UNIVERSITETET I OSLO

Matematisk Institutt

EXAM IN:	**STK 4020 – Model Selection and Model Averaging**
	Part I of two parts: The project
WITH:	**Nils Lid Hjort**
TIME FOR EXAM:	**30.v.–9.vi.2011**

This is the exam project set for STK 4160, spring semester 2011. It is made available on the course website as of *Monday 30 May 12:00*, and candidates must submit their written reports by *Thursday 9 June 14:00* (or earlier), to the reception office at the Department of Mathematics, in duplicate. The supplementary oral examinations take place Tuesday June 14 (practical details concerning this are provided elsewhere). Reports may be written in nynorsk, bokmål, riksmål, English or Latin, and should preferably be text-processed (TeX, LaTeX, Word), but may also be hand-processed. Give your name on the first page. Write concisely (in der Beschränkung zeigt sich erst der Meister; brevity is the soul of wit; краткость – сестра таланта). Relevant figures need to be included in the report. Copies of machine programmes used (in `R`, or `matlab`, or similar) are also to be included, perhaps as an Appendix to the report. Candidates are required to work on their own (i.e. without cooperation with any others), but are graciously allowed not to despair if they do not manage to answer all questions well.

Importantly, each student needs to submit *two special extra pages* with her or his report. *The first* (page A) is the 'erklæring' (self-declaration form), properly signed; it is available at the webpage as 'Exam Project, page A, declaration form'. *The second* (page B) is the student's one-page summary of the exam project report, which should also contain a brief self-assessment of its quality.

This exam set contains three exercises and comprises seven pages. The final page gives some potentially useful hints for **R** programming.

Exercise 1

DOUBLE, DOUBLE TOIL AND TROUBLE, fire burn and cauldron bubble: the so-called double exponential distribution, with parameter λ, has density $f(y,\lambda) = \frac{1}{2}\lambda \exp(-\lambda|y|)$, and has many uses in statistics and probability theory. It is used to fit data (typically with an additional parameter for location), but also as a background distribution for parameters, as for the so-called lasso method of high-dimensional regression. The present exercise concerns a certain extension of the double exponential distribution to allow for asymmetry.

(a) Suppose an i.i.d. sample $y_1, \ldots, y_n$ is observed from the distribution above. Find a formula for its maximum likelihood estimator $\widehat{\lambda}$, and determine the limit distribution of $\sqrt{n}(\widehat{\lambda} - \lambda)$.

(b) As focus parameter μ we shall take the upper probability that Y exceeds a given positive threshold y_0. Show that $\mu = \frac{1}{2}\exp(-\lambda y_0)$, under current model assumptions, and determine as above the limiting distribution of $\sqrt{n}(\widehat{\mu}_{\rm narr} - \mu)$, where $\widehat{\mu}_{\rm narr}$ is the maximum likelihood estimator for μ inside this narrow start model.

(c) Our model extension is given by the following density function:

$$f(y,\lambda,a) = \tfrac{1}{2}\lambda(1-a^2)\begin{cases}\exp\{-\lambda(1+a)|y|\} & \text{for } y \geq 0,\\ \exp\{-\lambda(1-a)|y|\} & \text{for } y \leq 0.\end{cases}$$

Verify that it is a density. What is the canonical parameter region for the extra parameter a? For a fixed value of λ, choose some values for a, and display the corresponding densities in a diagram. Comment briefly on this two-parameter model.

(d) Show that the Fisher information matrix for this model can be written in the form

$$J = \begin{pmatrix} 1/\lambda^2, & -c/\lambda \\ -c/\lambda, & d \end{pmatrix},$$

where

$$c = \frac{2a}{1-a^2} \quad \text{and} \quad d = \frac{2+2a^2}{(1-a^2)^2}.$$

For this point it may be useful to write the density in the form

$$f(y,\lambda,a) = \tfrac{1}{2}\lambda(1-a^2)\exp\{-\lambda(1+a)A(y) - \lambda(1-a)B(y)\},$$

in which

$$A(y) = \max\{0,y\} = \begin{cases} 0 & \text{if } y \leq 0,\\ |y| & \text{if } y \geq 0,\end{cases} \quad \text{and} \quad B(y) = \max\{0,-y\} = \begin{cases} |y| & \text{if } y \leq 0,\\ 0 & \text{if } y \geq 0.\end{cases}$$

(e) Show that the upper probability above threshold y_0 in this more general model can be expressed as

$$\mu = \Pr\{Y \geq y_0\} = \tfrac{1}{2}(1-a)\exp\{-(1+a)\lambda y_0\}.$$

Letting $\widehat{\mu}_{\rm wide}$ be the maximum likelihood estimator in this wider model, show that the limit distribution of $\sqrt{n}(\widehat{\mu}_{\rm wide} - \mu)$ is normal and find an expression for its variance. For the special case where the narrow model is actually correct, corresponding to $a = 0$, how much is lost in terms of precision, when one uses wide model estimation rather than narrow model estimation?

(f) Use the theory developed in the course to identify the tolerance radius around the double exponential model inside which inference using that narrow model is still more precise than using the wider model.

(g) In the local asymptotic framework where $a = \delta/\sqrt{n}$, identify the limiting distribution of $D_n = \sqrt{n}\widehat{a}_{\rm wide}$, again using theory developed in the course. What is the (approximate) relationship between selecting a model here via the AIC and via the size of D_n?

(h) Use the theory developed in the course to provide the joint limit distribution of

$$\begin{pmatrix} \sqrt{n}(\widehat{\mu}_{\rm narr} - \mu_n) \\ \sqrt{n}(\widehat{\mu}_{\rm wide} - \mu_n) \\ \sqrt{n}\widehat{a}_{\rm wide} \end{pmatrix} = \begin{pmatrix} \sqrt{n}(\widehat{\mu}_{\rm narr} - \mu_n) \\ \sqrt{n}(\widehat{\mu}_{\rm wide} - \mu_n) \\ D_n \end{pmatrix},$$

again in the local asymptotic framework where $a = \delta/\sqrt{n}$, and with μ_n equal to the expected absolute value of y.

(i) Finally identify the limit distribution of $\widehat{\mu}_{\rm final}$, the estimator used after applying the AIC, again in the $a = \delta/\sqrt{n}$ framework. Briefly discuss any relevant aspects of this limit distribution.

Exercise 2

MEN MENNESKENES HJERTER forandres aldeles intet i alle dager, says Sigrid Undset, though we sometimes try, particularly if our hearts are at risk for entering cardiovascular difficulties. The data set `ldl-data` (to be accessed from the course website) contains observations pertaining to 200 individuals from South Africa and their levels of various cardiovascular risk factors. I have taken these data from a certain larger study (there were in particular even more individuals in the protocols, along with more covariates recorded per person), but have organised this subset of persons and covariates in order to have a simpler yet meaningful data set to work through for the present exam project. I should also make clear that the individuals selected for this study were considered to come from certain higher-than-normal risk groups, so they are in particular not to be seen as a random sample from the healthy population.

The data set focuses on y, the LDL or low-density lipoprotein level (also associated with so-called 'bad cholesterol'), which is recognised as a strong predictor for coronary heart problems, and the questions taken up relate to how y may be understood or predicted from other covariates,

- x_1, age (in years) at the onset of the study;
- x_2, adiposity (a measure of fatness, but different from e.g. the bmi);
- z_1, tobacco use (equivalent to the number of cigarettes per day, I think);
- z_2, presence (1) or absence (0) of family history of related illness;
- z_3, alcohol use (where I have not been able to find the precise definition of the scale being used).

The notation indicates that we are to take x_1 and x_2 protected but that z_1, z_2, z_3 are candidates for exclusion and/or inclusion when we attempt to construct good models. The LDL is here measured in mmol/L (other scales are also in use).

(a) Run a full linear regression of the type

$$y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \gamma_1 z_{i,1} + \gamma_2 z_{i,2} + \gamma_3 z_{i,3} + \varepsilon_i \quad \text{for } i = 1, \ldots, n,$$

with where $n = 200$ is sample size and where the ε_i are taken i.i.d. $\mathrm{N}(0, \sigma^2)$. Give the maximum likelihood estimates of all seven parameters, along with 90% confidence intervals. Briefly discuss these results.

(b) We shall now treat the linear regression model containing only x_1 and x_2 as 'the narrow model' and the fuller model dealt with in point (a) as 'the wide model'. Now fit each of the eight models 0, 1, 2, 3, 12, 13, 23, 123, with notation reflecting which of the three additional covariates z_1, z_2, z_3 are to be included or not, and give a table that displays dimension (the number of parameters being estimated); log-likelihood maximum; the AIC; the BIC; and an approximation to the probability of each model in the Bayesian setup that takes each of the eight models equally likely a priori. Briefly discuss your findings.

(c) Next we are to carry out focussed model selection, attempting to find an optimal model for a typified high-risk individual. More precisely the object is to estimate $\mu = \mathrm{E}(y \,|\, x_0, z_0)$, the expected LDL level for a person who is 70 years old, who smokes 20 cigarettes a day and has an alcohol use corresponding to 20 on the given scale, and who has a family history of heart disease; we finally set his adiposity level at $\bar{x}_2$, the average x_2 value in the data set. Carry out a FIC analysis, complete with a suitable table and plot, including the eight different estimates of μ.

(d) Then carry out a similar FIC analysis for the case of a typified low-risk individual. We make him 20 years old, a non-smoker and non-drinker and with no family history of heart disease, and the same adiposity level as his older fellow of point (c). Discuss any notable differences regarding the best models.

(e) Going back to the high-risk man of point (c), compute model averaging estimates of μ using respectively 'smoothed AIC' and 'smoothed FIC' weights. Briefly comment on your findings.

(f) We are now imagining a certain adult person's life, with his age running through the sequence $20, 21, 22, \ldots, 69, 70$, and with his other covariates fixed – he smokes ten cigarettes a day; he has a family history with heart problems; his alcohol use corresponds to the value 10 on that scale; and his adiposity level is fixed at $\bar{x}_2$. Carry out model selection via 'weighted FIC', where we take each of his years from 20 to 70 equally seriously. For the finally selected model, display his expected LDL level as a function of his age. Include with this plot also a pointwise 90% confidence band, i.e. curves $\widehat{\mu}_{\mathrm{low}}(x_1)$ and $\widehat{\mu}_{\mathrm{up}}(x_1)$ so that the interval from lower to upper point covers the intended $\mu(x_1)$ with probability approximately 90%, for each x_1. (For this occasion you are allowed to be content to construct this band without taking into account the uncertainty involved in selecting the AFIC model in the first place. If you have time and energy you may attempt to correct the band appropriately.)

(g) Without spending too many forces to explore too many possibilities, attempt to build one or a couple of more models for these data, and check if this leads to an improvement, e.g. in terms of AIC.

Exercise 3

YOUR EXACT VALUE may not matter as much as the crowd you're in. This is also true with the LDL levels. The American Heart Association and other medical organisations operate with different categories of LDL, depending also on age and other factors. For the present purposes we shall care about three categories:

. type 1: LDL ≤ 3.3 (essentially normal);
. type 2: $3.3 <$ LDL ≤ 4.9 (somewhat high, carrying some risk);
. type 3: LDL > 4.9 (very high, person in need of great concern and perhaps corrective treatment).

Each of the models used in Exercise 2 (including those you may have been tempted to work with in point (g)) may be used to form models of

$$p_1(x_1,x_2,z_1,z_2,z_3) = \Pr\{y \text{ of type } 1 \,|\, x_1,x_2,z_1,z_2,z_3\},$$
$$p_2(x_1,x_2,z_1,z_2,z_3) = \Pr\{y \text{ of type } 2 \,|\, x_1,x_2,z_1,z_2,z_3\},$$
$$p_3(x_1,x_2,z_1,z_2,z_3) = \Pr\{y \text{ of type } 3 \,|\, x_1,x_2,z_1,z_2,z_3\},$$

simply by reading off category probabilities from the continuous distributions. Sometimes such models use 'too much statistical energy' modelling aspects of y that are eventually less important, however, so it is a reasonable challenge to model p_1,p_2,p_3 directly, basing the analysis on

$$N_{i,1} = \begin{cases} 1 & \text{if no. } i \text{ is of type 1,} \\ 0 & \text{if else;} \end{cases}$$
$$N_{i,2} = \begin{cases} 1 & \text{if no. } i \text{ is of type 2,} \\ 0 & \text{if else;} \end{cases}$$
$$N_{i,3} = \begin{cases} 1 & \text{if no. } i \text{ is of type 3,} \\ 0 & \text{if else.} \end{cases}$$

(a) Show that the likelihood for these mapped data may be represented as

$$L_n = \prod_{i=1}^n (p_{i,1}^{N_{i,1}} p_{i,2}^{N_{i,2}} p_{i,3}^{N_{i,3}}), \quad \text{where } p_{i,j} = \Pr\{\text{no. } i \text{ is of type } j\} \text{ for } j=1,2,3.$$

– One particular model for such data, with three ordered categories, is as follows, where we first transform the original covariates by subtracting their mean values:

$$x_{i,1}^* = x_{i,1} - \bar{x}_1, \quad x_{i,2}^* = x_{i,2} - \bar{x}_2, \quad z_{i,1}^* = z_{i,1} - \bar{z}_1, \quad z_{i,2}^* = z_{i,2} - \bar{z}_2, \quad z_{i,3}^* = z_{i,3} - \bar{z}_3.$$

This turns out to help both numerical calculations and interpretation (and the 'average individual' now has covariate vector $(0,0,0,0,0)$). The model takes

$$p_{i,1} = H(a_0 + \beta_1 x_{i,1}^* + \beta_2 x_{i,2}^* + \gamma_1 z_{i,1}^* + \gamma_2 z_{i,2}^* + \gamma_3 z_{i,3}^*),$$
$$p_{i,2} = H(b_0 + \beta_1 x_{i,1}^* + \beta_2 x_{i,2}^* + \gamma_1 z_{i,1}^* + \gamma_2 z_{i,2}^* + \gamma_3 z_{i,3}^*) - H(a_0 + \beta_1 x_{i,1}^* + \beta_2 x_{i,2}^* + \gamma_1 z_{i,1}^* + \gamma_2 z_{i,2}^* + \gamma_3 z_{i,3}^*),$$
$$p_{i,3} = 1 - H(b_0 + \beta_1 x_{i,1}^* + \beta_2 x_{i,2}^* + \gamma_1 z_{i,1}^* + \gamma_2 z_{i,2}^* + \gamma_3 z_{i,3}^*),$$

where $a_0 < b_0$ and where $H(u) = \exp(u)/\{1+\exp(u)\}$ is the logistic transform.

(Other cumulative distribution functions may be used as well, such as the normal, but this three-box model now becomes a natural generalisation of the logistic regression model for two boxes, which is why I tend to prefer this H.)

(b) Fit this model via maximum likelihood; give all parameter estimates, along with approximate standard deviations; and make an attempt at interpreting your findings.

(c) For the individual we considered in Exercise 2(f), estimate and display the curve $p_3(x_1)$, that person's probability of belonging to type 3, as a function of his age, as it ranges from 20 to 70. If you have time, include also a pointwise 90% confidence band for this curve.

(d) As for Exercise 2(b), fit each of the eight models 0, 1, 2, 3, 12, 13, 23, 123 corresponding to including or excluding z_1, z_2, z_3, and give a table with AIC, BIC and approximate Bayesian model probabilities. Comment on what you find.

(e) If you have time, consider one or two alternative or more general models, and check whether you succeed in increasing the AIC score beyond the best of the eight models considered so far.

Appendix: some useful R tricks

Here I list just a few potentially useful **R** programming details.

1. To read the LDL data into your **R** session, use e.g.
   `data <- matrix(scan("ldl-data", skip=14), byrow=T, ncol=7)`
2. To easily find parameter estimates in standard regression models, without necessarily programming the log-likelihood function etc., one may use
   `look <- glm(y ~ X + Z, family = gaussian)`
   followed by `look$coef`. Here `X` and `Z` may be matrices with several columns each.
3. `combinations` generates all subsets of an index set $\{1, \ldots, r\}$, so that e.g. `combinations(4)` returns the $2^4 = 16$ subsets of $\{1, 2, 3, 4\}$, indicated by 0's and 1's:
```
combinations = function(n)
{
comb = NULL
{for (i in 1:n)
comb = rbind(cbind(0,comb),cbind(1,comb))
return(comb)}
}
```
4. Applied FICology means having a practical grip also on the projection matrices π_S and the associated
$$Q_S = (\pi_S Q^{-1} \pi_S^{\mathrm{t}})^{-1} \quad \text{and} \quad G_S = \pi_S^{\mathrm{t}} Q_S \pi_S Q^{-1}.$$
   These may be constructed and saved one at a time, so to speak, in a brute force fashion, but may also perhaps more efficiently and conveniently be constructed on the go inside a `for` loop through all relevant subsets. In the following setup I have first used `subsets = combinations(qq)` or similar to generate the required subsets, and constructed `Id`, the identity matrix of dimension `qq`.
```
for (j in 2:nrow(subsets))
{
where <- (1:qq)[subsets[j, ] == 1]
dims <- length(where)
piS <- Id[where, ]
dim(piS) <- c(dims,qq)
QS <- solve(piS %*% Qinv %*% t(piS))
GS <- t(piS) %*% QS %*% piS %*% Qinv
# then further work here:
}
```
   The `for` loop in this setup starts with `j = 2` since I find it easiest to deal with the null model separately, corresponding to `j = 1`, for which `subsets[j, ] = c(0,0,0,0)` in this example.
5. To sort continuous data `yy` into appropriate boxes or windows, as e.g. required for Exercise 3, one may use the following:
```
N1 = 1*(yy <= 3.3)
N2 = 1*(yy > 3.3)*(yy <= 4.9)
N3 = 1*(yy > 4.9)
```