Exam Statistical Models

02 Jun 2014

Question part

The question part consists of 4 questions giving a total of 36 points. No literature help is allowed for this part. As soon as you are done please hand in your answers to the question part to the "Tentamenvakt" and you will in return get the problem part. The problem part consists of problems giving 64 points in total. Note: Once you hand in the question part it is not possible to get it back nor add any extra answers to it! *Lycka till!*

Question 1 (9 Points)

What is an exponential family (use text and equations in your answer)? Why are exponential families useful (use only text in your answer)?

Question 2 (9 points)

What is the principle of an "exact test" in the setting of an exponential family?

Question 3 (9 Points)

Let $f(x;\theta)$ denote a statistical model parametrized by parameter vector θ. Give the general formula for the score test, allowing nuisance parameters. Give the formula for Wald's test, allowing nuisance parameters.

Question 4 (9 Points)

Define what characterizes a univariate Generalized Linear Model (GLM). What is a GLM with additional dispersion? When is dispersion useful? What other weights are imaginable?

Exam Statistical Models

02 June 2014

Problem part

The problem part consists of 4 problems giving a total of 64 points. The following literature is allowed while solving the problem part: lectures notes of the course, printout of the course slides, printout of the Chapter 3 excerpt from Fahrmeir & Tutz provided on the webpage, your own handwritten annotations from the course, possible formula collections from other courses. No additional books or literature photocopies are allowed. *Lycka till!*

Problem 1 (26 Points)

Let $Y_1 \sim \text{Exp}(\lambda_1)$ and $Y_2 \sim \text{Exp}(\lambda_2)$ be two independent exponential random variables each having expectation $1/\lambda_i$, $i = 1, 2$.

(a) (5 Points) Derive the maximum likelihood estimator $\hat{\lambda}_i$ for $i = 1, 2$ and show that

$$\text{se}(\hat{\lambda}_i) = \frac{1}{y_i}.$$

(b) (5 Points) What is the maximum likelihood estimator and the associated standard error of $\phi_i = \log \lambda_i$?

(c) (6 Points) Of interest is the ratio $\theta = \lambda_1/\lambda_2$. First construct a 95% Wald confidence interval for $\log(\theta)$ based on the observations $y_1 = 1$ and $y_2 = 10$. Then transform this interval back onto the θ-scale. Test the null hypothesis $H_0 : \theta = 1$ against $H_1 : \theta \neq 1$ at the $\alpha = 0.05$ significance level. *Hint:* Use the results from part (b).

(d) (5 Points) Derive the profile loglikelihood of θ. Do so by re-parameterizing the loglikelihood function of λ_1 and λ_2 to use $\theta = \lambda_1/\lambda_2$ and $\lambda = \lambda_2$ instead.

(e) (5 Points) Sketch how a 95% confidence interval for θ based on the profile loglikelihood is found. For the above data this 95% confidence interval is $[0.396, 252.6]$. Compare this interval with the interval from part (c) and discuss possible reasons for why the two intervals are rather different.

Problem 2 (12 Points)

In a branching process, we have n independent observations following sequentially over time, each having a Poisson distribution with mean

$$\mu \cdot (\alpha^i - \alpha^{i-1}), \quad \text{where } \alpha > 1, \mu > 0, i = 1, \dots, n.$$

(a) (2 Points) Write down the joint density of the n observations.

(b) (4 Points) Show that this distribution is a member of the exponential family with canonical parameters

$$\boldsymbol{\theta} = (\theta_1, \theta_2)^T = (\log(\mu) + \log(\alpha - 1)\,,\ \log(\alpha))^T.$$

(c) (2 Points) Also state the corresponding canonical statistic $\boldsymbol{t} = (u, v)^T$ to the parameterization given in (b).

(d) (4 Points) Show that the mean value parametrization for the u component of the canonical statistic is

$$\mu_u(\boldsymbol{\theta}) = \frac{\exp(\theta_1 + n\theta_2) - \exp(\theta_1)}{\exp(\theta_2) - 1}.$$

Problem 3 (14 Points)

The papers of 200 students for an examination were marked separately by two different examiners, who classified each paper as either pass or fail. The results were as follows:

	Examiner B		
Examiner A	Pass	Fail	Total
Pass	y_{11}=136	y_{10}=2	$y_{1.}$=138
Fail	y_{01}=16	y_{00}=46	$y_{0.}$=62
Total	$y_{.1}$=152	$y_{.0}$=48	$y_{..}$=200

Interest centers on whether examiner A fails more students than examiner B.

(a) (6 Points) Use a multinomial distribution to argue that a good way to investigate this hypothesis is to investigate $H_0 : p = 0.5$ vs. $H_1 : p > 0.5$ in a probability model $y_{01} \sim \text{Bin}(y_{01} + y_{10}, p)$.

(b) (5 Points) The R output of an exact test investigating the above hypothesis for the given data is as follows:

```
> binom.test( x=16,n=18, p=0.5, alternative="greater")

Exact binomial test

data:  16 and 18
number of successes = 16, number of trials = 18, p-value = 0.0006561
alternative hypothesis: true probability of success is greater than 0.5
95 percent confidence interval:
 0.6897373 1.0000000
```

Explain and give an equation for how the p-value is calculated. Interpret the result for the above data on the two examiners.

(c) (3 Points) Describe how the above one-sided 95% confidence interval is calculated.

Problem 4 (12 Points)

Consider the modelling of an ordinal response variable Y with k categories using the cumulative approach. Specifically, we look at the modelling

$$P(Y \leq r|\boldsymbol{x}) = F(\theta_r + \boldsymbol{x}^T\boldsymbol{\gamma}), \quad r = 1, \ldots, k-1,$$

where F is a given cumulative distribution function (CDF), $\boldsymbol{x}$ and $\boldsymbol{\gamma}$ are p-dimensional column vectors of covariates and regression coefficients, respectively, and $-\infty = \theta_0 < \theta_1 < \ldots < \theta_k = \infty$.

(a) (4 Points) Assume that one is using the logistic CDF $F(x) = \exp(x)/(1 + \exp(x))$. Show that in this case the odds for the event $Y \leq r|\boldsymbol{x}$ for a selected r, $1 \leq r \leq k-1$, is

$$\frac{P(Y \leq r|\boldsymbol{x})}{P(Y > r|\boldsymbol{x})} = \exp(\theta_r + \boldsymbol{x}^T\boldsymbol{\gamma}).$$

(b) (4 Points) Consider a j, $1 \leq j \leq p$. Let $\tilde{\boldsymbol{x}}$ and $\boldsymbol{x}$ be two covariate configurations which are equal except at the j'th position, i.e. $\tilde{\boldsymbol{x}}_l = \boldsymbol{x}_l$ for $l \neq j$ and $\tilde{\boldsymbol{x}}_j = \boldsymbol{x}_j + 1$. Use this setup and the result from (a) to interpret the value of $\exp(\gamma_j)$.

(c) (4 Points) Assume now that one instead is using the extreme-minimal-value CDF $F(x) = 1 - \exp(-\exp(x))$. In this case one obtains the so called *grouped Cox model* with

$$P(Y \leq r|\boldsymbol{x}) = 1 - \exp(-\exp(\theta_r + \boldsymbol{x}^T\boldsymbol{\gamma})), \quad r = 1, \ldots, k-1.$$

An alternative representation of the grouped Cox model is

$$P(Y = r|Y \geq r, \boldsymbol{x}) = 1 - \exp(-\exp(\tilde{\theta}_r + \boldsymbol{x}^T\boldsymbol{\gamma})), \quad r = 1, \ldots, k-1,$$

with $\tilde{\theta}_r = \log(\exp(\theta_r) - \exp(\theta_{r-1}))$. Show that the two representations are equivalent.