UNIVERSITETET I OSLO

Matematisk Institutt

Exam in:	**STK 4160/9160 – Model Selection and Model Averaging**
	Part I of two parts: The project
With:	**Nils Lid Hjort**
Time for exam:	**3.–15.vi.2015**

This is the exam project set for STK 4160/9160, spring semester 2015. It is made available on the course website as of *Wednesday 3 June 12:00*, and candidates must submit their written reports by *Monday 15 June 13:00* (or earlier), to the reception office at the Department of Mathematics, in duplicate. The supplementary four-hour no-book written examination take place *Thursday June 11* (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 relevant parts 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). They are graciously allowed not to despair should they 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 plus one exercises and comprises six pages. The first three exercises are for both the STK 4160 and STK 9160 students, whereas the PhD students taking the STK 9160 version of the course also should do Exercise four.

Exercise 1

"Every now and then a man's mind is stretched by a new idea or sensation, and never shrinks back to its former dimensions" (says Oliver Wendell Holmes), and perhaps you'll never again use the plain exponential model in the same way after having worked through the expo-stretching mechanisms of the present exercise.

(a) Consider a parametric model with probability density $f(y,\theta)$ and cumulative distribution function $F(y,\theta)$, for some appropriate θ of dimension say p. This model can then be stretched via a positive stretching parameter γ, in this fashion:

$$G(y,\theta,\gamma) = \Phi(\gamma\Phi^{-1}(F(y,\theta))) \quad \text{for } y > 0.$$

Here Φ is the cumulative standard normal distribution function, with inverse Φ^{-1} (called respectively `pnorm` and `qnorm` in R), and $\gamma = 1$ corresponds to leaving the original model intact, without stretching. Show that G indeed is a cumulative distribution function and give a formula for its density function.

(b) We shall now consider the particular case where the start model is the exponential one, with $f(y, \theta) = \theta \exp(-\theta y)$ and θ a positive parameter. Show that the density for this stretched exponential becomes

$$g(y, \theta, \gamma) = \theta \exp(-\theta y)\gamma \exp\big[\tfrac{1}{2}(1-\gamma^2)\{\Phi^{-1}(1-\exp(-\theta y))\}^2\big] \quad \text{for } y > 0,$$

and that the quantiles $G^{-1}(p)$ can be expressed as

$$\mu(p) = \mu(p, \theta, \gamma) = \frac{1}{\theta}\{-\log(1 - \Phi(\gamma^{-1}\Phi^{-1}(p)))\} \quad \text{for } p \in (0,1).$$

Comment in particular on the form of the median. For the value $\theta = 4.444$, compute and display in the same diagram the cumulative functions $G(y, \theta, \gamma)$ for a few values of γ, and comment.

(c) Next consider J, the Fisher information matrix for the two-parameter model, computed at the narrow model where $\gamma = 1$. Show that

$$J = \begin{pmatrix} 1/\theta^2 & k/\theta \\ k/\theta & 2 \end{pmatrix},$$

where k is a constant, with numerical value 0.5956. For a sample of independent observations, of size n, for what range of γ values can inference based on the exponential model be expected to be more precise than when using the fuller two-parameter model?

(d) We wish to estimate the quantile $\mu = G^{-1}(p)$, for a fixed p, and maximum likelihood gives rise to the estimators $\widehat{\mu}_{\rm narr}$ using the exponential model and $\widehat{\mu}_{\rm wide}$ with the two-parameter model. Give an explicit formula for the former and explain briefly how you can compute the latter from a given dataset. – Assume now that $\gamma = 1 + \delta/\sqrt{n}$, with associated true quantile $\mu_n = G^{-1}(p, \theta, 1 + \delta/\sqrt{n})$. In the notation used for some of the main results of Chs. 6-7 in Claeskens and Hjort (2008), we have

$$\sqrt{n}(\widehat{\mu}_{\rm narr} - \mu_n) \to_d \Lambda_{\rm narr} = \Lambda_0 + \omega\delta,$$
$$\sqrt{n}(\widehat{\mu}_{\rm wide} - \mu_n) \to_d \Lambda_{\rm wide} = \Lambda_0 + \omega(\delta - D),$$

where $\Lambda_0 \sim {\rm N}(0, \tau_0^2)$ and $D \sim {\rm N}(\delta, \kappa^2)$ are independent. Identify and find formulae for the quantities κ, τ_0, ω, for this quantile parameter $\mu = G^{-1}(p)$.

(e) Consider the limiting risk functions

$$r_{\rm narr}(\delta) = {\rm E}\,\Lambda_{\rm narr}^2 \quad \text{and} \quad r_{\rm wide}(\delta) = {\rm E}\,\Lambda_{\rm wide}^2.$$

Compute and display these, as a function of δ, for the case of the $\frac{1}{4}$-quantile $\mu = G^{-1}(\frac{1}{4})$, for $\theta = 4.444$. Comment on what you find.

(f) Now attempt to supplement the formulae and curves computed above for the limiting case with 'the real thing', namely

$$r_{n,\mathrm{narr}}(\delta) = \mathrm{E}_n\{\sqrt{n}(\widehat{\mu}_{\mathrm{narr}} - \mu_n)\}^2 \quad \text{and} \quad r_{n,\mathrm{wide}}(\delta) = \mathrm{E}_n\{\sqrt{n}(\widehat{\mu}_{\mathrm{wide}} - \mu_n)\}^2,$$

for finite n. Here E_n means expectation under the model where $\gamma = 1 + \delta/\sqrt{n}$. You would need to compute these two curves via simulation, where you for each δ simulate a high enough number of $\widehat{\mu}_{\mathrm{narr}}$ and $\widehat{\mu}_{\mathrm{wide}}$. Simulated datasets are most easily generated via the inverse cumulative distribution method (using that if H is any continuous cumulative distribution function, then $X = H^{-1}(U)$ has distribution H, when U is a uniform on the unit interval). Carry out this scheme for a couple of sample sizes n, and report briefly on your findings, along with one or at most two figures.

(g) Points (d)-(e)-(f) pertain to the two estimation methods that use either the narrow model or the wide model. Supplement your results and your risk functions with calculations (both in the limit experiment and for finite n, if you have the time) for the post-aic-estimator

$$\widehat{\mu}_{\mathrm{AIC}} = \begin{cases} \widehat{\mu}_{\mathrm{narr}} & \text{if AIC chooses the narrow model,} \\ \widehat{\mu}_{\mathrm{wide}} & \text{if AIC chooses the wide model.} \end{cases}$$

Comment on your findings.

Exercise 2

ANYTHING'S POSSIBLE if you've got enough nerve (says J.K. Rowling). The dataset `nervedata` is accessible from the course website, providing the time intervals between successive pulses along a certain nerve fibre, measured in seconds. For the purposes of this exercise these time intervals, say $y_1, \ldots, y_n$ with $n = 799$, are taken to be independent and identically distributed. We also treat this distribution as having a continuous density on the positive half-line, and ignore discretisation issues (the data are recorded to the level of centiseconds). Such data are typically assumed to follow an exponential distribution.

(a) Fit the data to the exponential model $\theta \exp(-\theta y)$ via maximum likelihood. Using this model, find the point estimate and an associated 95% confidence interval for $\mu = F^{-1}(\frac{1}{4})$, the 0.25 quantile (with $F(y)$ the cumulative distribution function with the exponential model).

(b) Then fit the data to the stretched exponential model $g(y, \theta, \gamma)$ of the previous exercise, again using maximum likelihood. Assuming that this two-parameter model is adequate, give approximate 95% confidence intervals for both parameters, and also for the 0.25 quantile $\mu = G^{-1}(\frac{1}{4})$. Plot the empirical distribution function along with the fitted parametric cumulative functions $F(y, \widehat{\theta}_{\mathrm{narr}})$ and $G(y, \widehat{\theta}, \widehat{\gamma})$. The empirical distribution function is $F_n(y) = n^{-1} \sum_{i=1}^n I\{y_i \le y\}$.

(c) Compute AIC and BIC scores for the narrow and the wide model here, and comment on what you find.

(d) We shall now consider an extension of the framework above, allowing further modelling flexibility. By first extending the exponential model to the Weibull, and then stretching this further using the same type of stretch mechanism, we reach a model with cumulative distribution function say

$$H(y,\theta,b,\gamma) = \Phi(\gamma\Phi^{-1}(1-\exp(-(\theta y)^b))) \quad \text{for } y>0,$$

with θ, b, γ being free positive parameters. Fit both the Weibull model (having $\gamma = 1$) and the full three-parameter model. Provide also confidence intervals for b and γ, compute AIC and BIC scores, and comment on your findings.

(e) The three-parameter model above contains the exponential model as a special case. For what range of values of (b,γ) around $(1,1)$ can the exponential model be expected to lead to more precise inference, for all estimands, than when using the three-parameter model?

Exercise 3

THE SAYING THAT BEAUTY IS BUT SKIN-DEEP is but a skin-deep saying. This exercise concerns modelling and analysis of survival data. You need to access the dataset `melanoma-data` at the course website, pertaining to $n = 205$ Danish patients with melanoma (a form of skin cancer), each of whom went through an operation. The data matrix consists of

$$(i, y_i, \delta_i, x_{1,i}, x_{2,i}, z_{1,i}, z_{2,i}, z_{3,i}) \quad \text{for } i = 1,\ldots,n,$$

with the first column simply being the running index $i = 1,\ldots,n$ identifying the patients. Here

- y_i is time to death or to censoring for patient i, after operation, in years;
- δ_i is 1 if y_i is time to death (non-censoring) and 0 in case of censoring;
- $x_{1,i}$ is gender, female 1 and male 2;
- $x_{2,i}$ is age at operation, in years;
- $z_{1,i}$ is thickness of melanoma, in mm;
- $z_{2,i}$ is infection level, with values 1, 2, 3, 4, where 1 is high and 4 is low resistance;
- $z_{3,i}$ is ulceration, with values 1 for yes and 2 for no.

The task is to model, analyse and understand how the covariates x_1, x_2, z_1, z_2, z_3 influence the chances of survival over time.

The study in question lasted for several years, and $n - \sum_{i=1}^n \delta_i = 148$ of the patients, those having $\delta_i = 0$, were luckily alive at the end of the study period (or left the study for other reasons). For a patient with $\delta_i = 0$, the time to death after operation is 'censored', the information hence being that this time to death after operation is longer than the recorded y_i. For the $\sum_{i=1}^n \delta_i = 57$ patients who died within the study period, however, y_i is non-censored.

Let $f_i(y)$ and $F_i(y)$ be the density and cumulative distribution function for the survival time after operation for individual i. The associated hazard rate function is

$$h_i(y) = \frac{f_i(y)}{1 - F_i(y)} \quad \text{for } y > 0,$$

and has the interpretation that $h_i(y)\,\mathrm{d}y$ is the probability of dying in the time interval $[y, y+\mathrm{d}y]$, given that the individual has survived up to time y. One may recover f_i and F_i from knowledge of the hazard rate h_i and cumulative hazard rate $H_i(y) = \int_0^y h_i(t)\,\mathrm{d}t$, via

$$f_i(y) = h_i(y)\exp\{-H_i(y)\} \quad \text{and} \quad F_i(y) = 1 - \exp\{-H_i(y)\} \quad \text{for } y > 0.$$

You do not need to demonstrate these formulae here, but this way of working with survival distributions via hazard rates is both conceptually important and mathematically convenient, and will be used below.

(a) Suppose a parametric model is put up for the hazard rates, say $h_i(y) = h_i(y, \theta)$, with cumulatives $H_i(y, \theta) = \int_0^y h_i(t, \theta)\,\mathrm{d}t$. Show that the log-likelihood information for individual i is

$$\begin{cases} \log h_i(y_i, \theta) - H_i(y_i, \theta) & \text{if } \delta_i = 1, \\ -H_i(y_i, \theta) & \text{if } \delta_i = 0. \end{cases}$$

Show from this again that the full log-likelihood function can be expressed as

$$\ell_n(\theta) = \sum_{i=1}^n \{\delta_i \log h_i(y_i, \theta) - H_i(y_i, \theta)\}.$$

(b) Write at the moment and for simplicity w_i for the covariate vector associated with individual i, say of length r (we shall soon return to the the specific covariates for the Danish patients). A simple but sometimes effective model for the hazard rates is that these are constant over time, with

$$h_i(y) = \exp(w_i^{\mathrm{t}}\theta) = \exp(w_{1,i}\theta_1 + \cdots + w_{r,i}\theta_r).$$

Show that the log-likelihood function may be written

$$\ell_n(\theta) = \sum_{i=1}^n \{\delta_i w_i^{\mathrm{t}}\theta - \exp(w_i^{\mathrm{t}}\theta) y_i\}.$$

Also give an expression for the $r \times r$ matrix of second order derivatives with respect to θ.

(c) For the utterly simple model which takes $h_i(y) = \exp(\beta_0)$, constant over time and across all individuals, find the maximum likelihood estimator and its numerical value for the Danish dataset. This indicates in particular that a sensible numerical start value for the parameter corresponding to β_0 in various models below, when using `nlm` or other numerical algorithms, might be -3 or similar.

(d) For the Danish dataset, with five covariates x_1, x_2, z_1, z_2, z_3, consider the six-parameter model where the hazard rates are constant over time, but differing from patient to patient, as

$$h_i(y) = \exp(\beta_0 + x_{1,i}\beta_1 + x_{2,i}\beta_2 + z_{1,i}\gamma_1 + z_{2,i}\gamma_2 + z_{3,i}\gamma_3) \quad \text{for } i = 1, \ldots, n.$$

Fit this model to the data. Give estimates along with standard errors (estimates of standard deviations) for the six parameters, and comment on these.

(e) We now take covariates x_1 (gender) and x_2 (age at operation) as 'protected' but z_1, z_2, z_3 as 'open'. Consider the submodels corresponding to pushing these three open covariates in and out. Fit these candidate models, compute AIC and BIC scores, and comment on what you find. Also compute approximate model probabilities $\Pr(\text{model} \,|\, \text{data})$, explaining the underlying assumptions for such calculations.

(f) Now focus on a given patient, with covariates equal to say $(x_{1,0}, x_{2,0}, z_{1,0}, z_{2,0}, z_{3,0})$, where we wish to estimate his or her hazard rate with the best precision. Specifically, for patients

A: woman, age 50, average melanoma thickness, infection level 3, with ulceration,

B: man, age 60, average melanoma thickness, infection level 4, no ulceration,

provide estimates of their hazard rates for each candidate model, along with FIC scores and a FIC plot. Comment briefly on your findings.

(g) The exponential models for constant hazard rates used above are not necessarily good enough for all purposes. An extension of the full six-parameter model above takes

$$h_i(y) = \exp(\beta_0 + x_{1,i}\beta_1 + x_{2,i}\beta_2 + z_{1,i}\gamma_1 + z_{2,i}\gamma_2 + z_{3,i}\gamma_3)\kappa y^{\kappa-1} \quad \text{for } y > 0,$$

with κ a positive parameter. Fit this seven-parameter model to the data and comment on what you find. Briefly discuss how your FIC analysis would need to be modified to accommodate this extra parameter (and if you have the time, carry out such an analysis, for the two patients met above).

(h) Choose one of the two patients encountered above, and consider that person's full survival curve $\Pr(T \geq y \,|\, x_{1,0}, x_{2,0}, z_{1,0}, z_{2,0}, z_{3,0})$, with T denoting the time to death after operation. Estimate and display this curve, along with an approximate 90% pointwise confidence band, first for the six-parameter and then for the seven-parameter model.

Exercise 4: for the PhD students taking the STK 9160 exam

A recent article by Martin Jullum and Nils Lid Hjort, *Parametric or nonparametric: the FIC approach* (June 2015) has been uploaded to the course website. Give a short summary of some of the methods developed in that paper, and apply them to the analysis of a dataset of your own choice, along with a brief discussion of your findings.