $\alpha = 4$, calculate the value of the corresponding natural parameter θ. **3 von 3 Punkten**

3 von 3 Punkten

Give the value of θ (**rounded to three decimal places**).

-0.288 ✓

(ii) For $\alpha = 4$ and $k = 2$, calculate the expectation $\mathrm{E}(Y)$. **2 von 2 Punkten**

(ii) For $\alpha = 4$ and $k = 2$, calculate the expectation $\mathrm{E}(Y)$. **2 von 2 Punkten**

2 von 2 Punkten

Give the value of $\mathrm{E}(Y)$.

3 ✓

(ii) For $\alpha = 4$ and $k = 2$, calculate the variance $\mathrm{Var}(Y)$. **2 von 2 Punkten**

2 von 2 Punkten

Give the value of $\mathrm{Var}(Y)$.

6 ✓

(b) Consider three stochastically independent discrete random variables Y_1, Y_2, Y_3, each having a probability mass function given by $(*)$ for $k = 1$ and some parameters $\alpha_1, \alpha_2, \alpha_3 \in (1, \infty)$, respectively, which are chosen such that it holds:

$$\mathrm{Var}(Y_1) = 6 \quad , \quad \mathrm{Var}(Y_2) = 12 \quad , \quad \mathrm{Var}(Y_3) = 2\,.$$

Further, let $\boldsymbol{Y} = (Y_1, Y_2, Y_3)'$ fulfill a GLM given by $g(\mathrm{E}(\boldsymbol{Y})) = g(\boldsymbol{\mu}) = X\,\boldsymbol{\beta}$ with **canonical** link function g, parameter vector $\boldsymbol{\beta} = (\beta_1, \beta_2)' \in \mathbb{R}^2$ and design matrix

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

In this framework and for the variances of Y_1, Y_2, Y_3 given above, determine the expected Fisher information matrix

$$\mathcal{I}_F(\boldsymbol{\beta}) = \begin{pmatrix} i_{11} & i_{12} \\ i_{12} & i_{22} \end{pmatrix}$$

for the corresponding parameter vector $\boldsymbol{\beta}$.

Hint: Part (b) can be solved independently of part (a), without deriving a representation of the form $(**)$ for the given probability mass function.

II.2.23 Definition (observed information matrix)

Let $\ell(\boldsymbol{\beta})$ be the log-likelihood of a GLM with associated parameter vector $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$. The observed information matrix is the following $p \times p$ matrix

$$\mathcal{I}_F^{obs} = \left(-\frac{\partial^2 \ell}{\partial \beta_k \partial \beta_r} \right) = -\mathcal{H} \,,$$

this parameter vector corresponds to each single parameter

for example, gamma(a, b), there can be a vector that corresponds to alpha

where $\mathcal{H}$ is known as the *Hessian matrix*. It holds:

$$\mathcal{I}_F = E\left(\mathcal{I}_F^{obs}\right) = E\left(-\mathcal{H}\right).$$

II.2.24 Information matrix for GLMs with canonical link

For a GLM with canonical link function, since $\eta_i = \vartheta_i$, it follows that (s. Proof of II.2.19) $\frac{\partial \mu_i}{\partial \eta_i} = \frac{\partial b'(\vartheta_i)}{\partial \vartheta_i} = b''(\vartheta_i)$, $i = 1, \ldots, n$, leading to

$$\mathcal{H} = -\mathbf{X}'\mathbf{W}_c\mathbf{X},$$

where $\mathbf{W}_c = \mathrm{diag}(w_1, \ldots, w_n)$ with $w_i = \frac{b''(\vartheta_i)}{a(\phi; i)}$, **independent** of y. Hence

$$\mathcal{I}_F = E\left(-\mathcal{H}\right) = -\mathcal{H} = \mathcal{I}_F^{obs}$$

Further, let $\boldsymbol{Y} = (Y_1, Y_2, Y_3)'$ fulfill a GLM given by $g(\mathrm{E}(\boldsymbol{Y})) = g(\boldsymbol{\mu}) = X\,\boldsymbol{\beta}$ with **canonical** link function g, parameter vector $\boldsymbol{\beta} = (\beta_1, \beta_2)' \in \mathbb{R}^2$ and design matrix

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

In this framework and for the variances of Y_1, Y_2, Y_3 given above, determine the expected Fisher information matrix

$$\mathcal{I}_F(\boldsymbol{\beta}) = \begin{pmatrix} i_{11} & i_{12} \\ i_{12} & i_{22} \end{pmatrix}$$

for the corresponding parameter vector $\boldsymbol{\beta}$.

Hint: Part (b) can be solved independently of part (a), without deriving a representation of the form $(**)$ for the given probability mass function.

Calculate the entries of $\mathcal{I}_F(\boldsymbol{\beta})$. **6 von 6 Punkten**

Give the value of i_{11}. 2 von 2 Punkten

18 ✓

Give the value of i_{12}. 2 von 2 Punkten

-12 ✓

Give the value of i_{22}. 2 von 2 Punkten

14 ✓

dynexite.rwth-aachen.de

Vorherige Aufgabe

Nächste Aufgabe

Gesamtpunktzahl

7 von 7 Punkte

Let $\mu_1, \mu_2 > 0$ and $(X_n)_{n\in\mathbb{N}}$, $(Y_n)_{n\in\mathbb{N}}$ be two sequences of stochastically independent random variables with $X_i \sim \mathcal{P}(\mu_1)$ and $Y_i \sim \mathcal{P}(\mu_2)$ for $i \in \mathbb{N}$, where $\mathcal{P}(\mu_i)$ denotes the Poisson distribution with parameter μ_i for $i \in \{1, 2\}$.

For $n \in \mathbb{N}$, the corresponding arithmetic means of $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$, respectively, are denoted by

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \quad \text{and} \quad \overline{Y}_n = \frac{1}{n}\sum_{i=1}^{n} Y_i\,.$$

In each of the follwing two tasks, determine the asymptotic variance, i.e. the variances of the corresponding limit distributions for the sequences of random variables considered in these parts.

(a) For $\mu_1 = 4$, calculate the asymptotic variance σ_1^2 of the sequence $\left(\sqrt{n}\,(\overline{X}_n - \mu_1)\right)_{n\in\mathbb{N}}$.

3 von 3 Punkten

Give the value of σ_1^2.

3 von 3 Punkten

4 ✓

(b) For $\mu_2 = 1$, calculate the asymptotic variance σ_2^2 of the sequence $\left(\sqrt{n}\left(\sqrt{\overline{Y}_n} - \sqrt{\mu_2}\right)\right)_{n\in\mathbb{N}}$.

4 von 4 Punkten

Give the value of σ_2^2.

4 von 4 Punkten

0.25 ✓

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

Consider the following **ungrouped** data file called `Beetles` which contains data from one of the first studies using a binary regression model in 1935. The study divided a sample of $n_{tot} = 481$ adult flour beetles to 8 groups of size m_j, $j = 1, \ldots, 8$, with $\sum_{j=1}^{8} m_j = n_{tot}$. The beetles were exposed to gaseous carbon disulfide at 8 distinct dosages (in mg/liter), one for every group. The study observed for the beetles of all groups whether they are alive or dead after a 5 hours exposure, which gives us the corresponding response variable Y, taking values $y_i \in \{0, 1\}$, where the value 0 denotes the survival of beetle i, $i = 1, \ldots, n_{tot}$. The explanatory variable is the dosage in log-scale. Hence, if dose_i is the dose to which beetle i is exposed to, then $x_i = \log_{10}(\text{dose}_i)$, $i = 1, \ldots, n_{tot}$.

Beetles

(a) Give the number m_1 of beetles exposed to the dose for which $x_i = \log_{10}(\text{dose}_i) = 1.691$. Further, give the proportion of deaths among the beetles exposed to this particular dose. **1 von 1 Punkt**

m_1 **(requested precision: whole numbers)** 0.5 von 0.5 Punkten

59 ✓

proportion of deaths **(requested precision: 2 digits)** 0.5 von 0.5 Punkten

0.1 ✓

(b) Fit a generalized linear model using the canonical link function for the response Y and treating the explanatory variable as a continuous one. Calculate the sum of squared errors (SSE). **2 von 2 Punkten**

SSE **(requested precision: 2 digits)** 2 von 2 Punkten

59.14 ✓

(b) Fit a generalized linear model using the canonical link function for the response Y and treating the explanatory variable as a continuous one. Calculate the sum of squared errors (SSE). **2 von 2 Punkten**

SSE **(requested precision: 2 digits)** 2 von 2 Punkten

59.14 ✓

(c) Provide the corresponding AIC and BIC values for the model in (b). **2 von 2 Punkten**

AIC **(requested precision: 2 digits)** 1 von 1 Punkt

376.35 ✓

BIC **(requested precision: 2 digits)** 1 von 1 Punkt

384.71 ✓

(d) For the model fitted in (b), give the 95% asymptotic (Wald) confidence interval for the parameter corresponding to the explanatory variable. **0 von 2 Punkten**

lower bound of CI **(requested precision: 3 digits)** 1 Punkt

28.868 ✗ 28.576

upper bound of CI **(requested precision: 3 digits)** 1 Punkt

40.319 ✗ 39.996

(e) What is the proportion of correctly classified observations from the model fitted in (b) using 0.5 as threshold probability? **1 von 1 Punkt**

II.3.18 Example (cancer remission) - continues

Tests of Significance and CIs

☞ Test of no effect ($H_0 : \beta_2 = 0$)

- $z = \hat{\beta}_2/SE = \frac{0.145}{0.059} = 2.45$ $\quad$ ($z^2 = 5.96 \sim \mathcal{X}^2_1$, under H_0, called **Wald** statistic)

  Strong evidence of a positive association between cancer remission and labeling index (p-value $= 0.015$).

☞ Confidence Interval (CI) for β_2

- 95% Wald CI: $\hat{\beta}_2 \pm 1.96(SE) = (0.029, 0.261)$
  (based on inverting the test above, e.g., the 95% CI is the set of β_2 not rejected at the 5% level in testing H_0 against $H_1 : \beta_2 \neq 0$)

II.3.19 Remark

1. Beyond the Wald test statistic, there exist also the likelihood-ratio and the score test statistics. The three types of tests are asymptotically equivalent, when H_0 is true.
2. There exist other types of CIs, based on inverting the likelihood ratio and the score tests.
3. Methods extend to inference for multiple parameters.

(d) For the model fitted in (b), give the 95% asymptotic (Wald) confidence interval for the parameter corresponding to the explanatory variable. **0 von 2 Punkten**

1 Punkt

lower bound of CI **(requested precision: 3 digits)**

28.868 ✗ 28.576

1 Punkt

upper bound of CI **(requested precision: 3 digits)**

40.319 ✗ 39.996

(e) What is the proportion of correctly classified observations from the model fitted in (b) using 0.5 as threshold probability? **1 von 1 Punkt**

1 von 1 Punkt

proportion of correctly classified observations **(requested precision: 2 digits)**

0.83 ✓

(f) Fit a generalized linear model using a probit link. Calculate the sum of squared errors (SSE). Based on the value of SSE, would you prefer the model using a probit link, the model of (b) or both? **2 von 2 Punkten**

1 von 1 Punkt

SSE **(requested precision: 2 digits)**

59.21 ✓

1 von 1 Punkt

preference based on SSE

model of (b) ✓

(a) Choose the statement (statements) that is (are) **true**. **0 von 1 Punkt**

- [ ] The link function used for obtaining the logistic regression model is the log link.
- [ ] Logistic regression assumes a linear relationship between the response variable Y and the explanatory variables.
- [x] If we have a binary response variable, we always have to use logistic regression.
- [x] Logistic regression assumes a linear relationship between the logarithm of the odds of the response and the explanatory variables. ✓
- [ ] The link function used for obtaining the logistic regression model is the identity link.

(b) Choose the assumption (assumptions) that is (are) **not** an assumption in the GLM framework where Y is the response variable and X the explanatory variable. **2 von 2 Punkten**

- [ ] The link function links the expectation of the response with the linear predictor.
- [ ] The conditional probability density (or mass) function (pdf or pmf) of Y given $X = x$ belongs to the exponential dispersion family.
- [x] The response is binary. ✓
- [x] For a random sample of size n, the responses $Y_i, i = 1, \ldots, n$, are independent and identically distributed. ✓
- [x] The conditional probability density (or mass) function (pdf or pmf) of X given $Y = y$ belongs to the exponential dispersion family. ✓

(c) Choose the statement (statements) that is (are) **true**. **0 von 2 Punkten**

- [x] Generalized linear models allow the linear predictor to be non-linear in the parameters β.
- [ ] Generalized linear models are more sensitive to outliers than linear models.
- [x] Generalized linear models can fit complex relationships between the response and the explanatory variables. ✓
- [x] Generalized linear models can handle both continuous and categorical data while linear models can just handle one type of them.
- [ ] In a generalized linear model, the distribution of the error term has to be a normal distribution.

(d) Choose the statement (statements) that is (are) **true** for a GLM. **2 von 2 Punkten**

- [x] The link function links the expected value of the random response variable to the linear predictor. ✓
- [ ] The link function is used to transform the values of the response variable.
- [ ] The link function transforms the expected value of the random response variable to the natural parameter θ of the exponential dispersion family corresponding to the random response variable.
- [ ] For a poisson distributed random response variable, the canonical link is the logit link.

(e) Choose the statement (statements) that is (are) **true** for a GLM. **0 von 1 Punkt**

- [x] The degrees of freedom of a saturated model are always equal to 0. ✓

(d) Choose the statement (statements) that is (are) **true** for a GLM. **2 von 2 Punkten**

- [x] The link function links the expected value of the random response variable to the linear predictor. ✓
- [ ] The link function is used to transform the values of the response variable.
- [ ] The link function transforms the expected value of the random response variable to the natural parameter θ of the exponential dispersion family corresponding to the random response variable.
- [ ] For a poisson distributed random response variable, the canonical link is the logit link.

(e) Choose the statement (statements) that is (are) **true** for a GLM. **0 von 1 Punkt**

- [x] The degrees of freedom of a saturated model are always equal to 0. ✓
- [ ] The saturated model is the model for which the raw residuals are all equal to zero. ✗
- [ ] The saturated model is nested in the null model.
- [ ] The null model is the model for which the raw residuals are all equal to zero.

(f) Consider a simple logistic regression model with parameter vector $\boldsymbol{\beta} = (\beta_1, \beta_2)^T$ and model matrix $\boldsymbol{X} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}$. Choose the statement (statements) that is (are) **2 von 2 Punkten**

II.3.9 Remark (interpretation of β_2)

For two levels of x, denoted by x_1 and x_2,

$$\text{odds ratio} = \frac{\pi(x_1)/[1-\pi(x_1)]}{\pi(x_2)/[1-\pi(x_2)]} = \frac{e^{\beta_1+\beta_2 x_1}}{e^{\beta_1+\beta_2 x_2}} = e^{\beta_2(x_1-x_2)}$$

- For $x_1 - x_2 = 1$, the odds of a success at $x = x_1$ are e^{β_2} times the odds of success at $x = x_2$, i.e., odds multiply by e^{β_2} for every 1-unit increase in x.
- $\beta_2 = 0 \longleftrightarrow$ odds ratio $= 1 \longleftrightarrow$ no effect of x on Y

II.3.10 Remark (generalization to multiple logistic regression model)

$$\log\left(\frac{\pi(x)}{1-\pi(x)}\right) = \beta_0 + \beta_1 x_1 + \ldots + \beta_q x_q$$

In this case, $\exp(\beta_j)$ represents the odds ratio between Y and two levels of x_j that are 1-unit apart ($j = 2, \ldots, p$), adjusting for all other predictors in the model.

(f) Consider a simple logistic regression model with parameter vector $\boldsymbol{\beta} = (\beta_1, \beta_2)^T$ and model matrix $\boldsymbol{X} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}$. Choose the statement (statements) that is (are) **true**. **2 von 2 Punkten**

- [x] Increasing the explanatory variable by one unit, the odds of success for the response variable will be multiplied by $\exp(\beta_2)$.
- [x] If $\beta_2 = 0$ the success probability of the response variable is a constant function of the explanatory variable.
- [ ] The median effective level is the point where the success probability of the response variable is maximized.
- [ ] Increasing the explanatory variable by one unit, the odds of success for the response variable will increase additively by $\exp(\beta_2)$.
- [ ] If $\beta_2 = 0$ the success probability of the response variable is equal to zero.

Vorherige Aufgabe

Nachkorrekturantrag anlegen?

ÜBERSICHT 1 2 3 4 EINSICHT BEENDEN