UNIVERSITETET I OSLO

Matematisk Institutt

EXAM IN:	**STK 4160/9160:**
	Model Selection and Model Averaging
	Part II of two parts
WITH:	**Nils Lid Hjort**
AUXILIA:	**Calculator, plus one single sheet of paper**
	with the candidate's own personal notes
TIME FOR EXAM:	**Part I: The Project, 7–19/vi/2017;**
	Part II: Friday 2/vi s.y., 9:00–13:00, written exam

This exam set contains four exercises and comprises five pages, including an informative Appendix which may be helpful for a few of the exercise points.

Exercise 1

This exercise looks into some likelihood analyses for binomial models, including ways of selecting between one and two unknown binomial parameters in a two-sample situation.

(a) Suppose X is binomial (n, p), with n known and probability parameter p unknown. Write down the log-likelihood function $\ell(p)$, find the maximum likelihood (ML) estimator $\widehat{p}$. Apply the general theory of the course to put up the approximate normal distribution for $\widehat{p}$.

(b) Show that the maximised log-likelihood function can be expressed as

$$\ell_{\max} = nH(\widehat{p}) + \log \binom{n}{x},$$

where

$$H(p) = p \log p + (1-p)\log(1-p) \quad \text{for } p \in [0,1].$$

We define the value of H at the two endpoints 0 and 1 to be equal to zero, by continuity.

(c) Assume now that there are two independent binomial experiments, with $X \sim \text{bin}(n, p)$ and $Y \sim \text{bin}(n, q)$. Write down the full log-likelihood function $\ell(p, q)$ for the combined data.

(d) We now consider two models. Model M_0 takes $p = q$, so that both X and Y come from the same population, whereas model M_1 does not assume anything regarding p or q, hence left as two free parameters. Find explicit expressions for

$$\ell_{0,\max} = \max\{\ell(p,q)\colon \text{model } M_0\} \quad \text{and} \quad \ell_{1,\max} = \max\{\ell(p,q)\colon \text{model } M_1\}.$$

(e) Assume $n = 100$ in the two binomial experiments, and that one observes $x = 16$ and $y = 25$. Which model does AIC select? Does BIC prefer the same?

(f) The BIC formula stems from a certain approximation to the model probabilities

$$r_0 = \Pr(\text{model } M_0 \,|\, x, y) \quad \text{and} \quad r_1 = \Pr(\text{model } M_1 \,|\, x, y),$$

under regularity assumptions regarding the prior distributions for the parameters under the different models. In this particular case one may compute these probabilities exactly, i.e. without resorting to the BIC type approximation. Assume (i) that the two models are equally likely, a priori; (ii) that under M_0, $p = q$ has a uniform distribution on $(0, 1)$; and (iii) that under M_1, p and q are independent and uniformly distributed on $(0, 1)$. Set up formulas for r_0 and r_1 (again, with $n = 100$ and $x = 16$, $y = 25$), and attempt to reach explicit expressions for these probabilities. You may use the fact that

$$\int_0^1 u^a (1-u)^b \, \mathrm{d}u = \frac{a!\, b!}{(a+b+1)!},$$

valid for integers a and b.

Notate bene: You are not required *to compute* r_0 and r_1, as the formulas will involve a list of factorials, for which you would need a computer to have their logs summed, or perhaps, at least, Stirling's formula from c. 1718.

Exercise 2

We consider independent and identically distributed observations $y_1, \ldots, y_n$ on the unit interval, and are in particular interested in estimating the underlying median μ. Two models will be encountered below.

(a) The first and simplest model has

$$\text{cumulative } F(y, \theta) = y^\theta \quad \text{and} \quad \text{density } f(y, \theta) = \theta y^{\theta - 1},$$

with θ an unknown positive parameter. We call this the narrow model, in that there will be a wider model to be worked with below. Write down the log-likelihood function, for this narrow model; identify the maximum likelihood (ML) estimator $\widehat{\theta}_{\text{narr}}$; and also the ML estimator $\widehat{\mu}_{\text{narr}}$ for the median.

(b) Under the conditions of this narrow model, explain how the ML estimator for μ can be supplemented with a confidence interval of level approximately 95%.

(c) The extended wide model takes

$$\text{cumulative } \bar{F}(y, \theta, \gamma) = 1 - \{1 - F(y, \theta)\}^\gamma = 1 - (1 - y^\theta)^\gamma \quad \text{for } y \in [0, 1],$$

where γ is another unknown and positive parameter. Show that the density can be written

$$\bar{f}(y, \theta, \gamma) = \gamma (1 - y^\theta)^{\gamma - 1} \theta y^{\theta - 1},$$

and find a formula for the median $\mu = \mu(\theta, \gamma)$ for this model.

(d) Show that the Fisher information matrix for this two-parameter model, computed at the narrow model, takes the form

$$J = \text{Var} \begin{pmatrix} (1/\theta)(1+\log Z) \\ 1+\log(1-Z) \end{pmatrix},$$

where Z is uniform on the unit interval.

Since $-\log Z$ and $-\log(1-Z)$ are seen to be unit exponential, they have unit variances, so that

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

with $c = \text{cov}\{\log Z, \log(1-Z)\}$, a number which can be computed to be -0.6449 (and which happens to be the same as $1-\pi^2/6$).

(e) Let $\widehat{\mu}_{\text{narr}}$ and $\widehat{\mu}_{\text{wide}}$ be the ML estimators of μ based on respectively the narrow and the wide models. For what range of γ can the simpler narrow-based estimator be expected to be more precise than the wide-based estimator?

(f) Assume that the true mechanism generating the data is the density

$$f_{\text{true}}(y) = f(y, \theta_0, 1+\delta/\sqrt{n}),$$

for appropriate θ_0 and δ. Put up the limiting distributions for

$$\sqrt{n}(\widehat{\mu}_{\text{narr}} - \mu_{\text{true}}) \quad \text{and} \quad \sqrt{n}(\widehat{\mu}_{\text{wide}} - \mu_{\text{true}}),$$

with μ_{true} being the true median (you may use results spelled out in the Appendix).

(g) Consider the AIC-post-selection estimator for the median,

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

Put up the limiting disribution for $\sqrt{n}(\widehat{\mu}_{\text{AIC}} - \mu_{\text{true}})$, using general results from the curriculum.

Exercise 3

Consider the general normal linear regression model, where the potential influence of covariates $x_{i,1}, \ldots, x_{i,p}$ on outcome variable y_i is modelled as

$$y_i = x_{i,1}\beta_1 + \cdots + x_{i,p}\beta_p + \varepsilon_i = x_i^{\text{t}}\beta + \varepsilon_i,$$

for observations $i = 1, \ldots, n$, where the ε_i are independent and distributed as $\text{N}(0, \sigma^2)$.

(a) Write up the log-likelihood function, say $\ell_n(\beta_1, \ldots, \beta_p, \sigma)$. The ML estimator for the β part is

$$\widehat{\beta} = (X^{\text{t}}X)^{-1}X^{\text{t}}y = \Big(\sum_{i=1}^n x_i x_i^{\text{t}}\Big)^{-1} \sum_{i=1}^n x_i y_i,$$

minimising $Q(\beta) = \sum_{i=1}^n (y_i - x_i^{\text{t}}\beta)^2$; here X is the $n \times p$ matrix of covariate vectors. You do not need to show this (here), but find the ML estimator $\widehat{\sigma}$ for σ.

(b) Show that the maximised log-likelihood may be expressed as

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

(c) Explain how AIC and BIC select the potentially most relevant covariates among $x_{i,1}, \ldots, x_{i,p}$.

(d) Suppose one is specifically interested in

$$\mu = \mathrm{E}\,(Y_0 \,|\, x_0) = x_0^{\mathrm{t}} \beta = x_{0,1}\beta_1 + \cdots + x_{0,p}\beta_p,$$

the expected response for a new object or individual with given covariates $x_{0,1}, \ldots, x_{0,p}$. Explain, briefly, how FIC goes about selecting the most relevant covariates for this x_0.

Exercise 4

Write up the essence of the AFIC method (the average weighted focused information criterion), using a maximum of two pages. You are again free to use material, notation and results summarised in the Appendix below.

Appendix

The following is a mini-summary of some of the core material from Chapters 6–7 in Claeskens and Hjort (2008), pertaining to a certain local neighbourhood large-sample framework, set here in the simpler framework of i.i.d. data. Notation and results given here may be used in Exercises 3 and 4 without proofs or further detailed discussion.

Assume independent observations $Y_1, \ldots, Y_n$ stem from a distribution with density function $f(y, \theta, \gamma)$, with θ of dimension p and γ of dimension q. With γ_0 a suitable null point in the parameter range for γ, corresponding to $f(y, \theta, \gamma_0)$ being a p-dimensional 'narrow model', assume that $\gamma = \gamma_0 + \delta/\sqrt{n}$, i.e. that the data follow the density

$$f_n(y) = f(y, \theta_0, \gamma_0 + \delta/\sqrt{n}).$$

As in the core material of the chapters mentioned, we shall work with a focus parameter $\mu = \mu(\theta, \gamma)$, with true value $\mu_{\mathrm{true}} = \mu(\theta_0, \gamma_0 + \delta/\sqrt{n})$. The necessary notation includes the Fisher information matrix J of dimension $(p+q) \times (p+q)$, computed at the null model, with inverse J^{-1}. These matrices have blocks according to the usual notation

$$J = \begin{pmatrix} J_{00} & J_{01} \\ J_{10} & J_{11} \end{pmatrix} \quad \text{and} \quad J^{-1} = \begin{pmatrix} J^{00} & J^{01} \\ J^{10} & J^{11} \end{pmatrix},$$

with J_{00} being of size $p \times p$, $Q = J^{11}$ of size $q \times q$, etc. Also, let

$$\tau_0^2 = (\tfrac{\partial \mu}{\partial \theta})^{\mathrm{t}} J_{00}^{-1} \tfrac{\partial \mu}{\partial \theta} \quad \text{and} \quad \omega = J_{10} J_{00}^{-1} \tfrac{\partial \mu}{\partial \theta} - \tfrac{\partial \mu}{\partial \gamma},$$

with derivatives computed at the null model. With maximum likelihood estimators $\widehat{\mu}_{\rm narr} = \mu(\widehat{\theta}_{\rm narr}, \gamma_0)$ and $\widehat{\mu}_{\rm wide} = \mu(\widehat{\theta}_{\rm wide}, \widehat{\gamma}_{\rm wide})$, under respectively the narrow and the wide models, we have

$$\sqrt{n}(\widehat{\mu}_{\rm narr} - \mu_{\rm true}) \to_d \Lambda_0 + \omega^{\rm t}\delta \quad \text{and} \quad \sqrt{n}(\widehat{\mu}_{\rm wide} - \mu_{\rm true}) \to_d \Lambda_0 + \omega^{\rm t}(\delta - D),$$

where $\Lambda_0 \sim {\rm N}(0, \tau_0^2)$ and $D \sim {\rm N}_q(\delta, Q)$ are independent. More generally, with $\widehat{\mu}_S$ the maximum likelihood estimator under submodel S, where S is a subset of $\{1, \ldots, q\}$,

$$\sqrt{n}(\widehat{\mu}_S - \mu_{\rm true}) \to_d \Lambda_0 + \omega^{\rm t}(\delta - G_S D),$$

where

$$G_S = \pi_S^{\rm t} Q_S \pi_S Q^{-1} = \pi_S^{\rm t}(\pi_S Q^{-1} \pi_S^{\rm t})^{-1} \pi_S Q^{-1}$$

is a $q \times q$ matrix determined by this S, involving the projection matrices π_S, where $\pi_S v$ maps $v = (v_1, \ldots, v_q)$ to the subvector v_S with v_j for $j \in S$.