STK 4160-9160: Model Selection & Model Averaging
Spring 2021

Course Notes and Exercises
by Nils Lid Hjort

– This version: as of 12 January 2021 –

1. Basic machinery for the linear-normal regression model

The classic linear-normal regression model is the most traditional way of investigating statistically how data $y_1, \ldots, y_n$ on n individuals relate to say p-dimensional covariate vectors $x_1, \ldots, x_n$. The model takes the following form:

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

where the ε_i are i.i.d. $\mathrm{N}(0, \sigma^2)$. Thus the model has $p+1$ (typically unknown) parameters; p for the regression surface and one for the spread. In more compact linear algebra language, the model may be represented as

$$y = X\beta + \varepsilon \sim \mathrm{N}_n(X\beta, \sigma^2 I_n),$$

where X is the $n \times p$ matrix with x_i^{t} on its ith row, $\beta = (\beta_1, \ldots, \beta_p)^{\mathrm{t}}$ is the vector of regression coefficients, and y and ε are the $n \times 1$ vectors collecting together the y_i and ε_i.

(a) Show that the log-likelihood function takes the form

$$\ell_n(\beta, \sigma) = -n \log \sigma - \tfrac{1}{2} Q(\beta)/\sigma^2 - \tfrac{1}{2} n \log(2\pi),$$

where $Q(\beta) = \sum_{i=1}^n (y_i - x_i^{\mathrm{t}}\beta)^2 = \|y - X\beta\|^2$ is the residual sum of squares.

(b) Show that the maximum likelihood (ML) estimators are

$$\widehat{\beta} = \mathrm{argmin}(Q) = (X^{\mathrm{t}}X)^{-1}X^{\mathrm{t}}y \quad \text{and} \quad \widehat{\sigma} = \sqrt{Q_{\min}/n},$$

where

$$Q_{\min} = \min Q(\beta) = Q(\widehat{\beta}) = \sum_{i=1}^n (y_i - x_i^{\mathrm{t}}\widehat{\beta})^2 = \|y - X\widehat{\beta}\|^2$$

is the sum of the squared estimated residuals.

(c) Let

$$\Sigma_n = (1/n)\sum_{i=1}^n x_i x_i^{\mathrm{t}}$$

be the empirical variance matrix of the n covariate vectors. Show that $\widehat{\beta}$ is unbiased with variance matrix $\sigma^2\Sigma_n^{-1}/n$, and that it is multinormally distributed;

$$\widehat{\beta} \sim \mathrm{N}_p(\beta, \sigma^2\Sigma_n^{-1}/n).$$

It is assumed here that Σ_n has full rank, which is equivalent to there being at least p linearly independent covariate vectors; in particular, $n \geq p$ (so the present machinery does not work if $p > n$).

(d) Show that the estimated residuals may be expressed as

$$\widehat{\varepsilon} = y - X\widehat{\beta} = (I_n - H)y,$$

in terms of the so-called hat matrix

$$H = X(X^{\mathrm{t}}X)^{-1}X^{\mathrm{t}}.$$

Show that H is symmetric and that $H^2 = H$ (making H a so-called idempotent matrix), and that its trace is $\mathrm{Tr}(H) = p$.

(e) Deduce from the above that

$$\widehat{\varepsilon} \sim \mathrm{N}_n(0, \sigma^2(I - H))$$

and that $Q_0 = \|\widehat{\varepsilon}\|^2$ has mean equal to $(n-p)\sigma^2$. Thus $\widehat{\sigma}^2 = Q_0/n$ is actually underestimating the real variance σ^2; a repaired version most frequently used (e.g. in output from software packages) is

$$s = \sqrt{Q_0/(n-p)} = \sqrt{\frac{n}{n-p}}\widehat{\sigma}.$$

Show that in fact

$$Q_0 \sim \sigma^2\chi^2_{n-p},$$

and that Q_0 and $\widehat{\beta}$ are stochastically independent. This is the basis for all *exact inference* for the linear-normal model. (The STK 4160 course aims however at developing and applying methodology that relies on first-order large-sample approximations, valid for general parametric models, as opposed to only the linear-normal regression model worked with in this exercise.)

(f) Show that the log-likelihood maximum is

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

with consequent AIC value

$$\mathrm{AIC} = -2n\log\widehat{\sigma} - 2(p+1) - n - n\log(2\pi).$$

Hence selecting a linear regression model via the AIC method, among competing candidates, is equivalent to searching for the model that has the smallest value of

$$\mathrm{crit} = n\log\widehat{\sigma} + p.$$

(Note that $\widehat{\sigma}$ depends on the model at hand, and changes value and interpretation if one e.g. pushes a covariate component in or out of the model.)

(g) Assume an estimate and a confidence interval are required for the parameter

$$\mu = \mathrm{E}(Y \,|\, x = x_0),$$

the regression surface value at a given position $x = x_0$ in the covariate space. If we are willing to assume that the model is adequate, then $\mathrm{E}(Y \,|\, x)$ is the same as $x^{\mathrm{t}}\beta$; the following arguments continue from this assumption of absence of bias. Show that the ML estimator is $\widehat{\mu} = x_0^{\mathrm{t}}\widehat{\beta}$, and that

$$\widehat{\mu} = x_0^{\mathrm{t}}\widehat{\beta} \sim \mathrm{N}(\mu, \sigma^2 x_0^{\mathrm{t}}\Sigma_n^{-1}x_0/n).$$

(h) From the above follows

$$Z_n = \frac{x_0^{\mathrm{t}}\widehat{\beta} - x_0^{\mathrm{t}}\beta}{\sigma v_n/\sqrt{n}} \sim \mathrm{N}(0,1) \quad \text{where } v_n = (x_0^{\mathrm{t}}\Sigma_n^{-1}x_0)^{1/2},$$

which is not immediately employable since σ is unknown. Argue however that

$$Z_n^* = \frac{x_0^{\mathrm{t}}\widehat{\beta} - x_0^{\mathrm{t}}\beta}{\widehat{\sigma} v_n/\sqrt{n}} = \frac{\sigma}{\widehat{\sigma}} Z_n \approx_d \mathrm{N}(0,1),$$

since $\widehat{\sigma}$ is close to σ with high probability. A precise mathematical version of this is that $Z_n^* \to_d \mathrm{N}(0,1)$ as n increases. Deduce that

$$\Pr\{x_0^{\mathrm{t}}\beta \in x_0^{\mathrm{t}}\widehat{\beta} \pm 1.96\,\widehat{\sigma} v_n/\sqrt{n}\} \approx 0.95,$$

for example, i.e. we have an approximate 95% confidence interval for $\mu = x_0^{\mathrm{t}}\beta$. Again, the accurate mathematical version of the approximation statement is that the left hand side probability converges to precisely 0.95, as $n \to \infty$.

(i) A more careful probability calculation, when the model conditions are exactly in force, gives *the exact* distribution of Z_n^* above. Show that indeed $Z_n^* \sim t_{n-p}$ (the t distribution with degrees of freedom equal to $n-p$). An exact 95% confidence interval is therefore

$$x_0^{\mathrm{t}}\beta \in x_0^{\mathrm{t}}\widehat{\beta} \pm t_{0.975,n-p}\,\widehat{\sigma} v_n/\sqrt{n},$$

in terms of the 0.975 quantile of the t_{n-p}. We note that the difference between this quantile and the corresponding 1.96 for the standard normal is not big, as soon as $n-p$ is say 30 or bigger (check this, via `qt(0.975,n-p)` in **R**). Also, both confidence intervals (those of (h) and (i)) are large-sample valid even if the underlying error distribution deviates from the normal.

2. Illustrations via simulations

This exercise leads to some concrete illustrations of the general linear-normal regression machinery summarised in Exercise 1. The themes are (i) selection of a good model, balancing precision with complexity, and (ii) construction of different confidence intervals for the same statistical parameter.

(a) For $n = 250$, generate first covariate values $x_i \sim \text{unif}(0, 1)$, and then observations

$$y_i = m(x_i) + \varepsilon_i \quad \text{for } i = 1, \ldots, n,$$

where the true regression curve is taken to be $m(x) = \exp(\sin(\pi x^2))$, and where the ε_i are i.i.d. $\text{N}(0, \sigma_0^2)$, with $\sigma_0 = 0.333$. The point is that the real data generating mechanism corresponds to a regression curve that is smooth but never inside the world of polynomial curves. – Plot the data.

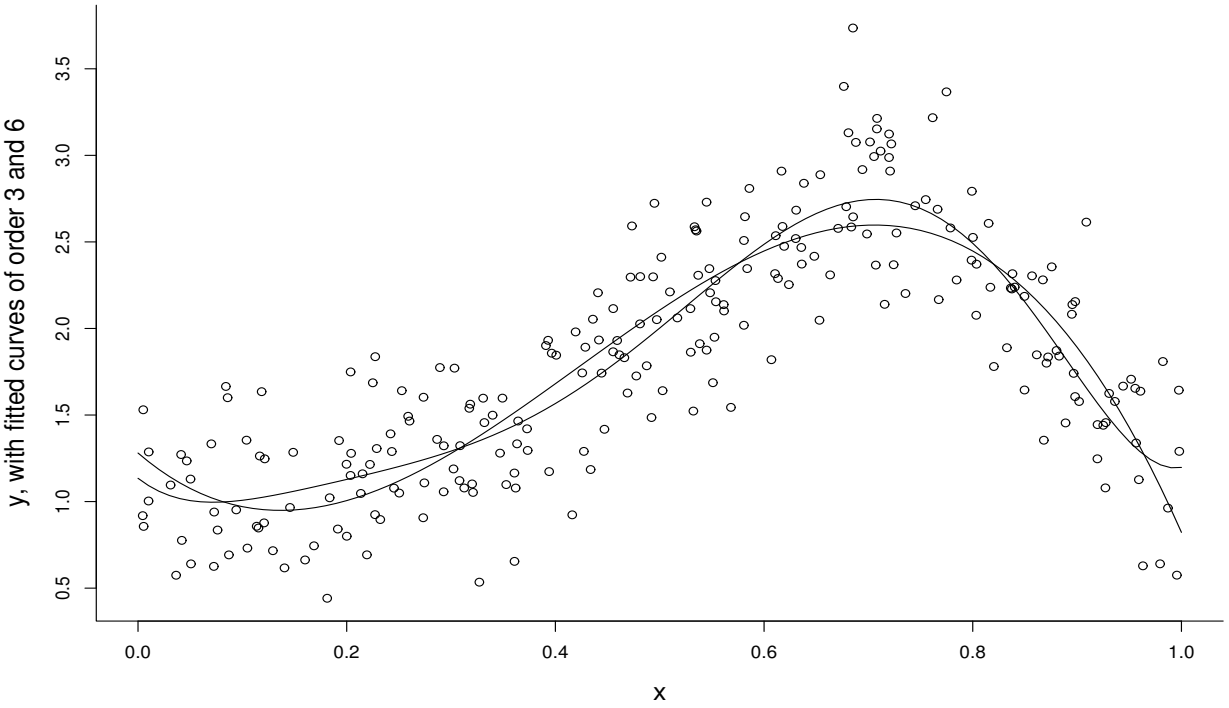

Figure 1: A total of $n = 250$ data points generated as described, along with fitted polynomial regression curves of order three and six.

(b) Fit each of the ten models $M_1, \ldots, M_{10}$ to the data, where model M_p is the linear-normal regression model corresponding to a polynomial order p model for the regression curve, i.e.

$$y_i = m_p(x_i) + \varepsilon_i' = \beta_0 + \beta_1 x_i + \cdots + \beta_p x_i^p + \varepsilon_i' \quad \text{for } i = 1, \ldots, n,$$

with ε_i' taken as $\text{N}(0, \sigma^2)$. For each of these ten models, compute the maximum log-likelihood valus $\ell_{n,\max}$ and the AIC value. You may use

`lm(yy ~ Xnow)` or `glm(yy ~ Xnow, family=gaussian)`

in **R**, where the `Xnow` matrix contains the columns $(x, x^2, \ldots, x^p)$, when you fit model M_p. Which of the ten models is best, as judged by the AIC?

(c) For each of the ten candidate models, compute also the model-based estimate of the parameter $\mu = m(x_0)$, where $x_0 = 0.75$. Create (and then study and comment upon) a table with the six columns model, maximum log-likelihood, model complexity, AIC value, $\widehat{\sigma}$, and $\widehat{m}(x_0)$.

(d) Extend the previous analysis to include not only the model based parameter estimate of $m(x_0)$, but also (still model based) 95% confidence intervals for this parameter. Include both the large-sample approximation and the exact t based versions. Thus create an extended table with ten columns, consisting of the six columns already given above, plus lower and upper confidence points (approximate) and lower and upper confidence points (exact). Comment on these confidence intervals.

3. Quotations and maxims

The point of this exercise is to gather a few maxims and quotations of significance, and then to ponder their relevance and implications for model selection and model averaging issues.

(a) 'All models are wrong, but some are useful' (most often attributed to G.E.P. Box).

(b) 'Entia non sunt multiplicanda praeter necessitatem' (more or less: entities should not be multiplied beyond necessity, called Ockham's razor, 1323, after the 14th century English logician and Franciscan friar William of Ockham).

(c) 'How odd it is that anyone should not see that all observation must be for or against some view if it is to be of any service' (C. Darwin).

(d) 'The purpose of models is not to fit the data, but to sharpen the questions' (S. Karlin).

(e) 'It can scarcely be denied that the supreme goal of all theory is to make the irreducible basic elements as simple and as few as possible without having to surrender the adequate representation of a single datum of experience' (A. Einstein, 1934; the somewhat vulgarised version of this is 'everything should be made as simple as possible, but not simpler').

(f) A famous exchange, after the 1782 premiere of KV 384 in Wien: Emperor Joseph II: "Gewaltig viel Noten, lieber Mozart." Mozart: "Gerade soviel Noten, Euer Majästät, als nötig sind."

4. Life-lengths in Roman era Egypt

Via the book's home page `feb.kuleuven.be/public/u0043181/modelselection/`, access the data set on mortality in ancient Egypt; see Example 2.6. In this exercise we take these 141 life-lengths (ranging from 1.5 to 96.0 years) to be i.i.d. from some underlying age-at-death distribution. Here we are considering four different models:

(i) the exponential distribution, with density $b\exp(-by)$;

(ii) the Gamma(a, b) distribution, with density $\{b^a/\Gamma(a)\}\, y^{a-1}\exp(-by)$;

(iii) the Gompertz model with parameters (a, b), with hazard rate $h(y) = a\exp(by)$ and a corresponding density of the form $f(y) = \exp\{-H(y)\}h(y)$, with $H(y)$ the cumulative hazard function, see the book;

(iv) the Weibull distribution, with cumulative distribution $F(y) = 1 - \exp\{-(y/a)^b\}$.

(a) Fit each of these four models using maximum likelihood. Also display the histogram of the data points along with the four estimated probability densities. (Use e.g. `hist(y,prob=T,breaks=12)` for the histogram part.)

(b) For each of the four models, compute approximate standard deviations for the parameter estimates involved, using the ML machinery formula

$$\operatorname{Var}\widehat{\theta} \doteq \widehat{J}_{\text{total}}^{-1},$$

where

$$\widehat{J}_{\text{total}} = -\ell_n''(\widehat{\theta}) = -\frac{\partial^2 \ell_n(\widehat{\theta})}{\partial\theta\partial\theta^{\mathrm{t}}},$$

the Hessian matrix associated with the log-likelihood function (cf. Section 2.2). Note that this matrix is found 'for free' via an application of the `nlm` algorithm in **R**, via `nlm(minusloglik, starthere, hessian=T)` for a pre-programmed `minusloglik` function and a suitable starting point `starthere` for the iterative method underlying the algorithm. The formula used here does assume that the parametric model used is adequate.

(c) We are interested in estimating the two parameters

$$\mu = \text{med}(F) = F^{-1}(\tfrac{1}{2}) \quad \text{and} \quad \kappa = F^{-1}(0.80),$$

the median and the 0.80 quantile point of the underlying life-time distribution. For each of the four models, find (i) estimates and (ii) approximate 90% confidence intervals for μ and for κ. You may use the formula

$$\operatorname{Var}\widehat{\mu} \doteq \widehat{c}^{\mathrm{t}} \operatorname{Var}\widehat{\theta}\,\widehat{c} \doteq \widehat{c}^{\mathrm{t}} \widehat{J}_{\text{total}}^{-1}\widehat{c}$$

from the large-sample theory of maximum likelihood, where $\widehat{\mu} = \mu(\widehat{\theta})$ is the ML estimate of $\mu = \mu(\theta)$, and where

$$\widehat{c} = \partial\mu(\widehat{\theta})/\partial\theta$$

is the vector of partial derivatives of $\mu(\theta)$, evaluated at the ML estimate. In practice one is free to use the numerical version

$$\widehat{c}_j = \frac{\mu(\widehat{\theta} + \varepsilon\, e_j) - \mu(\widehat{\theta} - \varepsilon\, e_j)}{2\varepsilon} \quad \text{for } j = 1, \ldots, p,$$

where $e_j = (0, \ldots, 1, \ldots, 0)^{\mathrm{t}}$ is the jth unit vector, and where one set e.g. $\varepsilon = 10^{-5}$. One may also use the `grad` operation of `library(numDeriv)`. Comment on any noticeable differences between these four estimates and confidence intervals, for μ and for κ. (We may expect more agreement 'in the middle' than in the tails.)

(d) For the four models, compute maximal log-likelihood values and AIC scores. Which model is best, and which is worst (as judged by this criterion)?

(e) After these analyses, which estimates and confidence intervals would you 'publish', for the median and 0.80 quantile point?

5. Modelling nerve impulse data

Access the site `www.stat.ncsu.edu/sas/sicl/data/nerve.dat` to find $n = 799$ nerve impulse data (the time intervals between successive pulses along a certain nerve fibre), measured in seconds, and ranging from 0.01 to 1.38. It is traditionally assumed that such data ought to follow an exponential distribution. The figure displays a histogram of the data, along with two fitted model densities; see the points below.

(a) Fit the following five parametric models to these data, via maximum likelihood estimation (and assuming independence). Display for each model the parameter estimates and their estimated standard deviations.

1. The exponential, with density $\theta \exp(-\theta x)$.
2. The gamma, with density $\{b^a/\Gamma(a)\}x^{a-1}\exp(-bx)$.
3. The cut-off normal, with density $K(b,c)^{-1}\exp(-bx-\frac{1}{2}cx^2)$ on $(0,\infty)$; show that

$$K(b,c) = \sqrt{2\pi}\exp(\tfrac{1}{2}b^2/c)\{1-\Phi(b/\sqrt{c})\}/\sqrt{c},$$

and that the cumulative distribution function is of the form

$$F(x,b,c) = \{\Phi(\sqrt{c}x + b/\sqrt{c}) - \Phi(b/\sqrt{c})\}/\{1-\Phi(b/\sqrt{c})\}.$$

4. The 'Beta-enveloped' exponential, with cumulative distribution function and density respectively equal to

$$F(x) = 1 - \mathrm{Be}(\exp(-x), a, b) \quad \text{and} \quad f(x) = \mathrm{be}(\exp(-x), a, b)\exp(-x).$$

Here $\mathrm{be}(u,a,b)$ and $\mathrm{Be}(u,a,b)$ are the density and cumulative of a Beta distribution with parameters (a,b).

5. The Weibull, with cumulative distribution function $F(x) = 1 - \exp\{-(x/a)^b\}$.

(b) Note that the exponential density is included as special cases in each of the models 2, 3, 4, 5 (corresponding respectively to the cases $a = 1$; $c = 0$; $b = 1$; $b = 1$). Compute and display confidence intervals for these four expo-extending parameters, and check whether the implied tests for exponentiality accept or reject this hypothesis.

(c) Use the parameter estimates to construct Figure 2, with the histogram supplemented with two or more fitted densities.

(d) Compute AIC and BIC scores for each of the models. Which model appears to be preferred? Assuming that the five models are a priori equally likely, what are the approximate posterior probabilities for the models?

(e) Define $p = \Pr\{X \geq 0.333\}$, in terms of a future random observation X from the nerve impulse distribution. For each of the five models, estimate p, and compute both $\widehat{\tau}$ and $\widehat{\tau}/\sqrt{n}$, with $\widehat{\tau}$ the estimate of the standard deviation of the limit distribution of $\sqrt{n}(\widehat{p} - p)$, using the theory of Exercise 4. Compare to the direct nonparametric estimate and its standard deviation. See the table below.

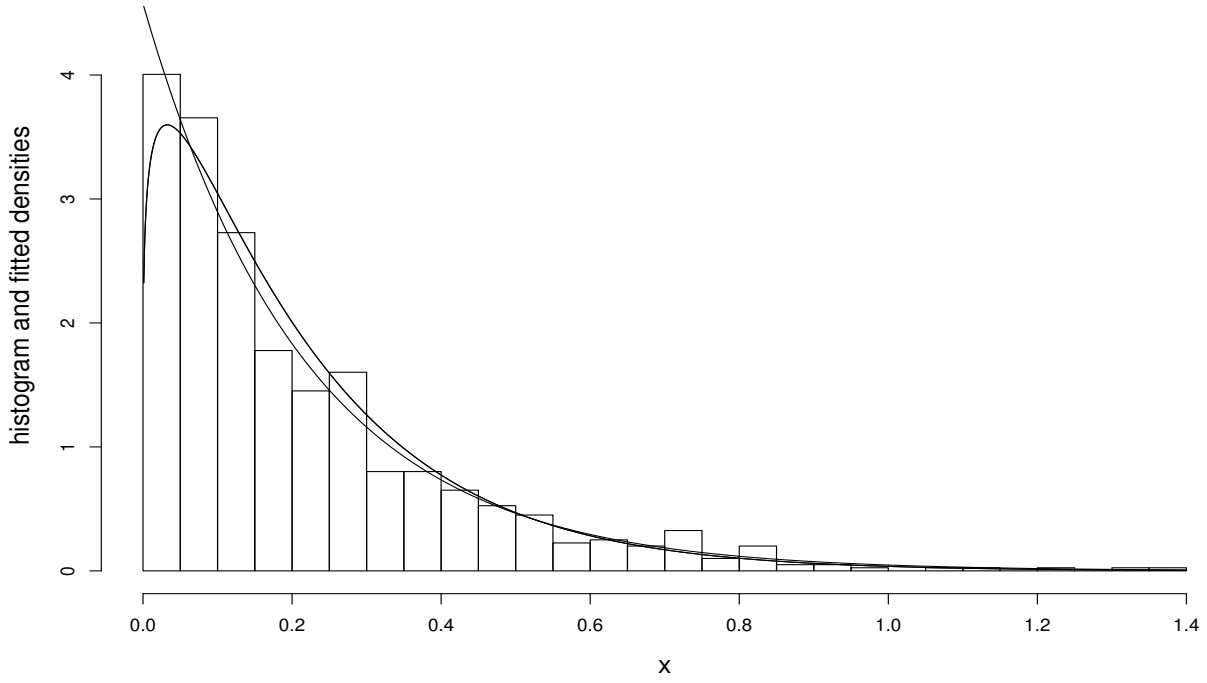

Figure 2: Histogram of the 799 nerve impulse data points, along with two fitted densities, corresponding to the exponential and the Beta envelope model.

(f) Try to find or invent an appropriate three-parameter model that may do even better than the best of the five models studied above, i.e. outscoring the Beta envelope model in this table; for the AIC selector, this is tantamount to finding a model with higher log-likelihood value than 423.167. (See Exercise 7.)

	dim	logLmax	AIC	BIC	phat	sd	sd/rootn	
1	1	415.987	829.973	825.290	0.2179	0.3320	0.0117	expo
2	2	422.150	840.300	830.933	0.2144	0.3317	0.0117	gamma
3	2	416.821	829.642	820.275	0.2212	0.3421	0.0121	cut-off
4	2	422.167	840.334	830.968	0.2143	0.3317	0.0117	envelope
5	2	420.009	836.018	826.651	0.2178	0.3323	0.0118	Weibull
6					0.2203	0.4144	0.0147	nonpmic

(g) Figure 2 indicates that to the extent the exponential distribution does not fully fit the data, the discrepancy may take place for the very smallest data points. For at least models 1, 2, 4, use the robustly weighted Kullback–Leibler estimation and model selection methods of Section 2.10.2 to (i) re-estimate the parameters and (ii)

re-compare the models, in terms of the wAIC criterion. Use weight function $w(y) = \exp(-3y)$. (Note that care must be taken regarding the final interpretation of such model comparison exercises, in that the smallest observations may not have been measured with sufficient accuracy; they are here recorded to 0.01 seconds precision level only.)

6. Predicting y2 from y1, y3, y4: The Adelskalenderen

Access and organise in your computer the *Adelskalenderen* data from the book's website, about the best speedskaters in the world. The table gives for each skater the personal bests y_1, y_2, y_3, y_4 over the four classical distances 500-m, 1500-m, 5000-m, 10000-m. Translate these times to the scale of seconds. The ranking is in terms of the canonical point-sum

$$y_1 + y_2/3 + y_3/10 + y_4/20,$$

in terms of 500-m times, so that e.g. Chad Hedrick's 35.58, 1:42.78, 6:09.68, 12:55.11 are translated into 35.580, 34.260, 36.968, 38.755, with samalogue sum 145.563. The data set available at the book's website stems from end-of-season 2006 (with Eskil Ervik #6 and Håvard Bøkko #15); you may check e.g. via links found in wikipedia that as of 15/iii/9, Ervik is #8 and Bøkko is #4 (whereas Shani Davis and Sven Kramer have passed Hedrick). – The present exercise is about predicting the 1500-m time from the other three results. For this purpose we focus on the $n = 250$ best skaters on the Adelskalenderen.

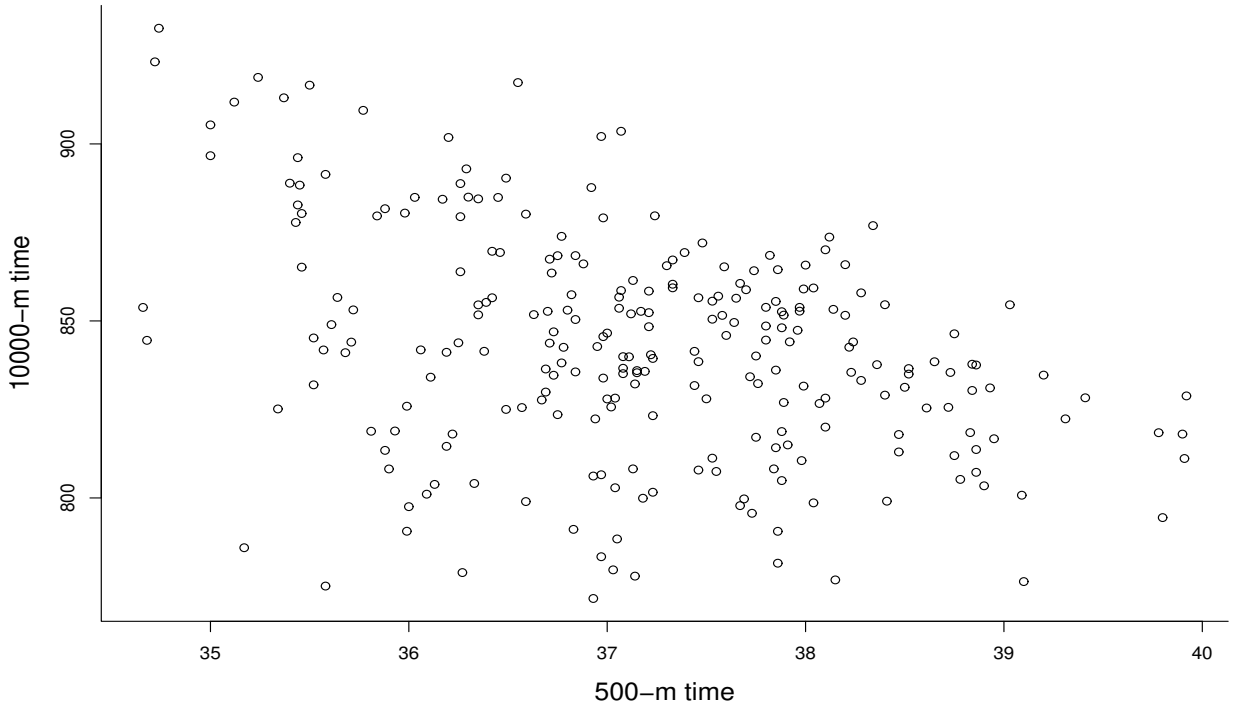

Figure 3: Personal best times for 500-m and 10000-m, for the 250 best skaters of the world (as per the Adelskalenderen, end-of-season 2006).

(a) Find Johan Olav Koss in the figure. Why is he not among the very best anymore?

(b) Consider at the outset seven different models for explaining y_2 in terms of y_1, y_3, y_4, namely those termed 1, 3, 4, 13, 14, 34, 134. Here e.g. '14' means the linear regression model that takes y_1, y_4 as covariates (but not y_3). Fit these seven models to the Adelskalenderen data (with the $n = 250$ top skaters), and compute AIC and BIC scores. Make a convenient table that displays model, dimension (the number of unknown parameters in the model), log-likelihood maximum, $\widehat{\sigma}$, AIC, BIC. Here $\widehat{\sigma}$ is the maximum likelihood estimate of the scale parameter in the appropriate $\mathrm{N}(x_i^{\mathrm{t}}\beta, \sigma^2)$ model. Identify the AIC winner and the BIC winner, among these seven models, and discuss their assumptions and properties.

(b) From the BIC scores, and assuming that the seven models were equally likely a priori, compute approximate posterior model probabilities.

(c) It turns out, not too surprisingly, that the best of these seven models is model 13, the one taking 500-m and 5000-m on board but not the 10000-m. Make some plots to check whether the *constant variability* assumption, implicit in the linear regression model 13, looks reasonable or not.

(d) Consider two skaters A and B. Skater A is fabulously strong, with personal bests 35.00 and 6.20.00 on the 500-m and 5000-m; skater B is rather more mediocre, with personal bests 37.00 and 6.40.00. For these two skaters, and using model 13, (i) provide a point estimate of the skater's 1500-m time (i.e. his predicted time), and (ii) supplement this predicted time with an approximate (or exact) 90% prediction interval.

– Inspired (or not) by point (c), we shall now investigate *three more models*, which I choose to call 13plus1, 13plus3, 13plus13. Here '13plus13' is the model that in addition to the linear mean of model 13 includes variance heterogeneity in y_1 and y_3, of the form

$$y_{2,i} = \beta_0 + \beta_1 y_{1,i} + \beta_3 y_{3,i} + \sigma_i \varepsilon_i \quad \text{for } i = 1, \ldots, n,$$

where

$$\sigma_i = \sigma \exp(\gamma_1 v_{1,i} + \gamma_3 v_{3,i}), \quad \text{with } v_{1,i} = y_{1,i} - \bar{y}_1 \text{ and } v_{3,i} = y_{3,i} - \bar{y}_3.$$

Finally, the ε_i here are i.i.d. $\mathrm{N}(0,1)$. The point of subtracting the means of y_1 and y_3 here is to give σ an easy interpretation, as the standard deviation parameter of y_2 data in the middle of the (y_1, y_3) ranges.

(e) For model 13plus13, write down a mathematical expression for the log-likelihood function (in terms of six unknown parameters); programme this function, and find the maximum likelihood estimates numerically (cf. the `nlm` algorithm discussed in Exercise 4). Check and comment on the confidence intervals for γ_1 and γ_3. For skaters A and B (see above), compute again their predicted 1500-m times, along with 90% prediction intervals. Comment on any differences with your answers under point (d).

(f) Find AIC and BIC scores for the four models 13, 13plus1, 13plus3, 13plus13. Which of these appears to be best? What are the (approximate) probabilities of these four models, given the data, assuming that they are equally before this question came to your attention?

(g) For model 13plus3, attempt to check its adequacy, perhaps via plots or goodness-of-fit testing. Discuss your findings.

7. An extended Gamma-Weibull distribution

Consider the density of the form

$$f(y, a, b, c) = k(a, b, c) y^{a-1} \exp(-by^c) \quad \text{for } y > 0,$$

where a, b, c are positive parameters. I have not seen such a thing in the literature before, but it appears to be a useful extension of both the Gamma and the Weibull families; see also page 142 in the book.

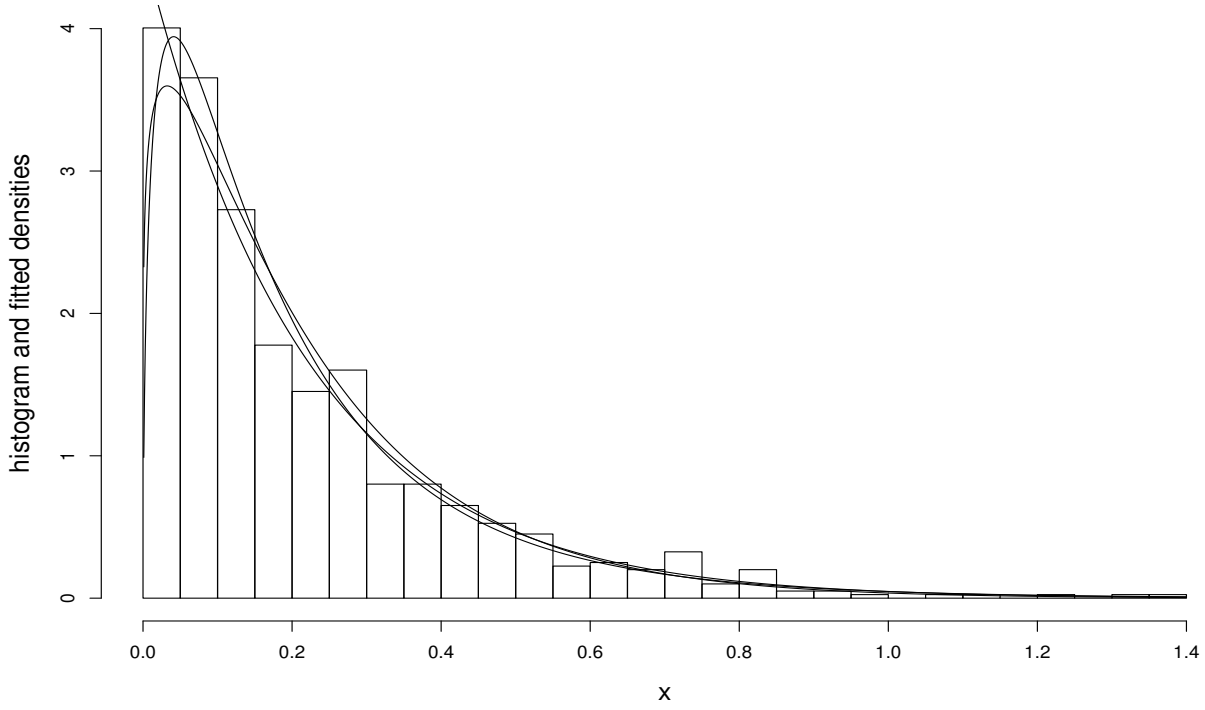

Figure 4: Histogram of the 799 nerve impulse data points, along with three fitted densities, corresponding to the exponential, the Gamma, and the extended Gamma-Weibull model (with the highest peak).

(a) Show that the Gamma and Weibull distributions are indeed special cases, and carry out the required integration exercise to show that

$$k(a, b, c) = c \frac{b^{a/c}}{\Gamma(a/c)}.$$

(b) For the $n = 799$ nerve impulse data points of Exercise 5, fit this three-parameter model, by maximising the log-likelihood function

$$\ell_n(a, b, c) = \sum_{i=1}^{n} \{\log c + (a/c) \log b - \log \Gamma(a/c) + (a-1) \log y_i - b y_i^c\}.$$

Construct the plot of Figure 4, with the fitted Gamma and the extended Gamma-Weibull densities. Also compute standard errors for the $(\widehat{a}, \widehat{b}, \widehat{c})$ in question. [I find $(1.6004, 7.2877, 0.6442)$ for $(\widehat{a}, \widehat{b}, \widehat{c})$.]

(c) Test the hypotheses $H_{0,G}$ and $H_{0,W}$ that the data come from respectively a Gamma distribution or a Weibull distribution. Compute AIC and BIC scores and demonstrate that the three-parameter Gamma-Weibull-extension family really works better than both the Gamma and the Weibull. Convert the BIC scores to approximate posterior model probabilities for the three models Gamma, Weibull, Gamma-Weibull-extension.

(d) Show that the three score functions (the log-density derivatives with respect to the three parameters), computed at the narrow Gamma model (i.e. at a position $(a, b, 1)$, where $c = 1$), take the form

$$\begin{aligned} U_1(y) &= \log(by) - \psi(a), \\ U_2(y) &= a/b - y = (1/b)(a - by), \\ V(y) &= (\log b)(by - a) - \{by \log(by) - \{1 + a\psi(a)\}. \end{aligned}$$

Here $\psi(a) = \Gamma'(a)/\Gamma(a)$ is the log-derivative of the gamma function. Note that under the conditions of the narrow Gamma model, the variable $Y^* = bY$ follows a Gamma distribution with parameters $(a, 1)$. Show also that

$$\mathrm{E}\, Y^* = a, \quad \mathrm{E}\, \log Y^* = \psi(a), \quad \mathrm{E}\, Y^* \log Y^* = 1 + a\psi(a),$$

which is consistent with the mathematical fact that the score functions always have mean zero.

(e) As made clear in Chapters 5 and 6, part of the theory of tolerance radii, comparisons of different model based estimators, compromise estimators, etc., is driven by the Fisher matrix of the wide model, computed under the narrow model. Here this matrix is

$$J = J_{\text{wide}} = \mathrm{Var} \begin{pmatrix} U_1(Y) \\ U_2(Y) \\ V(Y) \end{pmatrix} = \mathrm{Var} \begin{pmatrix} W_2 \\ -(1/b)W_1 \\ (\log b)W_1 - W_3 \end{pmatrix},$$

where

$$W_1 = Y^* - a, \quad W_2 = \log Y^* - \psi(a), \quad W_3 = Y^* \log Y^* - \{1 + a\psi(a)\}.$$

Formulae for the components of J may be put up, by computing variances and covariances of W_1, W_2, W_3; note that these have distributions depending only on a. One may use numerical integration or simulation of say a million score vectors from the required $bY \sim \text{Gamma}(a, 1)$.

(f) The nerve impulse data fitted to the Gamma distribution yield parameter estimates $(\widehat{a}_{\rm narr}, \widehat{b}_{\rm narr}) = (1.1738, 5.3704)$. Compute the matrices

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

at this position, with J_{00} and J^{00} being 2×2 matrices, etc. Compute in particular $\kappa = (J^{11})^{1/2}$ and the tolerance radius around the Gamma distribution with respect to the Gamma-Weibull-extension model. Conclude that for $|c-1| \leq 3.383/\sqrt{n}$, Gamma-based inference is better than that based on the extended three-parameter model.

8. Stretching the linear regression model

The classic linear regression model operates under these assumptions:

(i) linearity of the conditional mean function;
(ii) constancy of the variance level across covariates;
(iii) independence of observations given the covariates;
(iv) normality of the residuals (observations minus mean).

Each of these is often overlooked or left unchecked in statistical practice (depending on the data and their context, as well as on the practitioner), with various side effects and dangers. Often, assumption (iv) is less crucial than the others.

In the general spirit of Chapter 5, one may invent different extensions of the classic linear model, involving one or more extra parameters, both because these may be of interest and importance in their own right, and because we may study aspects of parametric robustness, tolerance radii with respect to different model departures, etc. The variance heterogeneity models studied in Exercise 6 are e.g. of this type, exemplifying a departure from assumption (iii), of the type

$$y_i = x_i^{\rm t}\beta + \sigma \exp(\gamma v_i)\varepsilon_i \quad \text{with } \varepsilon_i \sim {\rm N}(0,1) \text{ for } i = 1, \ldots, n,$$

for a suitable v_i. In this exercise we are concerned with a departure from assumption (i), by 'stretching the linearity'; see Figure 5.

(a) For the speedskating Adelskalenderen dataset (with the top $n = 250$ skaters, as in Exercise 6), perform first an ordinary linear regression model for the 10000-m time $y = y_4$ on the 500-m time $x = y_1$:

$$y_i = \beta_0 + \beta_1(x_i - \bar{x}_n) + \varepsilon_i \quad \text{for } i = 1, \ldots, n,$$

with $\varepsilon_i \sim {\rm N}(0, \sigma^2)$. (It is convenient to centre the covariate in this fashion, for better comparison with the extended model below, and since it gives β_0 an easy interpretation.) Plot the linear regression line along with the (y_1, y_4) point cloud, and compute the associated AIC score for this model [I find -2397.328].

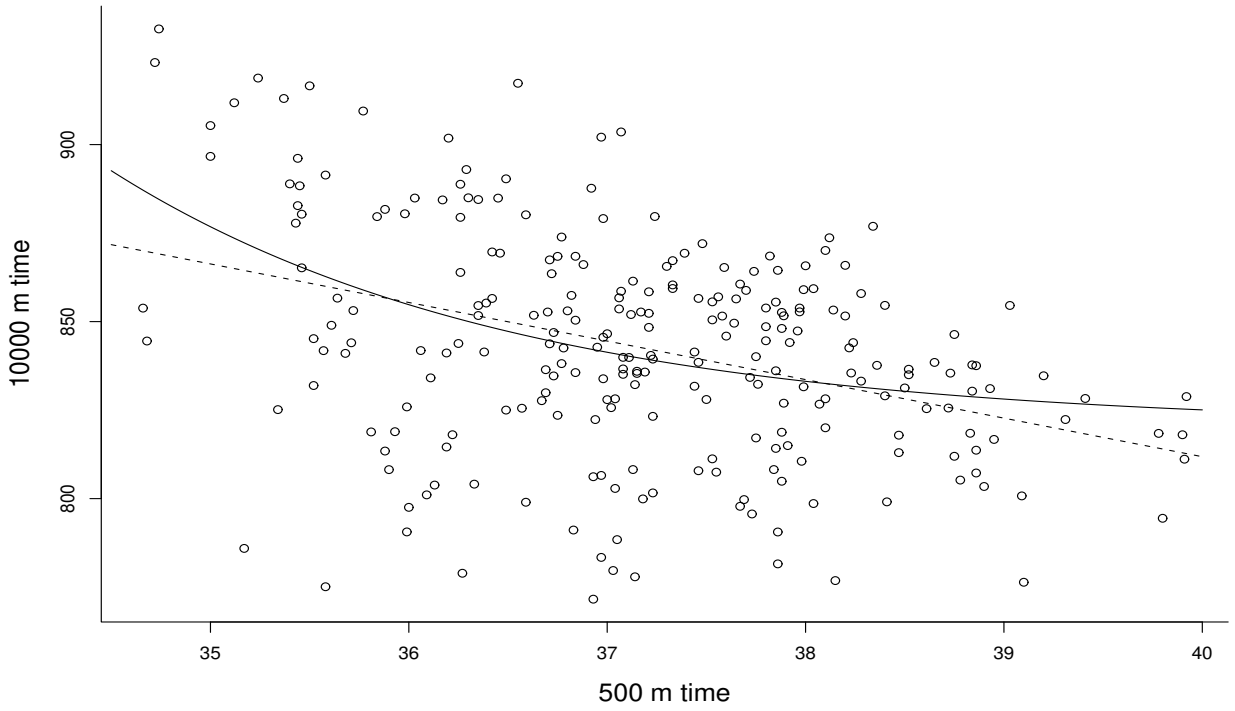

Figure 5: Personal best times for 500-m and 10000-m, for the 250 best skaters of the world (as per the Adelskalenderen, end-of-season 2006), along with fitted linear and a nonlinear regression lines. Can you spot Øystein Grødum?

(b) Study and plot the function

$$g(u,\gamma) = \frac{\exp(\gamma u) - 1}{\gamma}$$

for u in a suitable window around zero, for different values of γ. Show that its series expansion is $u + \frac{1}{2}\gamma u^2 + \frac{1}{6}\gamma^2 u^3 + \cdots$, so that in particular $g(u,0)$, defined by continuity, is simply the function $g(u,0) = u$.

(c) For the speedskating data, study and fit the extended regression model

$$y_i = \beta_0 + g(\beta_1(x_i - \bar{x}_n),\gamma) + \varepsilon_i = \beta_0 + \{\exp(\gamma\beta_1(x_i - \bar{x}_n)) - 1\}/\gamma + \varepsilon_i,$$

where again $\varepsilon_i \sim \mathrm{N}(0,\sigma^2)$. For small $|\gamma$, this is close to ordinary linear regression, with larger discrepancy with larger $|\gamma$. Plot also the resulting curved regression line, as in Figure 5, and compute the associated AIC value [I find -2396.210].

(d) In general terms, show that the ML estimates of (β_0,β_1,γ) are those minimising the sum of squares

$$Q(\beta_0,\beta_1,\gamma) = \sum_{i=1}^{n}\{y_i - \beta_0 - g(\beta_1(x_i - \bar{x}_n),\gamma)\}^2,$$

along with $\widehat{\sigma} = (Q_{\min}/n)^{1/2}$.

(e) Compute the log-derivatives of the model density function for y_i, and represent these as

$$\begin{aligned}
\partial \log f_i/\partial \beta_0 &= (1/\sigma)\varepsilon_i, \\
\partial \log f_i/\partial \beta_1 &= (1/\sigma)\varepsilon_i g'(\beta_1(x_i-\bar{x}_n),\gamma)(x_i-\bar{x}_n), \\
\partial \log f_i/\partial \sigma &= (1/\sigma)(\varepsilon_i^2-1), \\
\partial \log f_i/\partial \gamma &= (1/\sigma)\varepsilon_i g^*(\beta_1(x_i-\bar{x}_n),\gamma),
\end{aligned}$$

where ε_i is as in (c), whereas $g'(u,\gamma)$ and $g^*(u,\gamma)$ are the derivatives of $g(u,\gamma)$ with respect to respectively u and γ. Show that for $\gamma = 0$, these two derivatives are respectively 1 and $g^*(u,0) = \frac{1}{2}u^2$.

(f) Use the framework of Chapter 5 to find that the normalised information matrix of the wider four-parameter model, computed under the narrow three-parameter model, takes the form

$$J_n = J_{n,\text{wide}} = n^{-1}\sum_{i=1}^n \text{Var}\,\frac{1}{\sigma}\begin{pmatrix}\varepsilon_i \\ \varepsilon_i(x_i-\bar{x}_n) \\ \varepsilon_i^2-1 \\ \varepsilon_i v_i\end{pmatrix} = \frac{1}{\sigma^2}\begin{pmatrix}1 & 0 & 0 & \bar{v}_n \\ 0 & s_{n,x}^2 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ v_n & 0 & 0 & n^{-1}\sum_{i=1}^n v_i^2\end{pmatrix},$$

in terms of $s_{n,x}^2 = n^{-1}\sum_{i=1}^n (x_i-\bar{x}_n)^2$ and

$$v_i = g^*(\beta_1(x_i-\bar{x}_n),0) = \tfrac{1}{2}\beta_1^2(x_i-\bar{x}_n)^2.$$

(g) Deduce that

$$\kappa^2 = J_n^{11} = \frac{4\sigma^2}{\beta_1^4 s_{n,(x-\bar{x}_n)^2}^2},$$

in the notation of Chapter 5, where

$$s_{n,(x-\bar{x}_n)^2} = n^{-1}\sum_{i=1}^n (x_i-\bar{x}_n)^4 - \Big\{n^{-1}\sum_{i=1}^n (x_i-\bar{x}_n)^2\Big\}^2$$

is the empirical variance of the $(x_i-\bar{x}_n)^2$; with consequent tolerance radius

$$\kappa/\sqrt{n} = \frac{2\sigma}{\beta_1^2 s_{n,(x-\bar{x})^2}}\frac{1}{\sqrt{n}}$$

for the linear regression model with respect to the nonlinear stretching departure.

(h) Estimate γ, as well as κ and the tolerance radius, for the speedskating data above. Does it appear reasonable to use the wider model?

(i) Consider the parameter $\mu = \text{E}(Y \,|\, x_0)$, the expected 10000-m time for a skater with 500-m time $x_0 = 35.00$. Compute 90% confidence intervals for this μ, under the narrow and the wide models. Which would you prefer?

9. Stretching the tail of Gauß

The normal density model is and remains the most prominent of all, of course, for many reasons including tradition and mathematical convenience. Consequently, almost too many data sets are fitted to the normal, even data that exhibit non-normal features, like tails that are fatter than Gauß. Here we shall study the three-parameter extended density

$$f(y) = f(y, \xi, \sigma, \gamma) = \frac{1}{\sigma}\frac{1}{a(\gamma)} \exp\Big\{-\tfrac{1}{2}\Big|\frac{y-\xi}{\sigma}\Big|^{\gamma}\Big\} \quad \text{for } y \in \mathcal{R}.$$

With $\gamma = 2$ we are back to Gauß, with $a(2) = \sqrt{2\pi}$; also, $\gamma = 1$ corresponds to the double exponential model. Tails are fatter than Gauß for $\gamma < 2$ and slimmer than Gauß for $\gamma > 2$.

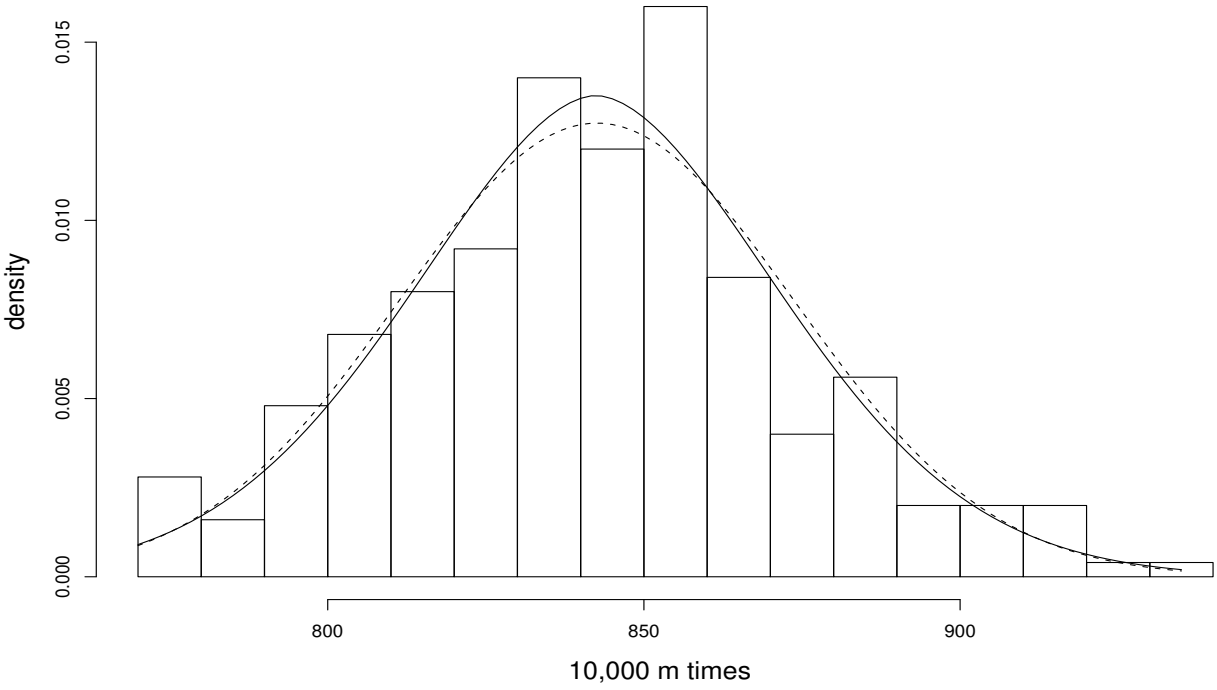

Figure 6: Histogram of personal best times 10000-m, for the 250 best skaters of the world (as per the Adelskalenderen, end-of-season 2006), along with two fitted densities: the normal (dotted line) and the three-parameter tail-extension (full line). Where is Lasse Sætre?

(a) Show that in fact

$$a(\gamma) = \int \exp(-\tfrac{1}{2}|x|^{\gamma})\,\mathrm{d}x = 2^{1+1/\gamma}\Gamma(1+1/\gamma).$$

(b) Access the Adelskalenderen data (cf. Exercises 6 and 8), and fit the $n = 250$ 10000-m personal best times to (i) the normal model and (ii) the three-parameter extension. Plot the consequent estimated densities on top of the histogram (cf. Figure 6).

(c) Compute AIC scores for the two models, and comment on these.

(d) Show that the Fisher information matrix of the wide three-parameter model, computed at the narrow two-parameter normal model, takes the form

$$J = \mathrm{Var}\begin{pmatrix} \varepsilon/\sigma \\ (\varepsilon^2-1)/\sigma \\ -\frac{1}{2}\varepsilon^2\log|\varepsilon| + b \end{pmatrix} = \begin{pmatrix} 1/\sigma^2 & 0 & 0 \\ 0 & 2/\sigma^2 & c/\sigma \\ 0 & c/\sigma & k \end{pmatrix},$$

where ε is standard normal. Here b is the constant making the score function component have mean zero (one finds $b = 0.1824$), and

$$c = \mathrm{cov}(\varepsilon^2 - 1, -\tfrac{1}{2}\varepsilon^2\log|\varepsilon|) = -0.8648, \quad k = \mathrm{Var}\,(-\tfrac{1}{2}\varepsilon^2\log|\varepsilon|) = 0.4242$$

(using numerical integration, or simulation). Deduce that the tolerance radius around the normal, in direction of tailstretching, is $\kappa/\sqrt{n}$, with $\kappa = 4.4599$.

(e) For the focus parameter $\mu = q_{0.90}$, the 0.90 quantile of the underlying distribution, compute ML estimates based on (i) the normal and (ii) the extended model. Which of the two estimates is to be preferred?

10. Gammaweibulling the exponential

Consider the three-parameter model density

$$f(y,a,b,c) = k(a,b,c)y^{a-1}\exp(-by^c) \quad \text{for } y > 0,$$

as in Exercise 7. Presently we are to view this as a two-parameter 'gammaweibull extension' of the traditional one-parameter Exponential model $b\exp(-by)$, corresponding to $(a,c) = (1,1)$. We may e.g. take an interest in comparing inference based on the narrow exponential model vs. that using the three-parameter extended model in connection with the 799 nerve impulse data of Exercise 7, where the ML estimate of b is 7.2877.

(a) Find the score function components $U(y), V_1(y), V_2(y)$, in the notation of the book's Chapter 5 (i.e. the derivatives of the log-density w.r.t. the three parameters, but evaluated at the null model), and find the associated Fisher information matrix $J = J_{\mathrm{wide}}$. Find also the 2×2 matrix $Q = J^{11}$. I find in fact

$$Q = \begin{pmatrix} 4.7616 & -5.8327 \\ -5.8327 & 10.5944 \end{pmatrix}$$

for this matrix (independent on the value of b); see point (d) below.

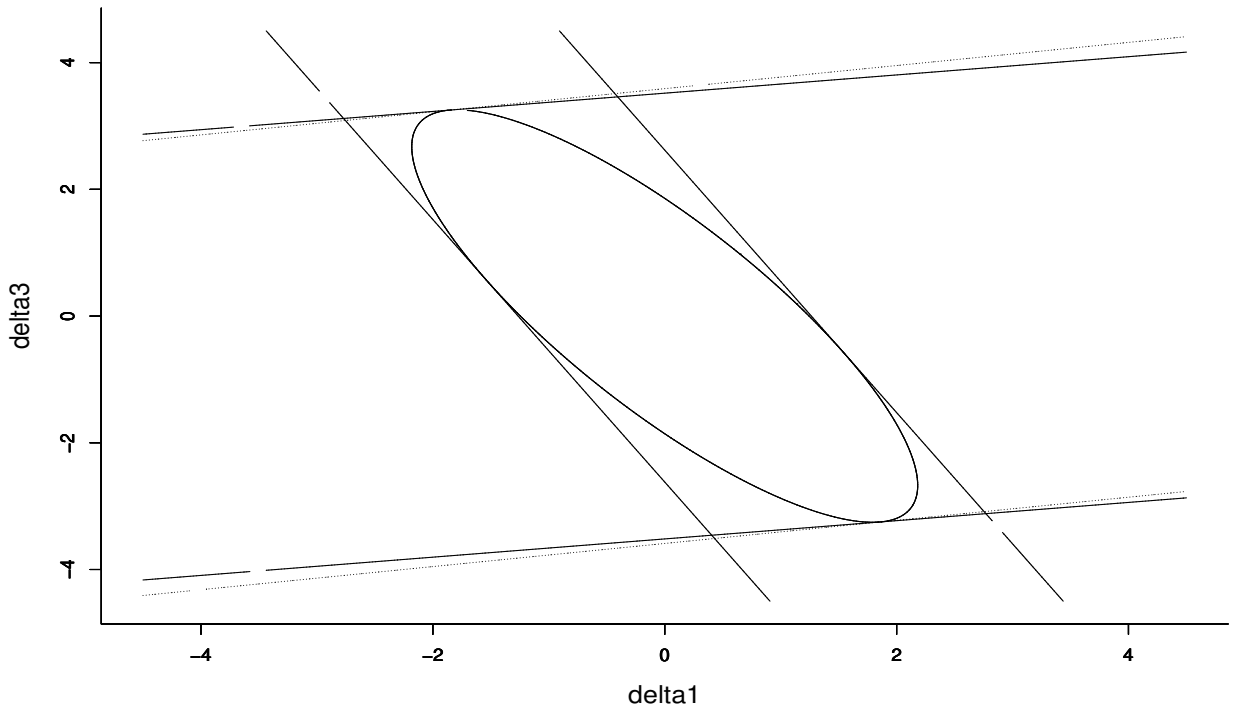

Figure 7: Tolerance ellipse, displaying the inner sanctum inside which the exponential model yields better inference than the three-parameter gammaweibull extension, in terms of $\delta_1 = \sqrt{n}(a-1)$ *and* $\delta_3 = \sqrt{n}(c-1)$*. Also shown are three tolerance strips, inside which exponential based inference is better, for estimands* $\mathrm{sd}(f)$ *(horizontal like), the probability that* $y \le x_0$ *with* $x_0 = 0.10$ *(North-Western to South-East direction), and the hazard rate at time* $x_h = 0.25$ *(again horizontal like). Calculations are carried out for* $b = 7.2877$*, the ML value for the nerve impulse data of Exercise 7.*

(b) Write $a = 1 + \delta_1/\sqrt{n}$ and $c = 1 + \delta_3/\sqrt{n}$. Draw the ellipse of (δ_1, δ_3) inside which inference based on the simple exponential model is better than using the wide model, for all estimands $\mu = \mu(a, b, c)$; cf. Figure 7.

(c) Consider now four different focus parameters, respectively the mean $\xi = \mathrm{E}\,Y$; the standard deviation $\sigma = \mathrm{sd}\,Y$; the probability $p = \Pr\{y \le x_0\}$ with $x_0 = 0.10$; and the hazard rate $h = f(x_h)/\{1 - F(x_h)\}$ at time position $x_h = 0.25$. For these, find numerical values for the ω vectors

$$\omega = J_{10} J_{00}^{-1} \frac{\partial \mu}{\partial \theta} - \frac{\partial \mu}{\partial \gamma},$$

in usual notation, where θ is b, the parameter of the narrow model, and $\gamma = (a, c)$ the extension parameters of the wider model. Draw the stripes of (δ_1, δ_3) values, inside which inference based on the narrow exponential model is better than using the broader three parameter model, for each of the four focus parameters.

(d) In some more detail, relating to finding the required Q matrix, show in fact that

$$\begin{aligned} U(y) &= -(1/b)W_1, \\ V_1(y) &= W_2, \\ V_2(y) &= (\log b)W_1 - W_3, \end{aligned}$$

in terms of

$$\begin{aligned} W_1 &= y - 1, \\ W_2 &= \log y - \psi(1), \\ W_3 &= y \log y - \{1 + \psi(1)\}, \end{aligned}$$

where y is standard exponential, and ψ is the digamma function (derivative of the log-gamma function); in particular, $\psi(1) = -\gamma_e$, where $\gamma_e = 0.5772...$ is the Euler constant. Write

$$\begin{pmatrix} U \\ V_1 \\ V_2 \end{pmatrix} = \begin{pmatrix} -1/b & 0 & 0 \\ 0 & 1 & 0 \\ \log b & 0 & -1 \end{pmatrix} \begin{pmatrix} W_1 \\ W_2 \\ W_3 \end{pmatrix} = AW,$$

with consequent formulae

$$J = \operatorname{Var} AW = AKA^{\mathrm{t}} \quad \text{and} \quad J^{-1} = (A^{\mathrm{t}})^{-1}K^{-1}A^{-1}$$

for J and its inverse. Then show that

$$Q = J^{11} = \begin{pmatrix} k^{22} & -k^{23} \\ -k^{23} & k^{33} \end{pmatrix} = \begin{pmatrix} 4.7616 & -5.8327 \\ -5.8327 & 10.5944 \end{pmatrix},$$

in terms of the elements $k^{i,j}$ of K^{-1}.

(e) Show that the ω vector in fact is equal to $(0,0)^{\mathrm{t}}$ for the special case of the focus parameter being $\mu = \mathrm{E}\, Y$, the mean. What are the consequences for inference for the mean?

(f) For the $n = 799$ nerve impulse data, find the ML estimates of the three parameters, and translate these to estimates of $\delta_1 = \sqrt{n}(a-1)$ and $\delta_3 = \sqrt{n}(c-1)$. Where are these estimates positioned, in relation to Figure 7? What are the anticipated consequences for model selection issues?

11. A variance heteroskedastic normal regression model

This exercise is concerned with a general version of the heteroskedastic regression model used in Exercise 6. The widest model takes the form

$$y_i = x_i^{\mathrm{t}}\beta + z_j^{\mathrm{t}}\gamma + \sigma \exp(\kappa v_i)\varepsilon_i \quad \text{for } i = 1, \ldots, n,$$

where the ε_i are i.i.d. and standard normal. Here x_i, z_i, v_i are covariates: x_i protected, of dimension p; z_i open or non-protected, of dimension q; and v_i at the outset of dimension 1 (though the generalisation to more than one κ parameter is reasonably straightforward).

Often v_i would be one of the x_is, possibly centred; see below. Thus the wide model has $p+q+2$ parameters, and the natural narrow model

$$y_i = x_i^{\mathrm{t}}\beta + \sigma\varepsilon_i$$

has $p+1$ parameters. The statistician may wish to decide which of the γ_j parameters need to be taken into the regression surface, and whether κ needs to be taken on board or not.

(a) Explain that we may assume that $\bar{v} = n^{-1}\sum_{i=1}^n v_i = 0$, without loss of generality. This yields a more easy interpretation of σ, as the standard deviation for y_i 'in the middle' of v_i space. It also aids numerical precision.

(b) Show that the values of (β, γ) that maximise the wide models likelihood, for given κ, are of the weighted least squares estimator form

$$\begin{aligned}\begin{pmatrix}\widehat{\beta}(\kappa)\\ \widehat{\gamma}(\kappa)\end{pmatrix} &= \Big\{n^{-1}\sum_{i=1}^n \frac{1}{\exp(2\kappa v_i)}\begin{pmatrix}x_i\\ z_i\end{pmatrix}\begin{pmatrix}x_i\\ z_i\end{pmatrix}^{\mathrm{t}}\Big\}^{-1}\Big\{n^{-1}\sum_{i=1}^n \frac{1}{\exp(2\kappa v_i)}\begin{pmatrix}x_i\\ z_i\end{pmatrix} y_i\Big\}\\ &= \{(XZ)^{\mathrm{t}}W(\kappa)^{-1}(XZ)\}^{-1}(XZ)^{\mathrm{t}}W(\kappa)^{-1}Y.\end{aligned}$$

Here (XZ) is the `cbind(X,Z)` covariate matrix of size $n \times (p+q)$, and $W(\kappa)$ is the $n \times n$ diagonal matrix with elements $\exp(2\kappa vi)$.

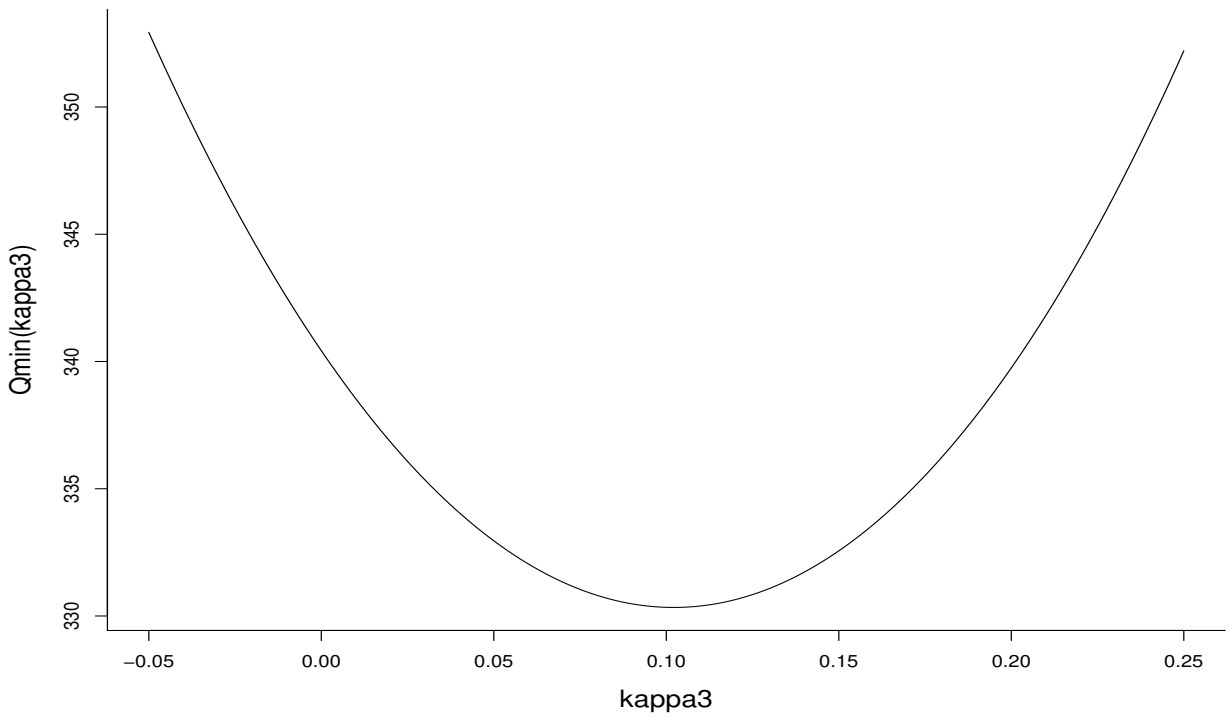

Figure 8: Plot of $Q_{\min}(\kappa_3)$ to find the ML estimate of κ_3 in the 3plus3 model. Note that $\kappa_3 = 0$ corresponds to the usual linear regression model.

(c) Show that the ML estimator of κ is the value minimising the function

$$Q_{\min}(\kappa) = \sum_{i=1}^{n} \frac{1}{\exp(2\kappa v_i)} \{y_i - x_i^{\mathrm{t}} \widehat{\beta}(\kappa) - z_i^{\mathrm{t}} \widehat{\gamma}(\kappa)\}^2,$$

and that ML estimator formulae for the other parameters become

$$\widehat{\beta} = \widehat{\beta}(\widehat{\kappa}), \quad \widehat{\gamma} = \widehat{\gamma}(\widehat{\kappa}), \quad \widehat{\sigma} = (Q_{\min}/n)^{1/2} = (Q_{\min}(\widehat{\kappa})/n)^{1/2}.$$

Usually the full log-likelihood function can be maximised directly, via e.g. `nlm`, but in cases where this is meeting numerical problems the above gives a recipe that is reduced to a one-dimensional search (finding the κ estimate, and then the others directly), rather than by a search in $(p+q+2)$-dimensional space.

(d) Consider the Adelskalenderen data of Exercise 6, where y_2 is to be modelled in terms of y_1, y_3, y_4, where we now take y_1 protected (each candidate model needs to include that variable). For numerical stability reasons with some of the models we take the trouble to centre the three explanatory variables into

$$y_1^* = y_1 - \bar{y}_1, \quad y_3^* = y_3 - \bar{y}_3, \quad y_4^* = y_4 - \bar{y}_4$$

first (we may of course afterwards back-transform regression coefficients to the original y_1, y_3, y_4 scale, if deemed convenient). We contemplate using any of the six models

0, linear regression with only y_1^*;
3, linear regression with y_1^* and y_3^*;
4, linear regression with y_1^* and y_4^*;
34, linear regression with y_1^* and y_3^* and y_4^*;
3plus3, regression with y_1^* and y_3^*, with κ_3;
34plus3, regression with y_1^* and y_3^* and y_4^*, with κ_3;

where inclusion of the κ_3 parameter means employing the heteroskedastic scheme of $\sigma_i = \sigma \exp(\kappa_3 v_{3,i})$, and with $v_{3,i} = (y_{i,3} \bar{y}_3)/\mathrm{sd}(y_3)$. For model 3plus3, there are five parameters $\beta_0, \beta_1, \beta_3, \sigma, \kappa_3$. (i) Programme the full logL function for direct maximisation, via the nlm routine; and (ii) find the ML estimates via the recipe above, which only requires minimising the $Q_{\min}(\kappa_3)$ function. The results are in essential numerical agreement, and are as follows (ML estimates, with standard error in parenthesis, for $\sigma, \beta_0, \beta_1, \beta_3, \kappa$):

$$1.149\,(0.109),\ -0.024\,(6.826),\ 2.045\,(0.073),\ 0.109\,(0.013),\ 0.102\,(0.041).$$

(e) In the situation just described, maximise likelihoods for the six models, and verify that the following table results, where the fourth column is the σ estimate in question.

```
model  dim logLmax shat     AIC      BIC
0       3 -488.143 1.705 -982.286 -992.851
3       4 -393.326 1.167 -794.653 -808.739 good
4       4 -433.219 1.369 -874.439 -888.525
34      5 -392.630 1.164 -795.260 -812.868 good
3plus3  5 -389.566 1.149 -789.131 -806.738 best
34plus3 6 -388.765 1.146 -789.529 -810.658 good
```

(f) For the Adelskalenderen data we do find a numerical estimate of the 6×6 information matrix $J_{n,\text{wide}}$ directly from data, via the Hessian matrix associated with using the `nlm` routine, but it is useful to find general mathematical expressions, also for understanding the behaviour of ML estimators. Start from

$$\log f_i = -\log\sigma - \kappa v_i - \tfrac{1}{2}\frac{1}{\sigma^2}\frac{1}{\exp(2\kappa v_i)}(y_i - x_i^{\mathrm{t}}\beta - z_i^{\mathrm{t}}\gamma)^2 - \tfrac{1}{2}\log(2\pi),$$

take partial derivatives, and deduce that

$$\begin{aligned}
\partial \log f_i/\partial\sigma &= (1/\sigma)(\varepsilon_i^2 - 1),\\
\partial \log f_i/\partial\beta &= (1/\sigma)x_i\varepsilon_i/\exp(\kappa v_i),\\
\partial \log f_i/\partial\gamma &= (1/\sigma)z_i\varepsilon_i/\exp(\kappa v_i),\\
\partial \log f_i/\partial\kappa &= v_i(\varepsilon_i^2 - 1),
\end{aligned}$$

Writing U_i for the two first and V_i for the two last of these score vector components, show from this that

$$J_{n,\text{wide}} = n^{-1}\sum_{i=1}^n \mathrm{Var}\begin{pmatrix} U_i \\ V_i \end{pmatrix} = \begin{pmatrix} 2/\sigma^2 & 0 & 0 & 0 \\ 0 & \Sigma_{n,00}/\sigma^2 & \Sigma_{n,01}/\sigma^2 & 0 \\ 0 & \Sigma_{n,10}/\sigma^2 & \Sigma_{n,11}/\sigma^2 & 0 \\ 0 & 0 & 0 & 2s_{n,v}^2 \end{pmatrix},$$

where

$$\Sigma_n = \Sigma_n(\kappa) = n^{-1}\sum_{i=1}^n \frac{1}{\exp(2\kappa v_i)}\begin{pmatrix} x_i \\ z_i \end{pmatrix}\begin{pmatrix} x_i \\ z_i \end{pmatrix}^{\mathrm{t}}$$

and $s_{n,v}^2 = n^{-1}\sum_{i=1}^n v_i^2$ is the empirical variance of the v_i.

(g) Deduce from this that $\widehat{\sigma}$, $\widehat{\kappa}$ and $(\widehat{\beta},\widehat{\gamma})$ are asymptotically independent, unbiased and normal, with

$$\begin{aligned}
\mathrm{Var}\,\widehat{\sigma} &\doteq \sigma^2/(2n),\\
\mathrm{Var}\,\widehat{\kappa} &\doteq (\tfrac{1}{2}/s_{n,v}^2)(1/n),\\
\mathrm{Var}\begin{pmatrix} \widehat{\beta} \\ \widehat{\gamma} \end{pmatrix} &\doteq \frac{\sigma^2}{n}\Sigma_n(\kappa)^{-1}.
\end{aligned}$$

Also, a useful and simple test for the hypothesis $H_0\colon \kappa = 0$ of variance homoskedasticity is to reject when $\sqrt{2}|\sqrt{n}\widehat{\kappa}| \geq 1.96$. The variance constancy hypothesis is indeed rejected for the speedskating data; modelling the standard deviation of the 1500-m time as exponentially increasing (but slowly) with 5000-m time pays off. Check with a plot that this looks reasonable.

12. FIC analysis for the variance heteroskedastic normal regression model

We continue with the framework and illustration of Exercise 11, but now turn our attention to FIC analysis.

(a) With a focus parameter $\mu = \mu(\sigma, \beta, \gamma, \kappa)$, show that the crucial ω parameter (cf. the book's Chs. 5 and 6) takes the form

$$\begin{aligned}\omega &= J_{n,10} J_{n,00}^{-1} \begin{pmatrix} \partial\mu/\partial\sigma \\ \partial\mu/\partial\beta \end{pmatrix} - \begin{pmatrix} \partial\mu/\partial\gamma \\ \partial\mu/\partial\kappa \end{pmatrix} \\ &= \begin{pmatrix} S_{n,10}\Sigma_{n,00}^{-1}\partial\mu/\partial\beta \\ 0 \end{pmatrix} - \begin{pmatrix} \partial\mu/\partial\gamma \\ \partial\mu/\partial\kappa \end{pmatrix}.\end{aligned}$$

In this connection, note a couple of practical aspects:

(i) The FIC theory developed in Ch. 6 works as long as a consistent estimator is used for the $J_{n,\text{wide}}$ matrix, which means convergence in probability to the correct limit in the local large-sample framework where γ and κ are both of size $O(1/\sqrt{n})$. We often prefer using a model-robust estimate of $J_{n,\text{wide}}$, i.e. arrived at via estimation in the wide model, which here corresponds to insert wide-model estimates $\widehat{\sigma}$ and $\widehat{\kappa}$ in the formula reached in (f) above. We may however also use a narrow-model based estimate, which corresponds to the simpler case of $\kappa = 0$ and the corresponding $\widehat{\sigma}_{\text{narr}}$.

(ii) The formula for ω above exploits the explicit form of $J_{n,\text{wide}}$ found above. Sometimes we find a numerical estimate of this matrix via the Hessian matrix associated with the `nlm` routine, which is consistent and hence fully satisfactory for the FIC machinery to work, but which does not agree fully with the $J_{n,\text{wide}}(\widehat{\sigma}, \widehat{\kappa})$ form, e.g. regarding some of the zeros. This is not a question of numerical accuracy, but of some statistics having certain mean values (resulting in certain zeros in the $J_{n,\text{wide}}$ formula), whereas the sample-based information matrix $J_{n,\text{hessian}}$ contains values of these statistics that may be close to zero, but not real zeros. In particular, if such a $J_{n,\text{hessian}}$ is used, then the numerical result for ω may agree approximately, but not fully, with the formula above.

(b) Also show that the minimal standard deviation parameter τ_0, in the notation of Ch. 6, becomes

$$\tau_0 = \sigma\{\tfrac{1}{2}(\tfrac{\partial\mu}{\partial\sigma})^2 + (\tfrac{\partial\mu}{\partial\beta})^{\mathrm{t}}\Sigma_{n,00}^{-1}\tfrac{\partial\mu}{\partial\beta}\}^{1/2}.$$

Also demonstrate that the $Q_n = J_{n,\text{wide}}^{11}$ matrix takes the form

$$Q_n = \begin{pmatrix} \sigma^2\Sigma_n^{11} & 0 \\ 0 & 1/(2s_{n,v}^2) \end{pmatrix},$$

where Σ_n^{11} is the appropriate lower right-hand $q \times q$ block of $\Sigma_n(\kappa)^{-1}$. There is again a choice between the model-robust version, which uses $\widehat{\kappa}$ for this matrix, and the narrow-based version, which simply uses $\kappa = 0$; both versions are satisfactory, though we typically favour the model-robust scheme.

13. A method for 'stretching' a given parametric model

Suppose $f(y, \theta)$ is a parametric model for a density of certain data $y_1, \ldots, y_n$ with associated cumulative distribution function $F(y, \theta)$; here θ is some p-dimensional parameter. Let furthermore $h(x, \gamma)$ be any probability density on $[0, 1]$, with cumulative $H(x, \gamma)$, with a certain inner parameter value $\gamma = \gamma_0$ corresponding to the uniform, i.e.

$$H(x, \gamma_0) = x \quad \text{and} \quad h(x, \gamma_0) = 1 \quad \text{for } x \in [0, 1].$$

We shall see how the start family can be 'stretched' via the h density, increasing the dimension of the model from p to $p + 1$.

(a) Explain why

$$G(y, \theta, \gamma) = H(F(y, \theta), \gamma)$$

defines a proper cumulative distribution function, and show that its density becomes

$$g(y, \theta, \gamma) = f(y, \theta) h(F(y, \theta), \gamma).$$

Note that this is simply the old $f(y, \theta)$ for the case when $\gamma = \gamma_0$.

(b) Show that the log-derivatives of g become

$$\begin{aligned} \partial \log g / \partial \theta &= u(y, \theta) + \partial \log h(F(y, \theta), \gamma) / \partial \theta, \\ \partial \log g / \partial \gamma &= \partial \log h(F(y, \theta), \gamma) / \partial \gamma, \end{aligned}$$

where $u(y, \theta)$ is the score function of the start model. Explain furthermore that at the null model, where $\gamma = \gamma_0$, then

$$\begin{aligned} \partial \log g / \partial \theta &= u(y, \theta), \\ \partial \log g / \partial \gamma &= v(F(y, \theta)), \end{aligned}$$

in terms of

$$v(x) = \partial \log h(x, \gamma_0) / \partial \gamma.$$

With notation and development as in the book's Chs. 5 and 6, this defines

$$J = \mathrm{Var} \begin{pmatrix} u(Y, \theta) \\ v(F(Y, \theta)) \end{pmatrix} = \begin{pmatrix} J_{00} & J_{01} \\ J_{10} & J_{11} \end{pmatrix},$$

computed at $\gamma = \gamma_0$, with consequences e.g. for how much γ disturbance the start model can tolerate. Note that the J_{00} here is the Fisher information matrix for the start model.

(c) As an example, consider

$$H(x, \gamma) = 1 - (1 - x)^\gamma \quad \text{with } h(x, \gamma) = \gamma (1 - x)^{\gamma - 1} \quad \text{on } [0, 1],$$

with $\gamma_0 = 1$ yielding uniformity. Show that $v(x) = 1 + \log(1-x)$. When coupled with the normal, to create the 'extended normal' density

$$\gamma\Big\{1 - \Phi\Big(\frac{y-\xi}{\sigma}\Big)\Big\}^{\gamma-1}\phi\Big(\frac{y-\xi}{\sigma}\Big)\frac{1}{\sigma},$$

show that

$$J = \mathrm{Var}\begin{pmatrix}(\varepsilon^2-1)/\sigma\\ \varepsilon/\sigma\\ v(\Phi(\varepsilon))\end{pmatrix} = \begin{pmatrix}2/\sigma^2 & 0 & a/\sigma\\ 0 & 1/\sigma^2 & b/\sigma\\ a/\sigma & b/\sigma & 1\end{pmatrix},$$

with ε a standard normal and

$$a = \mathrm{cov}\{\varepsilon^2 - 1, v(\Phi(\varepsilon))\} = -0.5956 \quad \text{and} \quad b = \mathrm{cov}\{\varepsilon, v(\Phi(\varepsilon))\} = -0.9032$$

(with numbers arrived at via numerical integration). Here I have taken the parameters in order of σ, ξ, γ.

(d) Letting $\tau^2 = \mathrm{Var}\, v(\Phi(\varepsilon))$, show that the tolerance radius, in the terminology of Ch. 5, is $\kappa/\sqrt{n}$, with

$$\kappa^2 = 1/(\tau^2 - \tfrac{1}{2}a^2 - b^2) = 12.0879^2.$$

Note that $X = \Phi(\varepsilon)$ simply has a uniform distribution, and that in fact $\tau = 1$.

(e) As a second example, consider

$$H(x,\gamma) = \frac{x}{\gamma - (\gamma-1)x} \quad \text{with} \quad h(x,\gamma) = \frac{\gamma}{\{\gamma-(\gamma-1)x\}^2}.$$

Show that this leads to the extended density

$$g(y,\theta,\gamma) = \frac{\gamma}{\{\gamma-(\gamma-1)F(y,\theta)\}^2}f(y,\theta),$$

where again $\gamma = 1$ gives back the start family, and that $v(x) = 2x - 1$. Show that this leads to a J matrix as above (the information matrix, computed at the null model), but now with $a = 0$ and $b = 0.5642$, resulting also in a lower tolerance radius, with $\kappa = 8.1596$.

(f) A third example takes

$$H(x,\gamma) = \Phi(\gamma\Phi^{-1}(x)) \quad \text{with} \quad h(x,\gamma) = \frac{\gamma\phi(\gamma\Phi^{-1}(x))}{\phi(\Phi^{-1}(x))}.$$

Show that this properly defines a cumulative distribution function. This stretch mechanism has the property that $H(\frac{1}{2},\gamma) = \frac{1}{2}$ for each γ, which means that the median remains invariant, i.e. does not change from its original value when the stretching is applied. Show that $v(x) = 1 - \Phi^{-1}(x)^2$.

(g) Explain why the third method above does not work for the case of starting with the normal distribution – but it may work very well for other families. Show that this strategy in general terms leads to

$$g(y, \theta, \gamma) = f(y, \theta)\gamma \exp\{\tfrac{1}{2}(1-\gamma^2)\Phi^{-1}(F(y, \theta))^2\},$$

and demonstrate that indeed it integrates to 1. Try out the method by plotting the densities for γ values close to 1, for some densities of a familiar model (see Figure 9). Also compute the tolerance threshold for the case of such stretching of the gamma distribution (where your answer will depend on the a parameter of the gamma).

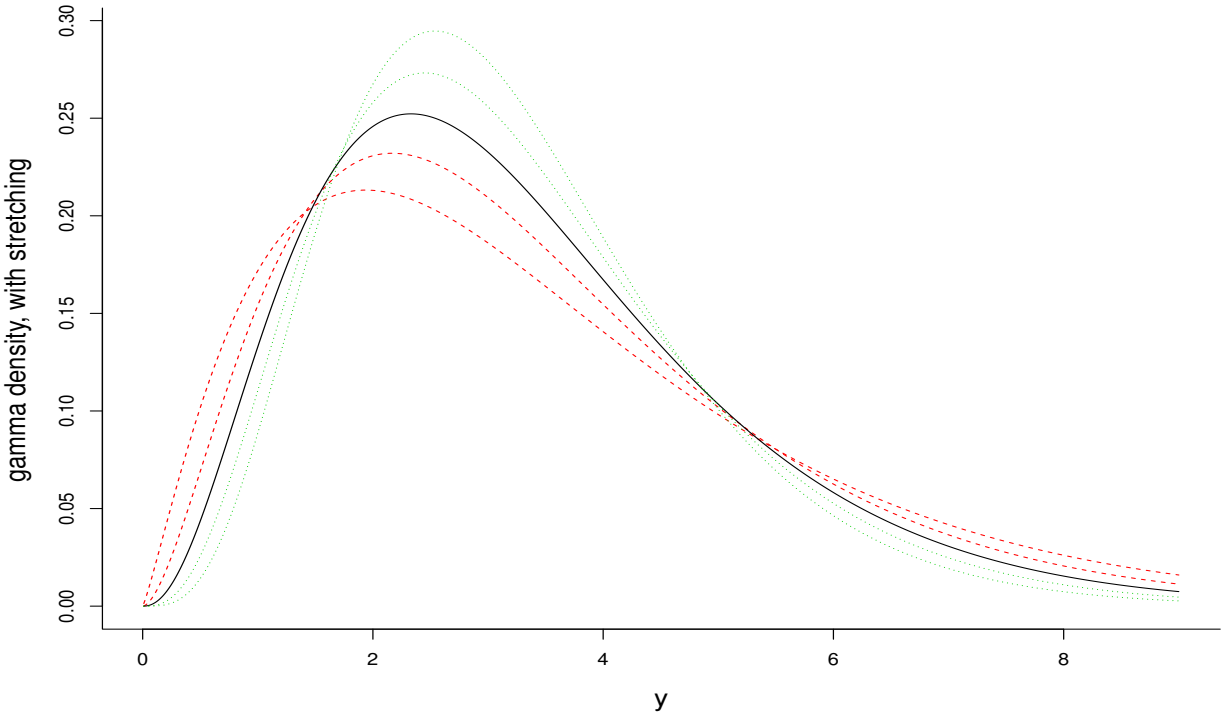

Figure 9: The gamma distribution density (full line, in the middle) along with four gamma-stretched versions, following the method of (f) and (g), with γ equal to 0.80, 0.90 (dashed lines, lower maxima), 1.10, 1.20 (dotted lines, higher maxima). Each of the densities have the same median. I have used $(a, b) = (3.33, 1)$ for the gamma density parameters.

(h) Generalise and modify the relevant results above to the situation where $H(x, \gamma)$ has a two-dimensional rather than merely a one-dimensional parameter.

14. Stretching the normal error distribution in regression models

We consider the linear regression model, of the familiar form

$$y_i = x_i^{\mathrm{t}}\beta + \sigma\varepsilon_i \quad \text{for } i = 1, \ldots, n,$$

but where we now intend to let the ε_i come from a model more general than the normal, following the model stretching methodology of Exercise 13.

(a) Specifically, assume that y_i is drawn from a distribution with cumulative distribution function

$$G(y, \xi_i, \sigma) = H(\Phi((y - \xi)/\sigma), \gamma), \quad \text{with } \xi_i = x_i^{\mathrm{t}}\beta,$$

with $H(x, \gamma)$ being a distribution function on the unit interval, and with $\gamma = \gamma_0$ corresponding to $H(x, \gamma_0) = x$. Show that this also corresponds to y_i having the density

$$g(y, \xi_i, \sigma)\phi\Big(\frac{y - \xi_i}{\sigma}\Big)\frac{1}{\sigma}h\Big(\Phi\Big(\frac{y - \xi_i}{\sigma}\Big), \gamma\Big).$$

(b) As far as information calculus in the wide model is concerned, under narrow model conditions, show that

$$J_n = n^{-1}\sum_{i=1}^{n} \mathrm{Var}\begin{pmatrix} (\varepsilon_i^2 - 1)/\sigma \\ \varepsilon_i x_i/\sigma \\ v(\Phi(\varepsilon_i)) \end{pmatrix} = \begin{pmatrix} 2/\sigma^2 & 0 & a/\sigma \\ 0 & \Sigma_n/\sigma^2 & b\bar{x}/\sigma \\ a/\sigma & b\bar{x}/\sigma & \tau^2 \end{pmatrix},$$

in the notation of Exercise 13(c)–(d); also, $\Sigma_n = n^{-1}\sum_{i=1}^n x_i x_i^{\mathrm{t}}$ and $\bar{x} = n^{-1}\sum_{i=1}^n x_i$. Deduce that the tolerance radius is $\kappa/\sqrt{n}$, where

$$\kappa^2 = (\tau^2 - \tfrac{1}{2}a^2 - b^2\bar{x}^{\mathrm{t}}\Sigma_n^{-1}\bar{x})^{-1}.$$

(c) Explore the range of possible values of κ, as dictated by the covariate distribution and its consequent size of $\bar{x}^{\mathrm{t}}\Sigma_n^{-1}\bar{x}$ above. There are cases where the normal model has large tolerance against such stretching, and other situations where it tolerates only moderate amounts of stretching.

15. Log-linear expansion stretching of a parametric model

Another approach for 'stretching' a given start model is via a log-linear expansion in certain basis functions. To exemplify, consider the functions

$$\psi_j(u) = \sqrt{2}\cos(j\pi u) \quad \text{for } u \in [0, 1],$$

for $j = 1, 2, 3, \ldots$, supplemented also with the unit function $\psi_0(u) = 1$.

(a) Show that these are orthonormal with respect to the uniform distribution, i.e. that

$$\int_0^1 \psi_j(u)^2\,\mathrm{d}u = 1 \quad \text{and} \quad \int_0^1 \psi_j(u)\psi_k(u)\,\mathrm{d}u = 0 \quad \text{for } j \neq k.$$

(b) For a given parametric model with density $f(y, \theta)$ and cumulative distribution $F(y, \theta)$, define the mth order expansion as

$$f_m(y, \theta, a) = f(y, \theta)\exp\Big\{\sum_{j=1}^{m} a_j\psi_j(F(y, \theta))\Big\}\Big/k_m(a),$$

where the integration constant is

$$k_m(a) = k_m(a_1, \ldots, a_m) = \int_0^1 \exp\Big\{\sum_{j=1}^m a_j \psi_j(u)\Big\}\, \mathrm{d}u.$$

Show that this actually defines a proper density, with the same support as that of the start model.

(c) Show by Taylor expansion that

$$k_m(a) = 1 + \tfrac{1}{2}\|a\|^2 + o(\|a\|^2),$$

giving a good approximation when the coefficients $a_1, \ldots, a_m$ are small.

(d) Show that the information matrix of the extended model, computed under the narrow null model where $a = 0$, takes the form

$$J_m = \mathrm{Var}\begin{pmatrix} u(y,\theta) \\ \psi(F(y,\theta)) \end{pmatrix} = \begin{pmatrix} J_{00} & J_{01} \\ J_{10} & J_{11} \end{pmatrix},$$

with J_{01} containing the covariances between $u_j(y,\theta)$ and $\psi_k(F(y,\theta))$, computed under the null model.

(e) Investigate the special case of a log-linear expansion of the normal model,

$$f_m(y,\xi,\sigma,a) = \phi\Big(\frac{y-\xi}{\sigma}\Big)\frac{1}{\sigma}\exp\Big\{\sum_{j=1}^m a_j \psi_j\Big(\Phi\Big(\frac{y-\xi}{\sigma}\Big)\Big)\Big\}\Big/ k_m(a).$$

Find an expression for J_m and hence for the consequent $Q_m = J_m^{11}$ matrix.

(f) Compare the model extension strategy of this exercise with those of Exercises 13 and 14.

16. FIC analysis for mothers and babies

For the 189 babies and mothers, let as before $x_1 = 1$; x_2 = mother's weight (in kg) before pregnancy; z_1 = age; z_2 indicator for race = 2; z_3 indicator for race = 3, with race being 1 (white), or 2 (black), or 3 (other). Keep x_1, x_2 protected and z_1, z_2, z_3 open, with $2^3 = 8$ submodels.

For each focus parameter μ given below: compute all eight estimates; estimate the mse; compute the FIC score; and give a FIC plot (with FIC or estimated mse on x axis and estimates on y axis). For μ, take

(i) probability p_{white} of low birthweight, for a white mother, of age 33, with weight 55;

(ii) probability p_{black}, for a black mother, also of age 33 and with weight 55;

(iii) the ratio $p_{\mathrm{black}}/p_{\mathrm{white}}$.

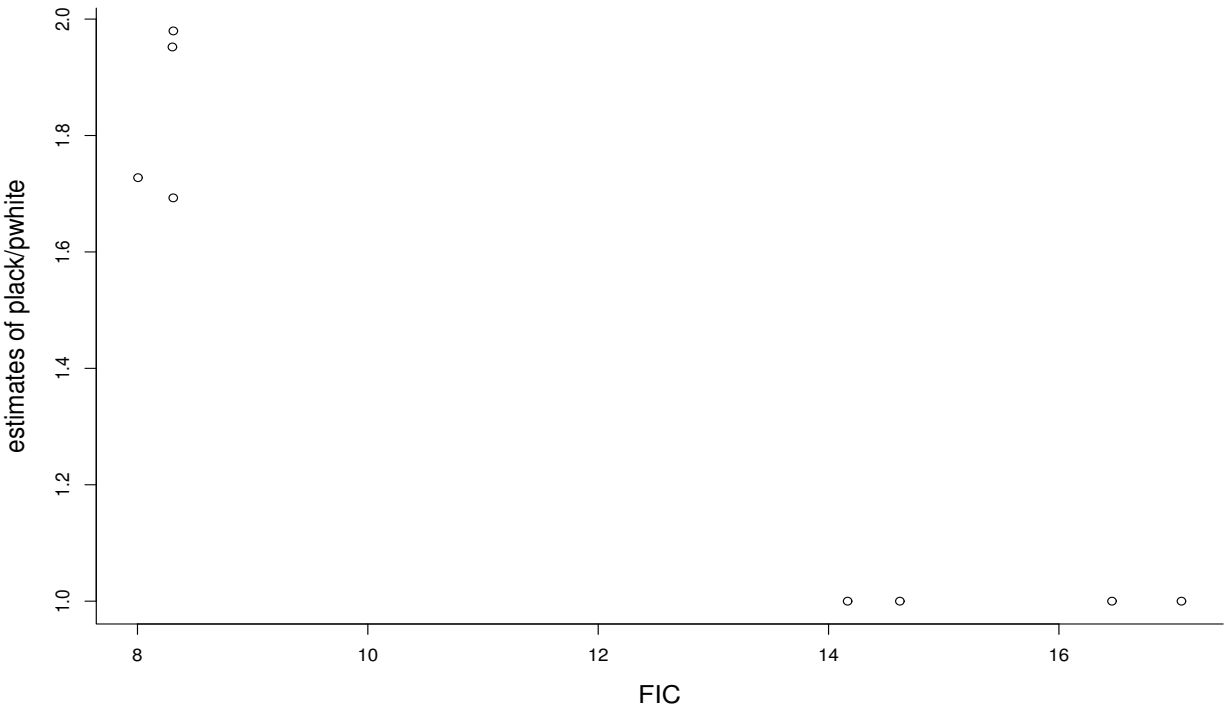

Figure 10: FIC plot for the ratio parameter $p_{\text{black}}/p_{\text{white}}$, for the eight candidate models.

FIC analysis should typically include both a table, as here, and a FIC plot, with FIC score on the horizontal axis (possibly transformed, e.g. to the root mean squared error scale, as here) and estimates on the vertical one; see Figure 10, pertaining to the third focus parameter, the ratio $p_{\text{black}}/p_{\text{white}}$. Here the four 'bad' estimates (as measured by FIC) are those equal to 1, and the four good ones take values between 1.693 and 1.980.

model	dim	estim	sd	bias1	bias2	rmse1	rmse2	
0	2	1.000	0.955	17.037	17.037	17.064	17.064	
1	3	1.000	2.194	14.453	14.453	14.619	14.619	
2	3	1.693	7.626	3.303	3.303	8.311	8.311	good
3	3	1.000	1.263	16.412	16.412	16.461	16.461	
12	4	1.728	7.665	2.307	2.307	8.005	8.005	best
13	4	1.000	2.301	13.977	13.977	14.165	14.165	
23	4	1.952	8.303	-0.236	0.000	8.300	8.303	good
123	5	1.980	8.311	0.000	0.000	8.311	8.311	good

The columns of the FIC table are respectively the model; its dimension; the parameter estimate; the estimated standard deviation; the estimated bias (in two versions); and the resulting root mean squared errors (in two version). Bias, standard deviation, root mean squared error are here computed in the limit experiment (i.e. on the $\sqrt{n}(\widehat{\mu}-\mu)$ scale), and the two bias estimates stem from taking the signed squared root of the unbiased and truncated estimate of squared bias.

17. FIC analysis for predicting y2 from y1, y3, y4

For the $n = 250$ best speedskaters on the Adelskalenderen, with results y_1, y_2, y_3, y_4, we study linear regressions of y_2 with respect to y_1, y_3, y_4, with y_1 protected, with a total of four candidate models (0, 3, 4, 34); cf. Nils Exercise 6.

(a) For each focus parameter μ of interest, compute the four estimates; estimate the root-mse (root mean squared error); compute FIC score; and display a FIC plot. For μ, take
   (i) the expected y_2, for a skater with 35.00, 6:20.00, 13:35.00;
   (ii) the probability that Øystein Grødum will skate a 1500-m at 1:48.00 or better (his personal bests, as we know, are 39.10, 6:15.50, 12:56.38).

(b) Extend your analyses to include the 13plus3 model of Exercise 6. Note that once a bigger model is being introduced, like here, the full FIC analysis needs to be revised, since FIC scores change (in interpretation and also in numerical values) also for the smaller models.

18. FIC analysis for polynomial regression

Going back to Exercise 2, with $n = 250$ simulated points from a certain nonlinear regression structure, let the list of candidate models be those corresponding to polynomial regressions of order 2, 3, 4, 5, 6 (with narrow = order two and wide = order six). Carry out FIC analysis for two estimands, (i) $\mu = \mathrm{E}(Y \,|\, x_0)$ for $x_0 = 0.75$; (ii) $\mu = \Pr\{Y > y_0 \,|\, x_0\}$ for $x_0 = 0.75$ and $y_0 = 3.0$.

19. FIC and AFIC for logistic regressions of increasing order

The following relates to the 'onset of menarche' data set from the books webpage (see Example 6.2), pertaining to 3918 Warszawa girls, and consisting of independent binomial data of the form $y_j \sim \mathrm{Bin}(m_j, p(x_j))$ for 25 age groups with mid-age numbers $x_1, \ldots, x_{25}$ ranging from 9.21 to 17.58. The candidate models we shall consider are logistic regression models of polynomial order 1, 2, 3, 4, i.e. from the narrow models $p(x) = H(\beta_0 + \beta_1 z)$ to the wides model $p(x) = H(\beta_0 + \beta_1 z + \beta_2 z^2 + \beta_3 z^3 + \beta_4 z^4)$, with H the logistic transform $\exp(u)/\{1 + \exp(u)\}$. Also, we transform from x to $z = x - 13.0$ for numerical stability, as in the book.

(a) Plot the estimated onset distribution curves corresponding to models of order one and four, along with the raw data estimates y_j/m_j at positions x_j; see Figure 11. Note that the fourth order model appears to follow the raw data better for the lower age range.

(b) Compute AIC and BIC scores.

(c) Carry out FIC analysis for focus parameter $\mu = p(x_0)$, with $x_0 = 11$ yr and $x_0 = 15$ yr.

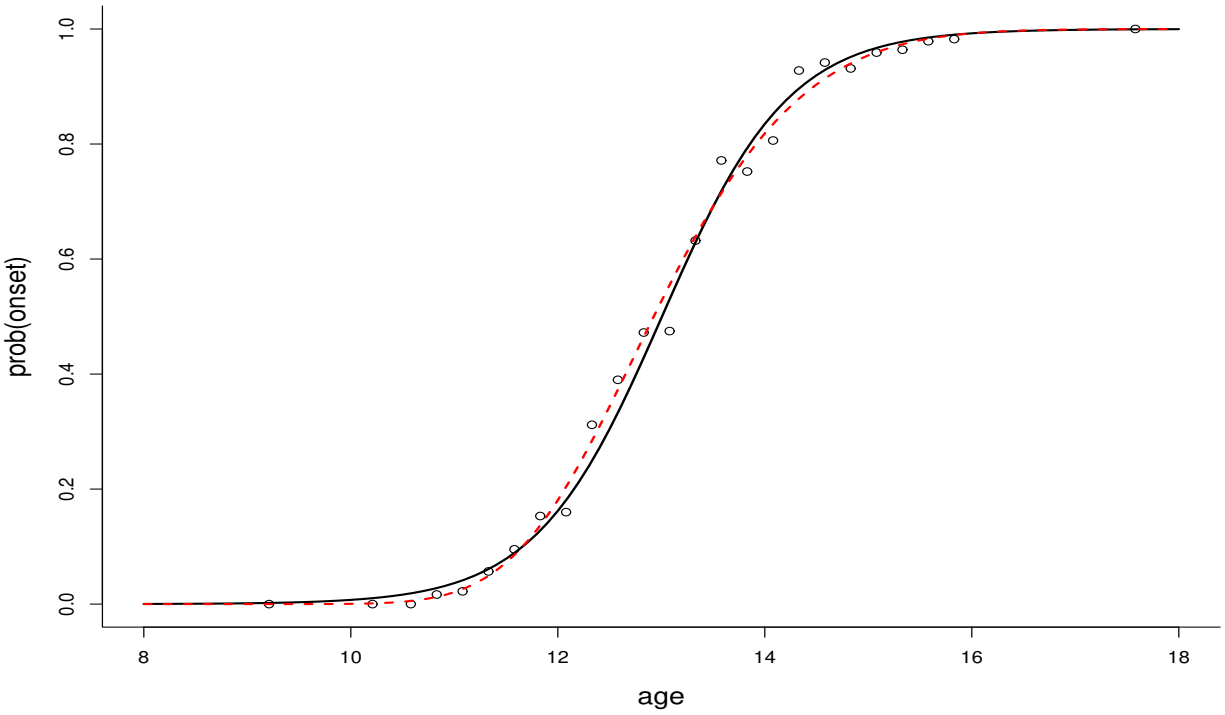

Figure 11: Onset of menarche distribution for Polish girls, via raw data estimates y_j/m_j and estimated logistic regressions of order one (full line) and four (dotted line).

(d) Carry out AFIC analysis for $\mu = p(x_0)$, with x_0 taking on 25 values uniformly spread from 11 yr 0 mnths to 13 yr 0 mnths, and with each such x_0 having the same importance. This requires the computation of (i) each ω_k vector, for $k = 1, \ldots, 25$; and (ii) the average weighted matrix

$$A = \int \omega(u)\omega(u)^{\mathrm{t}}\, W(\mathrm{d}u) = \sum_{k=1}^{25} w_k \omega_k \omega_k^{\mathrm{t}},$$

cf. Section 6.9 in the book. The first formula here is the generally valid one, which on this occasion, with a finite number of foci and with equal weights $\omega_k = 1/25$, becomes equal to the second expression.

(e) Then perform FIC analysis for μ being the 0.95 quantile of the onset distribution for menarche. The technical difficulty here is to find a way of estimating the ω vector.

20. FIC and AFIC for Poisson regression models

Find the 'birds on islands' data set from the book's web page, and let y be the number of bird species, x = distance from Ecuador (in km), z_1 = area (in thousands of sq km), z_2 = elevation (in thousands of m), z_3 = distance to nearest island (in km). Consider the eight Poisson regression models for y in terms of keeping x protected and z_1, z_2, z_3 open covariates.

(a) Construct a table with AIC and BIC values, and comment.

(b) Carry out FIC analysis for $\mu = \mathrm{E}(y \,|\, x_0)$ for $x_{\mathrm{low}} = 100$ km and $x_{\mathrm{high}} = 1300$ km, with z_1, z_2, z_3 kept at their average values. See the plot and the table below, arrived at for the $x_{\mathrm{low}} = 100$ km case. Here I have used the root-mse scale of $\mathrm{FIC}^{1/2}/\sqrt{n}$, i.e. estimating $(\mathrm{sd}^2 + \mathrm{bias}^2)^{1/2}$ on the scale of the $\widehat{\mu}$ estimates themselves. Thus the best model for this x_{low} case is model 1 (including z_1, discarding z_2, z_3), with estimated sd 3.828 and estimated bias 0. The model ranking is different for other x_0 values.

(c) Carry out AFIC analysis, averaged across say 100 x values spread from x_{low} to x_{high}, and weighted according to a weight function proportional to x. The focus is still $\mu = \mathrm{E}(Y \,|\, x)$, again with z_1, z_2, z_3 kept at their average values.

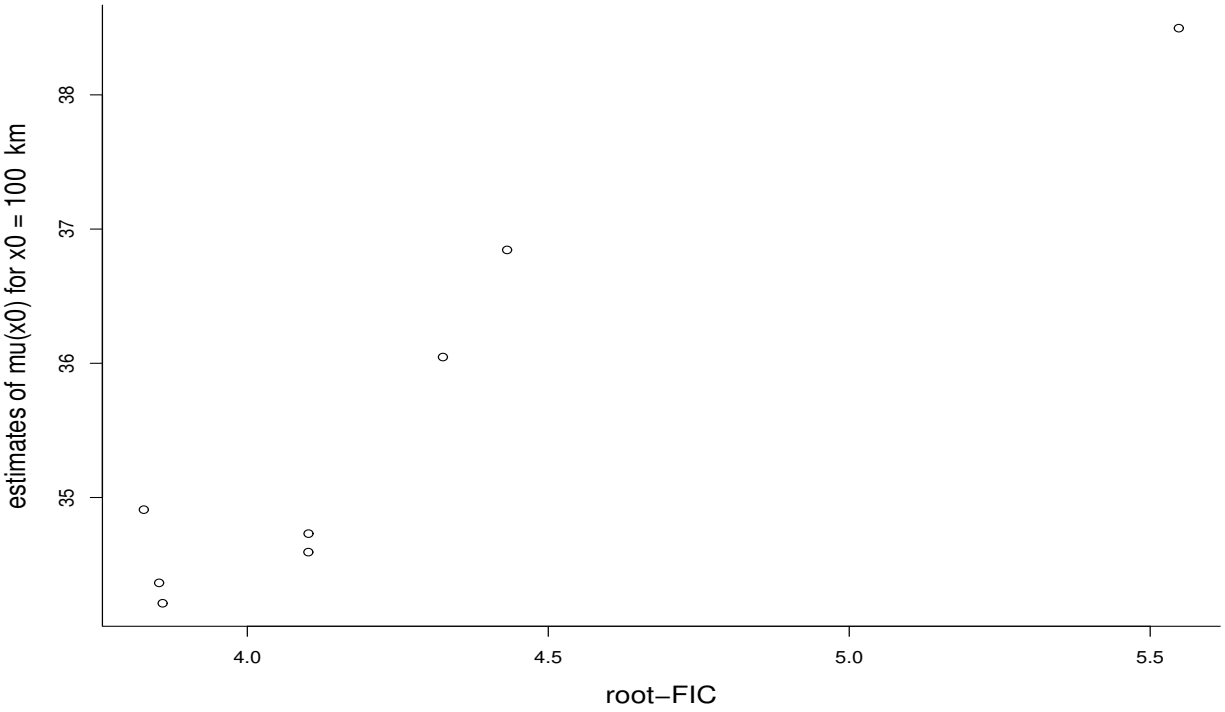

Figure 12: FIC for birds: All eight estimates of $\mathrm{E}(y \,|\, x_0)$*, for* $x_0 = 100$ *km, plotted with root-mse/*$\sqrt{n}$ *scores (root mean squared error on the scale of estimates, also equivalent to the FIC) along the horizontal axis. The more to the left in the diagram, the better is the estimate. See the table.*

model	dim	aic	muhat	sd	bias1	bias2	rmse1	rmse2	
0	2	-112.648	36.845	3.784	2.306	2.306	4.432	4.432	
1	3	-96.834	34.910	3.828	-1.436	0.000	4.089	3.828	winner
2	3	-102.166	34.364	3.853	-1.389	0.000	4.096	3.853	very good
3	3	-113.619	38.497	4.000	3.844	3.844	5.548	5.548	
12	4	-96.844	34.212	3.859	-1.334	0.000	4.083	3.859	also good
13	4	-98.817	34.731	4.099	0.142	0.142	4.102	4.102	
23	4	-103.193	36.047	4.050	1.518	1.518	4.325	4.325	
123	5	-98.768	34.593	4.101	0.000	0.000	4.101	4.101	

21. Simulation experiments with 50-50 AIC balance

To illustrate general issues pertaining to model selection, post-selection-estimation and post-selection-inference, etc., it is often useful to conduct well-designed simulation experiments.

(a) When there is only one extra parameter, from narrow to wide, show that AIC is large-sample equivalent to including γ if $|D_n/\kappa| \geq \sqrt{2}$, in notation of Chs. 5 and 6.

(b) Show that

$$\Pr\{\text{AIC selects wide}\} \to \text{pow}(\delta/\kappa),$$

where $\text{pow}(u) = \Pr\{|\text{N}(u,1)| \geq \sqrt{2}\}$, and that the 50-50 line, where AIC selects narrow or wide with the same probability $\frac{1}{2}$, is at $|\delta| = \kappa c_0$, with $c_0 = 1.408$.

(c) Set up a simple simulation experiment as follows, intended at having about 50-50 balance between narrow and wide. The narrow model is $y = \beta_0 + \beta_1 x + \text{noise}$, the wide is $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \text{noise}$, with x taken random from $\text{N}(0,1)$ and noise with standard deviation 1. Show that if $\beta_1 = (c_0/\sqrt{2}/\sqrt{n}$, then AIC is in the limit in the 50-50 balance situation. Simulate situations from such a model, say with $\beta_0 = 3$ and $\beta_1 = 1$, and compute estimates and 'quiet scandal' type confidence intervals for $\mu(x_0) = \text{E}(Y \mid x_0)$, with $x_0 = 2.0$. Comment on your findings. Study also compromise estimators using smooth AIC weights.

22. Estimating an onset distribution

For the problems studied in Exercise 19, concerned with the onset distribution for menarche, the information gathered was rather 'indirect' – rather than asking 'precisely how old were you, when event A first took place in your life', one merely asks 'how old are you now, and has event A ever happened to you'. This is a practical though indirect way of learning about such onset distributions, in broad terms. The present exercise looks at some theory in this regard.

(a) Suppose there are independent event times $y_1, \ldots, y_n$ following some population distribution $f(y,\theta)$. If one actually observes these y_i, therefore, one has the usual likelihood apparatus for such data, and arrives at a maximum likelihood estimator $\widetilde{\theta}$, with

$$\sqrt{n}(\widetilde{\theta} - \theta_0) \to_d \text{N}_p(0, J^{-1})$$

at the true parameter θ_0; for simplicity of discussion we here take the model to be correct. Here $J = J(\theta_0)$ is the Fisher information matrix of the model. Explain how you may compute J for the case of data y_i following the logistic model, with cumulative distribution function

$$F(y,\beta) = H(\beta_0 + \beta_1 y), \quad \text{where } H(u) = \frac{\exp(u)}{1+\exp(u)}.$$

(b) Now assume that one cannot observe the y_i directly, with access only to indirect data of the type (x_i, a_i), with

$$a_i = I\{y_i \le x_i\} = \begin{cases} 1 & \text{if } y_i \le x_i, \\ 0 & \text{if } y_i > x_i. \end{cases}$$

This corresponds to the situation above, where one asks a girl of age x_i whether the event A has taken place in her life or not. Show that the log-likelihood function may be expressed as

$$\ell_n(\theta) = \sum_{i=1}^{n} \{a_i \log p_i + (1 - a_i) \log(1 - p_i)\},$$

and that its derivative becomes

$$\ell_n^*(\theta) = \sum_{i=1}^{n} \frac{a_i - p_i}{p_i(1 - p_i)} p_i^*,$$

with

$$p_i = p_i(\theta) = F(x_i, \theta) \quad \text{and} \quad p_i^* = \frac{\partial p_i}{\partial \theta}.$$

How the x_i are generated depends on the study design; they may occur according to a certain sampling plan, or occur 'naturally' and reflect the underlying population density $f(y)$.