Exam Statistical Models

28 May 2013

Question part

The question part consists of 4 questions giving a total of 40 points. No literature help is allowed for this part. As soon as you are done please hand in your answers to the questions part to the "Tentamenvakt" and you will in return get the problem part which consists of problems giving 60 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!

Question 1 (10 Points)

Show that an exponential family is preserved under repeated sampling from the same distribution, and tell in what sense (what will change and what does not change?)

Question 2 (10 Points)

Write down the score function and the Fisher information for a full exponential family in canonical parametrization.

Question 3 (10 Points)

Define what characterizes a univariate Generalized Linear Model (GLM). How do ideas extend to multivariate GLMs? Also explain how to interpret a non-intercept coefficient in a univariate Poisson log-linear GLM.

Question 4 (10 Points)

Consider a one-parameter parametric model $f(y;\theta)$. An iid. sample of size n has been observed from the model. Motivate from the likelihood ratio test the construction of a $(1-\alpha)\cdot 100\%$ likelihood based confidence interval for θ and describe how the interval is computed.

Exam Statistical Models

28 May 2013

Problem part

The problem part consists of 3 problems giving a total of 60 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, printout of the Höhle (2010) article linked on the webpage, your own handwritten annotations from the course, possible formula collections from other courses. No additional books or literature photocopies are allowed.

Problem 1 (12 points)

Consider the Poisson distribution with probability mass function (PMF) $f(y;\lambda)$ and $E(y)=\lambda$.

(a) (9 points) Show that the saddlepoint approximation for the PMF is

$$f(y;\lambda) \approx \left[\sqrt{2\pi}\,\exp(-y)\,y^{y+1/2}\right]^{-1} \lambda^y \exp(-\lambda).$$

(b) (3 points) How can this approximation be additionally improved?

Problem 2 (33 points)

The so called *double Poisson distribution* has probability mass function

$$f(y;\theta,\psi) = c(\theta,\psi)\sqrt{\psi}\,\frac{\exp(-\psi\exp(\theta))}{y!}\left(\frac{y}{e}\right)^y \left(\frac{\exp(\theta)e}{y}\right)^{y\psi}, \quad y=0,1,\ldots \tag{1}$$

where $\theta \in \mathbb{R}$, $\psi > 0$ and $e=\exp(1)$. Furthermore, $c(\theta,\psi)$ is a function depending only on θ and ψ, which normalizes the density. In general the exact form of $c(\theta,\psi)$ can be very hard to determine and you should not attempt to do so.

(a) (5 points) Show that the double Poisson distribution belongs to the two-parameter exponential family with canonical parameter vector $(\theta\psi,\psi)^T$ and canonical statistic

$$t(y) = (y, y(1-\log(y)))^T.$$

Furthermore, show that the overall normalization function of the double Poisson is

$$C(\theta,\psi) = [c(\theta,\psi)\sqrt{\psi}\exp(-\psi\exp(\theta))]^{-1}.$$

(b) (6 points) Assume $c(\theta, \psi) = 1$, which in many cases is a very good approximation, and show that for the double Poisson distribution

$$E(t) = \begin{bmatrix} \exp(\frac{\alpha}{\psi}) \\ -\frac{1}{2\psi} + \exp(\frac{\alpha}{\psi}) \left(1 - \frac{\alpha}{\psi}\right) \end{bmatrix},$$

where we have used a parametrization with $\alpha = \theta\psi$ and ψ.

(c) (6 points) What is the expectation of a double Poisson variable under the $c(\theta, \psi) = 1$ assumption? What is the variance? Use this to interpret the parameter ψ.

(d) (3 points) Using a mixed representation (μ, ψ) it is possible to show (don't do it!) that the expected Fisher information for a random sample of size n under the above assumption is

$$I(\mu, \psi) = \begin{bmatrix} \frac{n\psi}{\mu} & 0 \\ 0 & \frac{n}{2\psi^2} \end{bmatrix}.$$

What is the advantage of the mixed representation here compared to $I(\alpha, \psi)$?

(e) (6 points) Show that the MLE for a random sample of size n under the simplifying assumption $c(\theta, \psi) = 1$ is

$$\hat{\mu} = \overline{y} \quad \text{and} \quad \hat{\psi} = \frac{n}{D(\boldsymbol{y}, \overline{y})},$$

where $D(\boldsymbol{y}, \boldsymbol{\mu}(\hat{\beta}))$ denotes the Poisson deviance of a GLM with expectation vector $\boldsymbol{\mu}(\hat{\beta})$, i.e.

$$D(\boldsymbol{y}, \boldsymbol{\mu}(\hat{\beta})) = 2\sum_{i=1}^{n} \left\{ y_i \log\left(\frac{y_i}{\mu_i(\hat{\beta})}\right) - (y_i - \mu_i(\hat{\beta})) \right\}.$$

(f) (7 points) Again assuming $c(\theta, \psi) = 1$ and considering iid. count data $y_1, \ldots, y_n$. Formulate the hypothesis and derive the Wald test statistic for the investigation of whether the double Poisson distribution provides a better fit to the data than the ordinary one-parameter Poisson distribution. What is the asymptotic distribution of this test statistic?

Problem 3 (15 points)

An alternative to estimate the parameters of an exponential family from a single iid. sample of size n is by using a multinomial type likelihood. In the discrete response case let $0, 1, 2, \ldots, k$ denote the possible outcomes and let n_i, $i = 0, \ldots, k$ denote the number of individuals in the population who have outcome $Y = k$, i.e. $n = \sum_{i=0}^{k} n_i$. As an example the data `tab4` in Table 1 contain the number of accidents to $n = 647$ female workers in an ammunition factory during the 1st World War.
The trick is now to model the cell frequencies using a Poisson log-linear GLM, i.e. $n_i \sim \text{Po}(\lambda_i)$, $i = 1, \ldots, k$, with

$$\log \lambda_i = \theta^T t(y_i) - \log C(\theta) + h(y_i). \tag{2}$$

We observe that $h(y_i)$ does not depend on the parameters to be estimated and hence can be included as offset in the modelling. Furthermore, $-\log C(\theta)$ is the intercept in the model. Actually, retaining an intercept in the Poisson model ensures that the marginal total is fixed to be n (c.f. Poisson trick!). A consequence of this is that we do not have to know the analytic form of $-\log C(\theta)$ in order to fit the model.

y	n	ylogy
0	447	0.000
1	132	0.000
2	42	1.386
3	21	3.296
4	3	5.545
5	2	8.047
6	0	10.751
7	0	13.621
8	0	16.636

Table 1: Data `tab4` on the number of accidents to 647 female workers in an ammunition factory. In the original data there were no workers with more than 5 accidents. We extend the dataset up to $k = 8$ and assume that ignoring outcomes higher than 8 only induces a negligible error.

(a) (7 points) Show that when using the above model to describe the conditional distribution $n_1, \ldots, n_k | n$, this results in a multinomial type likelihood with class-specific probabilities $\lambda_i / \sum_{j=1}^{k} \lambda_j$. What happens with $C(\theta)$ when inserting (2) as λ_i?

(b) (3 points) When the underlying exponential family distribution is the simple Poisson distribution, $\text{Po}(\exp(\theta))$, the above procedure applied to the data from `tab4` generates the following R output shown below[1]. Use the output to compute a 95% Wald confidence interval for θ in the simple Poisson model.

```
> summary(m1)

Call:
glm(formula = n ~ 1 + y + offset(-lfactorial(y)), family = poisson,
    data = tab4)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-4.3880  -0.1070  -0.0067   1.9862   4.3453

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  6.00712    0.04759  126.23   <2e-16
y           -0.76524    0.05764  -13.28   <2e-16

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 286.425  on 8  degrees of freedom
Residual deviance:  55.098  on 7  degrees of freedom
```

(c) (5 points) For the one sample case the approach can also be used to fit the double Poisson distribution from Exercise 2 without the $c(\theta, \psi) \approx 1$ assumption by using Poisson GLM methodology. In this case the estimated parameters are α and $(\psi - 1)$ (don't do the math to show this!):[2]

```
> summary(m2)
```

[1] The function `lfactorial(y)` computes $\log(y!)$.

[2] The function `I(y-ylogy)` means to treat its argument "as-is", i.e. a new covariate $y - y\log(y)$ is computed and then used in the GLM.

```
Call:
glm(formula = n ~ 1 + y + I(y - ylogy) + offset(-lfactorial(y)),
    family = poisson, data = tab4)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-0.89177  -0.71327  -0.22746   0.07055   1.31000

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   6.10193    0.04716 129.387  < 2e-16
y            -0.58507    0.04762 -12.285  < 2e-16
I(y - ylogy) -0.64021    0.08016  -7.986 1.39e-15

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 286.4245  on 8  degrees of freedom
Residual deviance:   4.2805  on 6  degrees of freedom
```

Use the previous and subsequent R output to compute the likelihood ratio test statistic for the hypothesis formulated in Exercise 2(f). Discuss your results with respect to goodness of fit of the simple Poisson model for the accident data.