Norges teknisk–
naturvitenskapelige universitet
Institutt for matematiske fag

English

Contact during exam:
Håkon Tjelmeland 73 59 35 38

EXAM IN TMA4300 MODERN STATISTICAL METHODS

Thursday June 7th 2007
Time: 09:00–13:00

Aids: Calculator HP30S.
Statistiske tabeller og formler, Tapir forlag.
K. Rottman: Matematisk formelsamling.
One yellow paper (A4 with stamp) with own formulas and notes.

Grading: June 28th 2007.

Oppgave 1

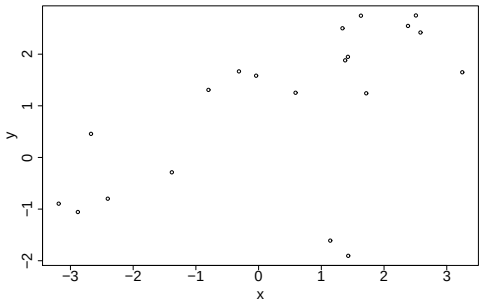
Let $x \sim \mathrm{N}(0,1)$ and $y|x \sim \mathrm{Poisson}(e^x)$. In the (marginal) distribution for y we want to find the parameter

$$\theta = \mathrm{E}[\sqrt{y}].$$

We assume we have available functions that generates independent realisations from the standard normal and Poisson distributions.

Write pseudo code and equations necessary to calculate the value of an estimator for θ, $\widehat{\theta}$, by Monte Carlo simulation.

Using x as a control variate it is possible to define an estimator for θ, $\tilde{\theta}$, with smaller variance than $\widehat{\theta}$. Define $\tilde{\theta}$ and write pseudo code and equations necessary to calculate the value of the new estimator $\tilde{\theta}$.



Figur 1: Observed data in Exercises 2 and 3.

Oppgave 2

Figure 1 shows an observed data set, $\{(x_i, y_i)\}_{i=1}^n$.

In this exercise we consider a regression analysis and use a Bayesian model. We assume a regression model with three parameters, $\alpha \in \mathbb{R}$, $\beta \in \mathbb{R}$ and $\lambda > 0$, where

$$\mathrm{E}[y_i|\alpha, \beta, \lambda] = \alpha + \beta(x_i - \bar{x}),$$

with $\bar{x} = (1/n)\sum_{i=1}^n x_i$. We assume the y_i's to be independent (given the parameters α, β and λ), and to get a robust analysis we assume y_i to have a double exponential distribution (Laplace distribution) with mean $\alpha + \beta(x_i - \bar{x})$ and intensity λ, i.e.

$$\pi(y_i|\alpha, \beta, \lambda) = \frac{\lambda}{2} \exp\left\{-\lambda|y_i - (\alpha + \beta(x_i - \bar{x}))|\right\}.$$

We assume α, β and λ to be apriori independent. For α and β we assume prior distributions

$$\alpha \sim \mathrm{N}(0, 10^2) \quad \text{and} \quad \beta \sim \mathrm{N}(0, 10^2),$$

respectively, and for λ we assume the prior distribution

$$\pi(\lambda) = \frac{\lambda}{4} \exp\left\{-\frac{\lambda}{2}\right\} \qquad \text{for} \qquad \lambda > 0.$$

a) Visualise the Bayesian model as a graphical model.

Show that the full conditional distribution for λ can be expressed as

$$\pi(\lambda|\alpha, \beta, y_1, \ldots, y_n) \propto \lambda^{n+1} \exp\left\{-\lambda\left(\frac{1}{2} + Q(\alpha, \beta, y_1, \ldots, y_n)\right)\right\},$$

and find thereby an expression for $Q(\alpha, \beta, y_1, \ldots, y_n)$.

b) Show that a realisation from the full condition distribution for λ can be generated by first sampling v from a χ^2 distribution with a suitable degrees of freedom ν and thereafter setting

$$\lambda = v \cdot r,$$

for a suitable value r. Find the suitable values for ν and r. [*Hint: Find first the density function for the resulting* λ *as a function of* ν *and* r *and compare thereafter this expression with the full conditional distribution found in the previous item.*]

c) Write pseudo code generating samples from the posterior distribution $\pi(\alpha, \beta, \lambda | y_1, \ldots, y_n)$ by a Metropolis–Hastings algorithm. Specify what proposal distributions you use and find expressions for the corresponding acceptance probabilities as functions of the parameters in the problem. The acceptance probability expressions should be simplified as much as possible.

Let $\{\alpha^k, \beta^k, \lambda^k\}_{k=1}^K$ be values simulated in the MCMC algorithm you specified in item **c**), where K is the number of iterations run.

d) Specify how you from $\{\alpha^k, \beta^k, \lambda^k\}_{k=1}^K$ (and if necessary other simulated values, which you then have to specify how to sample) can estimate

1. $\mathrm{P}(\beta > 0 | y_1, \ldots, y_n)$
2. $\mathrm{E}[y_0 | y_1, \ldots, y_n]$, where y_0 is a new observation for $x = x_0$
3. $\pi(y_0 | y_1, \ldots, y_n)$, where y_0 is again a new observation for $x = x_0$

Oppgave 3

In this exercise we consider the same regression problem as in Exercise 2, but now in a non-Bayesian setting and using bootstrapping in the analysis. Thus, the observed data are $\{(x_i, y_i)\}_{i=1}^n$ and the model for y_i is

$$y_i = \alpha + \beta(x_i - \bar{x}) + \varepsilon_i,$$

where $\varepsilon_1, \ldots, \varepsilon_n$ are independent and from a zero mean double exponential distribution with intensity λ,

$$\pi(\varepsilon_i) = \frac{\lambda}{2} \exp\left\{-\lambda |\varepsilon_i|\right\}.$$

The maximum likelihood estimators for α and β in this model can be found by minimising

$$R(\alpha, \beta, y_1, \ldots, y_n) = \sum_{i=1}^{n} |y_i - (\alpha + \beta(x_i - \bar{x}))|$$

with respect to α and β. In this exercise we will assume we have available a (numerical) minimisation algorithm that do this minimisation and return the corresponding parameter estimates $\widehat{\alpha}$ and $\widehat{\beta}$. When $\widehat{\alpha}$ and $\widehat{\beta}$ are obtained, the maximum likelihood estimator for λ, $\widehat{\lambda}$, can easily be found as

$$\widehat{\lambda} = \frac{n}{R(\widehat{\alpha}, \widehat{\beta}, y_1, \ldots, y_n)}.$$

a) To generate a bootstrap sample $\{(x_i^\star, y_i^\star)\}_{i=1}^n$ one can imagine (at least) three methods, one parametric and two non-parametric. For each of these three methods, describe in detail how the bootstrap samples can be generated (you may use pseudo code).

Discuss briefly advantages and disadvantages with the three resampling methods. Are there specific situations where one of the three methods is preferable?

Let $\{(x_i^{\star(b)}, y_i^{\star(b)})\}_{i=1}^n, b = 1, \ldots, B$ be generated bootstrap samples (using one of the methods discussed in item **a**).

b) Write pseudo code and equations necessary that describe how one from $\{(x_i^{\star(b)}, y_i^{\star(b)})\}_{i=1}^n, b = 1, \ldots, B$ can

1. estimate $\mathrm{SD}(\widehat{\beta})$
2. estimate a confidence interval (percentile interval) for $\mathrm{E}[y_0] = \alpha + \beta x_0$, where y_0 is a new observation for $x = x_0$

Finally we will consider how to generalise the idea used when defining the percentile confidence interval to define also a bootstrap (percentile) prediction interval for a new observation y_0 for $x = x_0$.

c) For each of the three resampling methods discussed in item **a**), discuss the possibility of using bootstrapping to estimate a prediction interval for y_0. For the resampling method(s) where this is possible, give pseudo code and necessary equations that give how to compute the interval. For the resampling method(s) where it is not possible to generate a prediction interval, explain why it is not possible.