UNIVERSITETET I OSLO

Matematisk Institutt

Exam in:	**STK 4160/9160:**
	Model Selection and Model Averaging
	Part II of two parts: The project
With:	**Nils Lid Hjort**
Time for exam:	**5.–17.vi.2019**

This is the exam project set for STK 4160/9160, spring semester 2019. It is made available on the course website as of *Wednesday 5 June 11:11*, and candidates must submit their written reports by *Monday 17 June 14:28* (or earlier), to the Inspera System at the Department of Mathematics. The supplementary four-hour no-book written examination took place *Tuesday June 4*. 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 student-web identification number 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, by handing in your report to the Inspera system you guarantee that you've read, understood, and confirmed the points of the *self-declaration form*; see the last page of this document. Also, your report should contain *one separate extra page*, the student's one-page summary of the exam project report, which should briefly tell its readers about how the work has proceeded, and also contain a brief self-assessment of its quality. You may make this the very last page of your report.

This exam set contains two exercises and comprises as many as nine pages (though there are various figures, and the two last pages are an appendix with R details and the self-declaration form).

Exercise 1

Itj fårrå nålles. Too many people die in car accidents, still, though luckily the accident rates go down in most countries. You need to access the dataset `sweden-accidents` from the course website, which I've found via a Swedish colleague and then amended for the purposes of the present exam project; the data originate from official Swedish police records. It consists of the columns

$$\texttt{year}, \texttt{x1}, \texttt{x2}, \texttt{x3}, \texttt{y},$$

for years 1955 to 2010, with y the number of traffic deaths for each of these years (counting both people killed in cars and by cars), along with covariates

– x_1, which is simply the year minus 1950, making some of the numerical computations simpler;

– x_2, traffic volume, measured as the total number of cars owned in Sweden, in millions;

– x_3, amount of petrol (bensin) sold per year, measured in billions of litres.

There are several models one might use to analyse how traffic volume, petrol use, and also calendar time itself influence the fatalities statistics (cars have become safer, etc.), but for the present occasion we shall primarily use Poisson regression models.

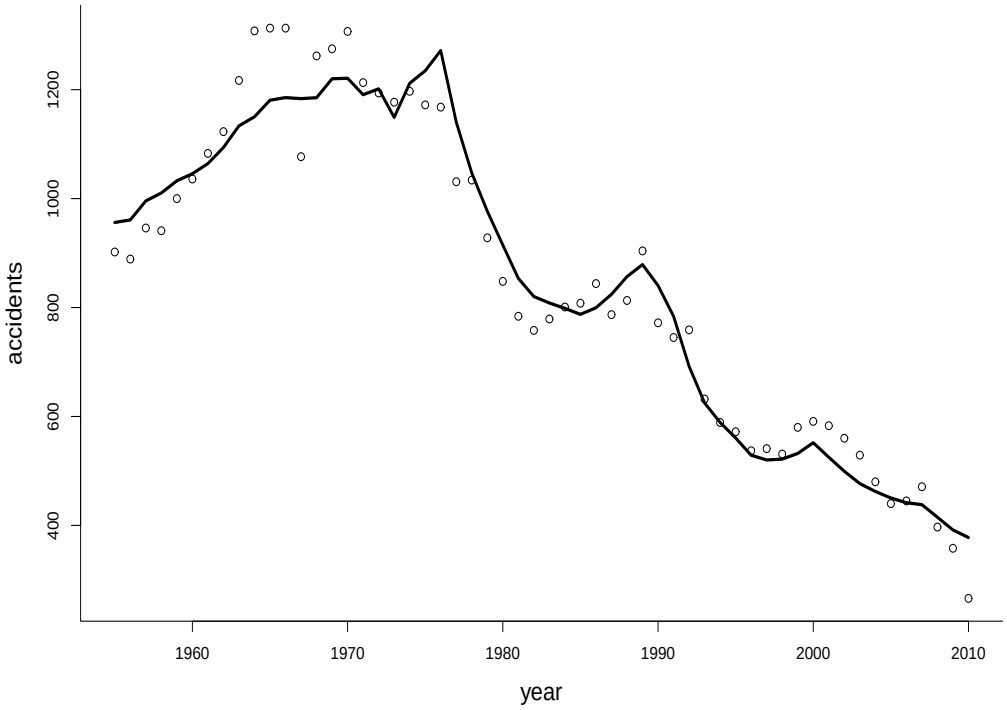

Figure A: The number of car accident fatalities in Sweden, 1955 to 2010, along with the fitted curve from the Poisson regression model with covariates x_1, x_2, x_3.

(a) Suppose in general terms that Y_i has the Poisson distribution with parameter $\xi_i = \exp(x_i^{\mathrm{t}}\beta)$, involving a covariate vector $x_i = (x_{i,1}, \ldots, x_{i,p})$ of length p, and that the n count observations $Y_1, \ldots, Y_n$ are independent given all the covariate information. Write down the log-likelihood function and its two first derivatives. Show that minus the normalised Hesse matrix becomes

$$J_n(\beta) = -\frac{1}{n}\frac{\partial^2 \ell_n(\beta)}{\partial\beta\partial\beta^{\mathrm{t}}} = \frac{1}{n}\sum_{i=1}^{n} \xi_i x_i x_i^{\mathrm{t}}.$$

(b) For the Swedish data we start with the simplest Poisson regression model, say M_1, using

$$\xi_i = \exp(\beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \beta_3 x_{i,3}).$$

Find the maximum likelihood estimates, give standard errors (estimated standard deviations) for these, and reproduce a version of Figure A. You may programme the log-likelihood function from scratch, or use something along the lines of
`summary( glm(y ~ x1 + x2 + x3, family=poisson) )`
in R.

(c) Give a plot of the estimated residuals $r_i = (y_i - \widehat{\xi}_i)/\widehat{\xi}_i^{1/2}$, and comment on what you might learn from this.

(d) There are reasons for not being entirely happy with this start model M_1 (which leads us to building other models in a little while). Find an expression for the matrix

$$K_n(\beta) = \frac{1}{n}\sum_{i=1}^{n} u(y_i, x_i, \beta)u(y_i, x_i, \beta)^{\mathrm{t}},$$

with $u(y_i, x_i, \beta)$ the score function associated with datum y_i. Compute both $J_n(\widehat{\beta})$ and $K_n(\widehat{\beta})$, and comment.

(e) Introduce the extra covariates

$$z_2 = w_2^2, \quad z_3 = w_2^3, \quad z_4 = w_2^4,$$

where $w_2 = (x_2 - \bar{x}_2)/\mathrm{sd}(x_2)$ is the normalised version of covariate x_2; working with w_2 here instead of x_2 aids interpretation of regression coefficients and makes numerical methods more stable. Find the AIC values for models M_1 (the one above, with x_1, x_2, x_3); M_2, which uses x_1, x_2, x_3, z_2; M_3, which uses x_1, x_2, x_3, z_2, z_3; M_4, which uses $x_1, x_2, x_3, z_2, z_3, z_4$. Find similarly the model robust versions of AIC scores. What are your conclusions, so far?

(f) In addition to the AIC and AIC-robust scores dealt with above, you are now asked to carry out cross-validation, for the four models. For model M_1, for example, your task is to compute both

$$\mathrm{xv}_{1,a} = \frac{1}{n}\sum_{i=1}^{n} \log f_1(y_i, x_i, \widehat{\beta}_{1,(i)}) \quad \text{and} \quad \mathrm{xv}_{1,b} = \frac{1}{n}\sum_{i=1}^{n} |y_i - \widehat{\xi}_{1,i,(i)}|,$$

in slightly laborious notation. Here f_1 denotes the model density for model M_1, with $\widehat{\beta}_{1,(i)}$ the maximum likelihood estimator for the model parameter vector of this model M_1, obtained by pushing out (x_i, y_i) from the data, leading to the estimated ξ_i via $\widehat{\xi}_{1,i,(i)}$, etc. Comment on what you find using these two cross-validation schemes.

(g) Using the available accident data for years up to 2010, we shall now attempt to estimate ξ_{n+1}, the expected number of fatalities for the year 2011. Set values of covariates x_2 and x_3 for 2011 equal to those for 2010, making

$$\xi_{n+1} = \mathrm{E}\,\{Y_{n+1} \,|\, (x_{n+1,1}, x_{n+1,2}, x_{n+1,3})\}$$

well-defined. Estimate this ξ_{n+1} using models M_1, M_2, M_3, M_4. Carry out FIC analysis to compare and rank these four candidate models, using M_1 as the 'narrow model' and M_4 as the 'wide model'. Compare your estimates with the actual number Y_{n+1} of traffic fatalities in Sweden 2011, which you should be able to find on the internet.

(h) The count data here might have more variability than that dictated by the Poisson model. One way to model such a potential aspect of data is as follows. First, I write $\xi \sim \text{Gamma}(a, b)$ to indicate that ξ has the Gamma distribution with density $\{b^a/\Gamma(a)\}\xi^{a-1}\exp(-b\xi)$, with mean a/b and variance a/b^2. In general terms, with a p-dimensional covariate vector x_i, suppose now that

$$Y_i \,|\, \xi_i \sim \text{pois}(\xi_i) \quad \text{whereas} \quad \xi_i \sim \text{Gamma}(c\exp(x_i^{\mathrm{t}}\beta), c).$$

In particular, a high value of c corresponds to the previous situation, where Y_i is close to a Poisson. Show that

$$\mathrm{E}\, Y_i = \exp(x_i^{\mathrm{t}}\beta) \quad \text{and} \quad \mathrm{Var}\, Y_i = \exp(x_i^{\mathrm{t}}\beta)(1 + 1/c).$$

Find an expression for the log-likelihood function $\ell_n(\beta, c)$.

(i) Consider the model $M_{1,\text{plus}}$, which has the same covariates x_1, x_2, x_3 as for model M_1, but has the extra gamma-mixture $\xi_i \sim \text{Gamma}(c\exp(x_i^{\mathrm{t}}\beta), c)$ as in the above description. Find maximum likelihood estimates $(\widetilde{\beta}, \widetilde{c})$, along with the associated maximal value and the AIC scores. Comment on what you find.

(j) Via models and methods in this exercise we've been able to describe and interpret certain aspects of the Swedish traffic deaths statistics, 1955 to 2010, and their relation to basic traffic volume numbers. Do you have further suggestions, regarding either modelling or analysis of other data of relevance?

Exercise 2

YOU CANNOT STEP INTO THE SAME RIVER TWICE, and Heraclitus would also have agreed that skating the 10000 metres is not quite the same as skating the 5000 metres twice. These are the results for the 5k and 10k times from the World Allround Championship 2019, held March 2-3 in Calgary:

hero	5k	10k
1 Patrick Roest	6:08.27	12:51.17
2 Sverre Lunde Pedersen	6:10.10	12:56.91
3 Sven Kramer	6:08.83	13:00.93
4 Douwe de Vries	6:12.72	13:01.44
5 Ted Jan Bloemen	6:13.20	12:53.15
6 Sindre Henriksen	6:26.64	13:30.71
7 Danila Semerikov	6:13.75	13:18.92
8 Haralds Silovs	6:24.80	13:54.14

In this exercise we shall take an interest in how the 10k pans out compared to the 5k. For each skater, i.e. for those eight allowed to skate both the 5k (on the Saturday) and the 10k (on the Sunday), we may record

$$y_j = \frac{\text{time on the 10k for skater } j}{\text{time on the 5k for skater } j} \quad \text{for } j = 1, \ldots, 8.$$

For Pedersen, for example, this was $(12 \cdot 60 + 56.91)/(6 \cdot 60 + 10.10) = 2.099$. Clearly this ratio varies among types of skaters (some are long-distance athletes, others prefer the shorter distances), perhaps with the ice and weather conditions (rainy, windy outdoor Amsterdam 2018 is not the same as Calgary indoor Olympic hall 2019), and over time (top skaters are generally better in 2020 than in 2000, also in terms of equipment).

I have gone to the trouble of collecting such (5000 m, 10000 m) data, for all skaters in question, for the twenty World Allround Championships from 2000 Milwaukee and 2001 Budapest up to 2018 Amsterdam and 2019 Calgary (and yes & indeed, I've been to a few of these, qua journalist for *SpeedSkating World*). You need to access the files `fivektenk-results` and `fivektenk-data` from the course website. The first gives all results, in the format indicated above, with names and result times in the customary format; the second file is the one to use to access the data in a simple fashion, and to crank out the required

$$y_{i,j} = \frac{\text{time on the 10k for skater } j \text{ in event no. } i}{\text{time on the 5k for skater } j \text{ in event no. } i} \quad \text{for } i = 1, \ldots, k,\ j = 1, \ldots, m_i.$$

Here $k = 20$ is the number of events, and m_i is currently kept as low as 8 (to the consternation of many skaters and fans), though the number of skaters entering the fourth distance was $m_i = 12$ up to 2012 (and always equal to 16, with eight pairs, in the more distant past).

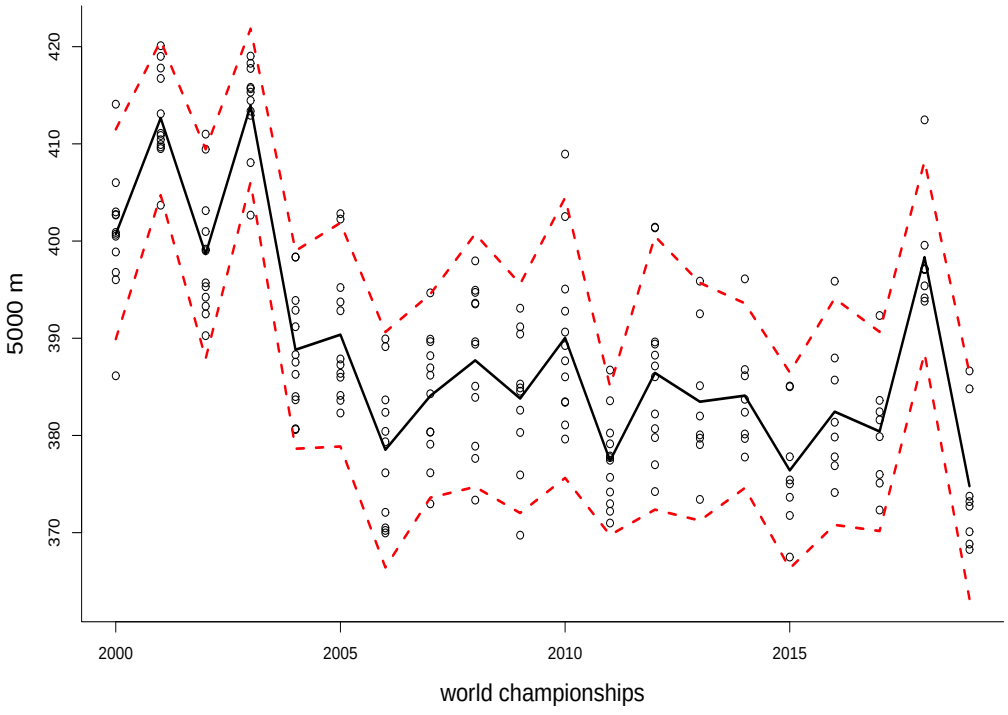

Figure B: 5000 m results (here given in seconds, so 6:20 is 380 seconds, etc.), for the skaters with four distances, World Allround Championships 2000 to 2019. The solid line is the empirical mean, per event, and the band is a 90% confidence interval for the mean parameter, per event.

These annual World Allround Championships are the most important events of the season, for the skaters and the zillions of fans (apart perhaps from the Olympics, in such years). The athletes skate 500 m and 5000 m on the Saturday, then the 1500 m and 10000 m on the Sunday, with the final ranking based on the accumulated point-sum, measured in 500 m times, i.e. $x = x_1 + x_2/10 + x_3/3 + x_4/20$, with x_1, x_2, x_3, x_4 the times clocked over the four distances. Only the very best of the best are allowed to skate the fourth distance. There is an ongoing active debate on how many pairs should be allowed to do the fourth distance, also as a part of a more general discussion, related to how to make the glorious 10000 m appeal to a broader public, and how to save the allround event from being pushed away by single-distance championships. The investigation in this exercise is a small part of this general discussion, related also to how 'compact' and how variable the $y_{i,j}$ ratios are (if all such ratios are pretty close to some given number, from year to year, the 10k is arguably 'too predictable', etc.).

(a) Consider Figure B, where I've plotted the 5000 m results for all $n = \sum_{i=1}^{k} m_i = 208$ skaters, for the $k = 20$ events, along with the curve of means, and the 90% confidence band for the underlying means, assuming normality. Construct a similar figure for the 10000 m results. (I do such things in R by first using `plot`, to get the points, and then adding curves via `matlines`. There are several other ways.)

(b) Then construct a similar figure for the 10k/5k ratios $y_{i,j}$, the observations under scrutiny in this project. For the 10k in Amsterdam 2018 I've included heroic Pedersen, despite his fall, and also Semerikov's 'winning' time, since I disagree with ISU's decision to disqualify him; for the further analysis you might wish to remove Søndrål from the 2000 data, however, since, as we recall, he fell and lost many seconds.

(c) For this and the next point, we work with the model where all observations are taken independent, and with $y_{i,j} \sim \mathrm{N}(\xi_i, \sigma_i^2)$ for the skaters from event i. At the outset there are therefore $2k = 40$ parameters. Before we come to ways of modelling these 20 + 20 parameters, compute maximum likelihood estimates for all $\widehat{\xi}_i$ and $\widehat{\sigma}_i$, and display these in two appropriate figures, both ornated with 90% confidence intervals.

(d) You should now work through the following list of as many as six candidate models, regarding how to view these ξ_i and σ_i parameters. For each, find and list parameter estimates and their model-based standard errors, and compute AIC values. Comment on what you learn from this. For models M_4 and M_5, use $w_i = (t_i - \bar{t})/\mathrm{sd}(t)$, the normalised version of $t_i = \mathrm{year}_i$; this aids interpretation and avoids potential numerical difficulties.

  (i) Model M_1: take all ξ_i to be equal to a common ξ, and all σ_i equal to a common σ.
  (ii) Model M_2: take all σ_i equal to a common σ, but let the ξ_i be different.
  (iii) Model M_3: take all ξ_i equal to a common ξ, but let the σ_i be different.
  (iv) Model M_4: take $\xi_i = a + bw_i$, with all the σ_i equal to a common σ.
  (v) Model M_5: take $\xi_i = a + bw_i$ and $\sigma_i = \sigma_0 \exp(\gamma w_i)$.
  (vi) Model M_6: with no restrictions at all, allowing the $2k$ parameters to be free.

(e) Then there is a seventh model I wish you to consider and work through, with a certain different viewpoint for the data. Instead of thinking that data and parameters from year to year are more or less unrelated, we view the ξ_i as stemming from a background population of such mean values. We take $\xi_i \sim \mathrm{N}(\xi_0, \tau_0^2)$, or in other words and symbols $\xi_i = \xi_0 + \delta_i$, where the δ_i are from $\mathrm{N}(0, \tau_0^2)$. This leads to

$$y_{i,j} = \xi_0 + \delta_i + \varepsilon_{i,j} \quad \text{for } i = 1, \ldots, k,\ j = 1, \ldots, m_i,$$

with the $k + n$ variables being independent. Show that the m_i-vector y_i, with components $y_{i,j}$, becomes a multinormal,

$$y_i \sim \mathrm{N}_{m_i}(\xi_0 \mathbf{1}_i, \Sigma_i),$$

with

$$\Sigma_i = \begin{pmatrix} \tau_0^2 + \sigma^2 & \cdots & \tau_0^2 \\ \vdots & \cdots & \vdots \\ \tau_0^2 & \cdots & \tau_0^2 + \sigma^2 \end{pmatrix} = \tau_0^2 \mathbf{1}_i \mathbf{1}_i^{\mathrm{t}} + \sigma^2 I_{m_i},$$

where $\mathbf{1}_i$ is the m_i-vector of $(1, \ldots, 1)$ and I_{m_i} the identity matrix. Find parameter estimates for (ξ_0, τ_0, σ), along with the AIC score, and comment.

(f) On Sunday March 1, 2020, at Vikingskipet in Hamar, eight fabulous skaters will be ready for the 10000 m (after having raced well enough on the three first distances), among them Pedersen, Roest, Kramer (check `youtube.com/watch?v=dmm_k5V3DcU`). What will their y_j ratios look like, based on your analyses?

Appendix: Some R tricks

Reading data from my data files into your computer: For the Swedish accident data, you may use

```
sweden <- matrix(scan("sweden-accidents",skip=10),byrow=T,ncol=5)
```

For the skaters data, you may use

```
skaters <- matrix(scan("fivektenk-data",skip=5),byrow=T,ncol=8)
```

Then you may read off 5k and 10k times, as well as other required statistics, in the following fashion, or with similar tricks:

```
year <- skaters[ ,1]
fivek <- 60*skaters[ ,3] + skaters[ ,4] + skaters[ ,5]/100
tenk <- 60*skaters[ ,6] + skaters[ ,7] + skaters[ ,8]/100
yy <- tenk/fivek
events <- 2000:2019
mm <- 0*events
mean5k <- 0*events
kk <- length(events)
for (j in 1:kk)
{
now5k <- fivek[year==events[j]]
mean5k[j] <- mean(now5k)
mm[j] <- length(now5k)
}
```

A way to programme log-likelihood functions, for the 20-event skating data: Here I illustrate for a too simple model, the one having $y_{i,j} \sim \mathrm{N}(\xi_i, 1)$, and for that model you actually don't need a log-likelihood programme. The scheme may be amended to work for more complicated models, however.

```
aux11 <- 0*(1:kk) # loglik bits, to be added up
logL11 <- function(para)
{
xi <- para[1:kk]
for (i in 1:kk)
{
now <- yy[year==events[i]]
aux11[i] <- -0.5*sum((now-xi[i])^2)- 0.5*mm[i]*log(2*pi)
}
sum(aux11)
}
starthere11 <- c(mean(yy)+0*(1:kk))
logL11(starthere11) # to check that things work
minuslogL11 <- function(para)
{-logL11(para)}
nils11 <- nlm(minuslogL11,starthere11,hessian=T)
```

after which you may read off maximum likelihood estimates, the Hesse matrix, etc.

Handing in your exam project report to the Inspera system is taken as guarantee that you've also read, understood, and confirmed the following points, regarding your work:

Egenerklæring / Self declaration:

Jeg erklærer herved at min hjemmeeksamensprosjektrapport, som er levert for kurset

STK 4160 eller **STK 9160**

ved Matematisk institutt, Universitetet i Oslo,

1. ikke har vært brukt til en annen eksamen ved et annet institutt eller universitet eller høyskole, innenlands eller utenlands;
2. ikke refererer til andres arbeid uten at dette er oppgitt;
3. ikke refererer til eget tidligere arbeid uten at dette er oppgitt;
4. ikke er forfattet av Juksemaker Pipelort;
5. har oppgitt alle referanser i litteraturlisten;
6. ikke er et samarbeid med en eller flere andre.

Jeg er kjent med at brudd på disse bestemmelsene er å betrakte som fusk, og at dette strider mot universitetets reglement.

Oslo, juni 2019