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:	**7.–19.vi.2017**

This is the exam project set for STK 4160/9160, spring semester 2017. It is made available on the course website as of *Wednesday 7 June 11:11*, and candidates must submit their written reports by *Monday 19 June 11:59* (or earlier), to the reception office at the Department of Mathematics, *in duplicate*. The supplementary four-hour no-book written examination took place *Friday June 2* (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 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, 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 two plus one exercises and comprises seven pages. The *first two 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 three.

Exercise 1

Polynomials have many names (yes, that was a joke). In this exercise we shall work with certain log-polynomial models for densities on the unit interval $[0, 1]$.

The data used to illustrate some of the methods are $n = 901$ so-called mhaq measurements; these relate to certain scores from modified health assessment questionnaires, from patients at the Division for Women and Children at Oslo University Hospital at Ullevål. The original scale for these scores is from 0.00 (the patient is essentially close to pain-free) to 3.00 (the patient has extensive pain, and has trouble accomplishing various tasks); here they have for convenience been transformed to the continuous scale $[0, 1]$, however. The

data-file is available at the course website under the name `mhaq-data`. We treat these $n = 901$ observations as being independent and drawn from the same continuous density on $[0, 1]$.

(a) Suppose $y_1, \ldots, y_n$ are independent observations from the same density on $[0, 1]$. A simple model for such data is of the form

$$f(y, \theta_1, \theta_2) = (1/k) \exp(\theta_1 y + \theta_2 y^2),$$

with k the required normalisation constant. It is mathematically practical to write this as

$$f(y, \theta_1, \theta_2) = \exp\{\theta_1 y + \theta_2 y^2 - c(\theta_1, \theta_2)\}.$$

Show that then

$$c(\theta_1, \theta_2) = \log\Big\{\int_0^1 \exp(\theta_1 y + \theta_2 y^2)\, \mathrm{d}y\Big\}.$$

(b) For this model, find expressions for the score functions (the log-derivative with respect to θ_1 and θ_2) and the Fisher information matrix $J(\theta_1, \theta_2)$.

(c) For this two-parameter model, write down the log-likelihood function, and fit the model to the $n = 901$ mhaq observations, via maximum likelihood. For this you need both numerical integration and numerical maximisation, and some of the ingredients given in this exam set's Appendix might be useful for you.

(d) Assuming first that the two-parameter model is correct, give approximate 95% confidence intervals for θ_1 and θ_2, based on theory developed in the course. Similarly give approximate 95% confidence intervals for the density itself, at positions $y = 0.25, 0.50, 0.75$.

(e) Then give such approximate 95% confidence intervals, for the two parameters and for the density value at positions $0.25, 0.50, 0.75$, without assuming that the model is correct.

(f) Then fit the data to log-polynomial models of order three and four, using the same techniques. Also compute the maximised values of the corresponding log-likelihood functions of order two, three, four, say $\ell_{2,\max}, \ell_{3,\max}, \ell_{4,\max}$. Use these to compute AIC values for these three models. Construct a figure similar to my Figure 1 here. Which of the three fitted densities would you say does the best job?

(g) Explain how you would go about selecting the best of these log-polynomial models, of order two, three, four, for the purpose of estimating the density at position $y_0 = 0.33$. Carry out this scheme.

(h) Look for Karl Weierstraß's approximation theorem (from about 1885), on your bookshelf or on the world's wild web. Explain its significance for log-polynomial modelling of densities on the unit interval.

(i) Suggest other parametric models for these $n = 901$ mhaq data; and if you find time, fit one or more of these models and see if one of your suggestions does better than the log-polynomial models of order two, three, four pointed to above. Should you be tempted to explore higher polynomial orders, it may be useful for numerical stability and optimisation to go from $y, y^2, y^3, \ldots$ to $T_1(y), T_2(y), T_3(y), \ldots$, with $T_j(y) = 2^j(y - \frac{1}{2})^j$. (Fitting models with the $T_j(y)$ instead of the y^j produces the same density estimates and the same AIC values.)

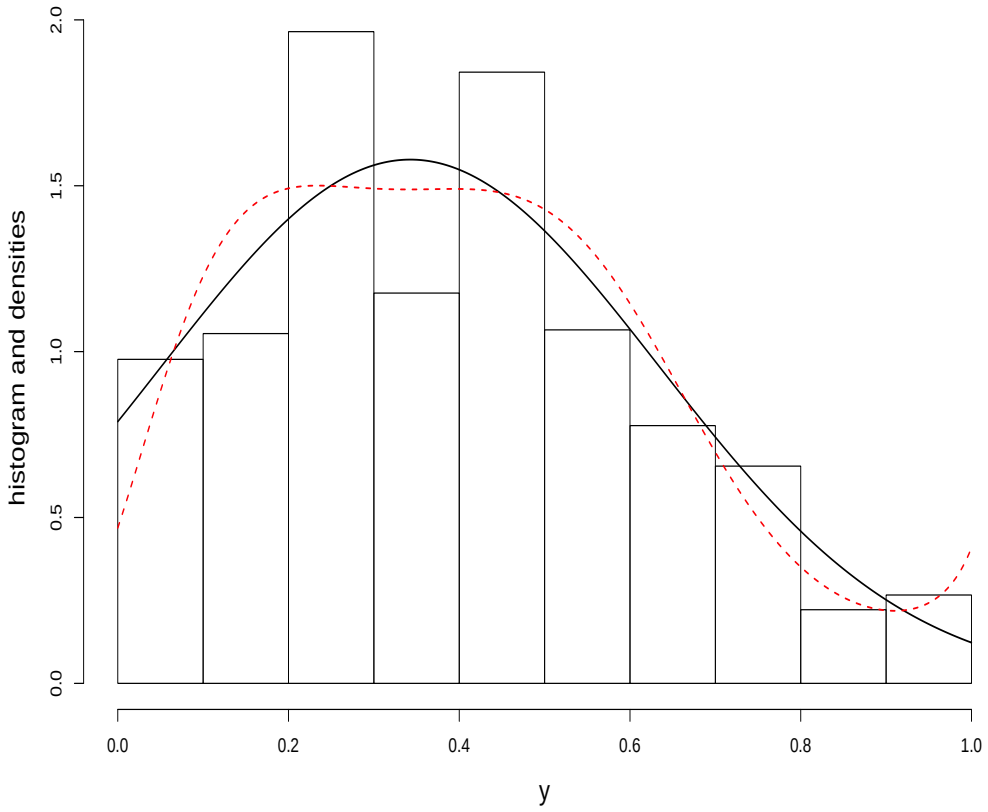

Figure 1: Histogram for the 901 mhaq measurements, along with fitted log-polynomial density of order two, and one more fitted density from another model.

Exercise 2

Многая лета! And such prayers appear to be heard, as witnessed in the table below, giving expected life-lengths for women and men, born in and living in the Nordic countries, from 1960 to 2015. We do not go into the precise construction and calculation of these numbers here, but their interpretation is indeed built on the 'expected life-length' concept. Boys born in Sweden in the year 2015, for example, will, if Sweden stays about the same over the next hundred years, have life-lengths following a certain distribution (with all relevant components and quantities estimated in 2015, not ninety years later, with good precision), and the mean of this distribution is 80.7. (I have manually extracted these data from `www.worldlifeexpectancy.com/history-of-life-expectancy`.)

	1960	1970	1980	1990	2000	2010	2015
women:							
Norway	75.9	77.3	79.2	79.8	81.3	83.1	83.7
Sweden	74.9	77.2	78.9	80.4	81.9	83.5	84.0
Denmark	74.0	75.9	77.2	77.7	79.2	81.3	82.5
Finland	72.4	74.4	77.9	78.9	80.9	83.1	83.8
men:							
Norway	71.3	71.0	72.3	73.4	75.7	78.8	79.8
Sweden	71.2	72.2	72.8	74.8	77.3	79.5	80.7
Denmark	70.4	70.9	71.2	72.0	74.5	77.1	78.6
Finland	65.4	66.2	69.2	70.9	74.0	76.6	78.3

Figure 2 displays plots of these data for the women. We shall at the outset treat these data as four plus four independent linear regressions, with the model

$$y_i = a + b\,\text{year}_i + \sigma\varepsilon_i \quad \text{for } i = 1, \ldots, 7$$

being in operation for each, and with the ε_i seen as independent and standard normal. Thus for each country, and for men and for women, there is such a regression structure, dictated by parameters (a, b, σ). In total, therefore, there are as many as 24 different parameters at work behind the data displayed in the table. We shall however search for meaningful submodels with fewer effective parameters. For concreteness, most of the analyses below focus on the women, where we hope to find good models with less than twelve parameters.

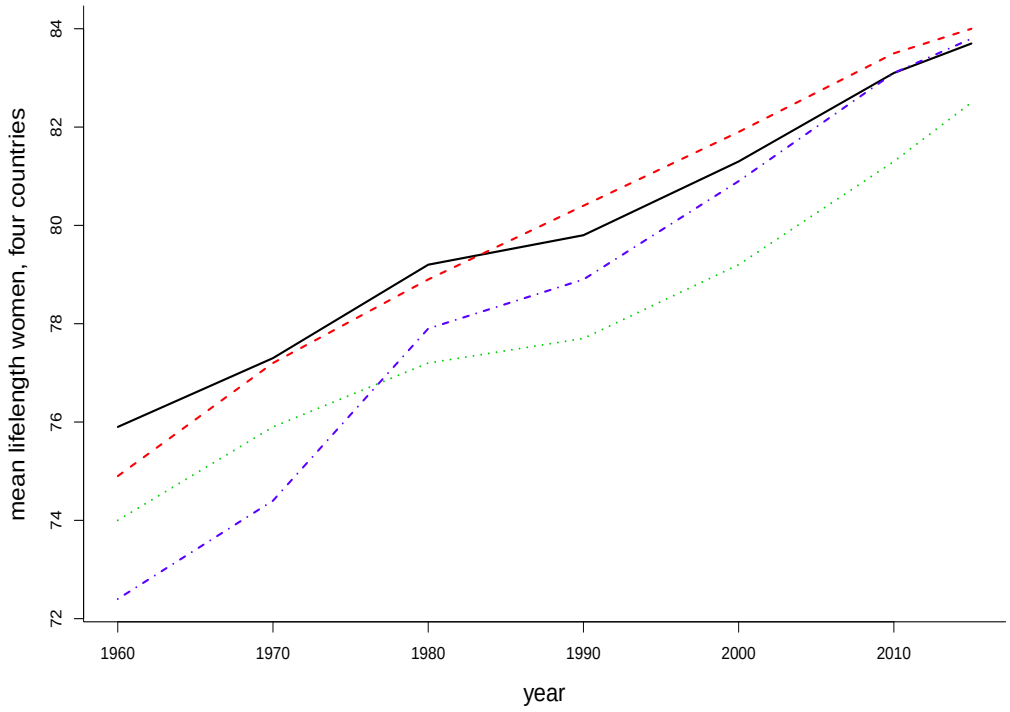

Figure 2: Evolution of mean life-lengths for women, from 1960 to 2015, for the four Nordic countries.

(a) For numerical stability reasons it is helpful to write the regressions as

$$y_i = a + bx_i + \sigma\varepsilon_i \quad \text{for } i = 1, \ldots, n,$$

with $x_i = \text{year} - 1960$ for the seven time-points. For one of these regressions, say for Swedish women, write down the log-likelihood function, and explain how the consequent maximum likelihood (ML) estimates $(\widehat{a}, \widehat{b}, \widehat{\sigma})$ may be found. Give numerical values for the parameter estimates, for the Swedish women, and also 90% confidence intervals for both the b and the σ parameter. (Parameter estimates for the regression line may be found using `lm(y ~ x)` or `glm(y ~ x)` in R; see the Appendix.)

(b) Show also that a formula for the maximised log-likelihood function is

$$\ell_{\max} = -n \log \widehat{\sigma} - \tfrac{1}{2} n - \tfrac{1}{2} n \log(2\pi).$$

For the Swedish women, give the numerical value for the maximised log-likelihood, and also for the AIC score (for that partial dataset, and for that model).

(c) Consider the data for the women, for the four countries. For the biggest of these models, the one using three regression model parameters for each of the four countries, compute the maximised log-likelihood and the AIC score.

(d) For initial comparison, study the smallest of all relevant models for the women, the one employing the same (a, b, σ) across all four countries. Compute the maximised log-likelihood value and the AIC score. Comment on this initial finding.

(e) When analysing the full dataset for women, across four countries, there are many submodels worth considering. For the present purposes we single out the following.

M_1 The biggest model, with different (a, b, σ) for each of the four countries.

M_2 The model having equal b for the four countries, and otherwise varying a and σ parameters.

M_3 The model having equal σ for the four countries, and otherwise varying a and b parameters.

M_4 The model lumping together Norway and Sweden, with a common (a, b, σ) for these two, and then separate regressions for Denmark and Finland.

For each of these four models, compute and quote the relevant parameter estimates, the maximised log-likelihood values, and the AIC scores. What do you conclude, so far?

(f) If you wish to, invent and present *one more model* for the women data, and check if it does better than the best of the above four models, in terms of AIC scores.

(g) Use first the data for the Norwegian women only, and then the Norwegian men only, to give estimates of the mean life-lengths of women and men born in Norway in the year 2025. For the women, supplement your estimate with estimates based on what you find to be a good or perhaps the best model, among those worked with in point (e).

(h) For concreteness and simplicity we have focused on model building and analyses for the women part of the data above. Similar efforts can of course be invested into the data for the men. Due to the practical limitations of exam time, etc., we do not go down that road here and now. But suggest *one or two models* that relate the women and men data sets, in an effort to reduce the number of free parameters in the eight regressions from 24 to something smaller. Explain what could be reasonable goals with such model building efforts, and how you might go about selecting among these models.

(i) For a model of your own choice, among the many regression models under consideration, provide a 90% prediction interval for Y_{2025}, the not-yet-seen mean life-length for Norwegian women and for Norwegian men, both in the year 2025.

Exercise 3: for the PhD students taking the STK 9160 exam

An article by Bruce E. Hansen, *Challenges for Econometric Model Selection*, published in *Econometric Theory* in 2005, has been uploaded to the course website. Write up two-three pages summarising methods and viewpoints discussed in the paper, supplemented with views of your own. In particular, explain if there are points made by Hansen you might not agree with, or where you have additional insights or points of criticism. If you have the time, do not hesitate to illustrate methods or points using datasets you might find yourself, or on simulated data.

Appendix: Some R tricks

Defining some functions:

```
T1 <- function(y)
{y}
T2 <- function(y)
{ y^2 }

cc2 <- function(theta)
{
inte2 <- function(y)
{ exp(theta[1]*T1(y) + theta[2]*T2(y)) }
log( integrate(inte2,0,1)$value )
}
```

Maximising a log-likelihood function:

```
logL2 <- function(theta)
{
hei <- theta[1]*T1(yy) + theta[2]*T2(yy) - cc2(theta)
sum(hei)
}

minuslogL2 <- function(theta)
{ -logL2(theta) }

hello2 <- nlm(minuslogL2,c(0,0),hessian=T)
ML2 <- hello2$estimate
Jhat2 <- hello2$hessian
```

Partial derivatives:

```
here <- c(4.049,-5.909)
eps <- 0.001
oi1 <- ( cc2(here + eps*c(1,0)) - cc2(here - eps*c(1,0)) ) / (2*eps)
oi2 <- ( cc2(here + eps*c(0,1)) - cc2(here - eps*c(0,1)) ) / (2*eps)
```

Linear regression:

```
well <- glm(yy ~ xx)
estimates <- well$coef
residuals <- well$res
spread <- sqrt( mean(residuals^2) )
```