Exercises and Lecture Notes
STK 4060, Spring 2022

Version 0.21, 18-Feb-2022

Nils Lid Hjort

Department of Mathematics, University of Oslo

Abstract

These are Exercises and Lecture Notes for the course on time series, modelling and analysis, STK 4060 (Master level) or STK 9060 (PhD level), for the spring semester 2022. The collection will grow during the course.

1. The variance or an average of correlated variables

A classical and crucial result from traditional statistics is that if $x_1,\ldots,x_n$ are independent with the same distribution, then $\mathrm{Var}\,\bar x_n=\sigma^2/n$, for the data average $\bar x=(1/n)\sum_{i=1}^n x_i$, where σ^2 is the variance of a single observation. This is rather different for models with dependence. Suppose now that $x_1,\ldots,x_n$ is a stationary sequence, with $\mathrm{cov}(x_k,x_{i+h})=\sigma^2\rho(|h|)$, for some correlation function $\rho(h)=\mathrm{corr}(x_i,x_{i+h})$.

(a) Show that

$$\mathrm{Var}\,\bar x_n=\frac{\sigma^2}{n}\Big\{1+2\sum_{h=1}^n(1-h/n)\rho(h)\Big\}=\frac{\sigma^2}{n}\sum_{h=-n}^n(1-|h|/n)\rho(j).$$

(b) For the special case of $\rho(h)=\rho^h$, called autocorrelation of order 1, show that

$$\mathrm{Var}\,\bar x_n=\frac{\sigma^2}{n}\Big\{1+2\sum_{h=1}^{n-1}\rho^h-(1/n)\sum_{h=1}^{n-1}h\rho^h\Big\}=\frac{\sigma^2}{n}\Big\{\frac{1+\rho}{1-\rho}+O(1/n)\Big\}.$$

With a positive autocorrelation, therefore, the variance of $\bar x_n$ becomes clearly bigger than under independence.

(c) Suppose you observe such a stationary time series $x_1,\ldots,x_n$, with autocorrelation function $\rho(h)=\rho^h$ for $h=1,2,3,\ldots$, and with unknown mean μ, variance σ^2, and autocorrelation parameter $\rho\in(-1,1)$. If you do the traditional $\bar x_n\pm1.96\,s_n/\sqrt{n}$ interval for μ, recommended in 99 statistics books, with s_n the empirical standard deviation, what will be its confidence coverage level?

(d) Give estimators for μ,σ,ρ, constructed from the observed time series.

(e) Give a more careful and appropriate 95 percent confidence interval, taking autocorrelation into account. Note in particular that such a confidence interval *is wider* than the traditional one, when the autocorrelation is positive.

2. An autoregressive time series model

Construct a time series $x_1, x_2, \ldots, x_n$ as follows, via i.i.d. $\varepsilon_1, \ldots, \varepsilon_n$ being standard normal. Let $x_1 = \varepsilon_1$ and then $x_{t+1} = \rho x_t + \varepsilon_{t+1}$ for $i = t, 2, \ldots$, where ρ is a value inside $(-1, 1)$.

(a) Take $n = 100$ and $\rho = 0.345$, and simulate such a time series in your computer. Check what the `acf(xdata)` does, playing also a bit with other combinations of n and ρ.

(b) Write $\mathcal{F}_t$ for all observed history up to and including time point t. Show that $\mathrm{E}\,(x_{t+1} \,|\, \mathcal{F}_t) = \rho x_t$ and $\mathrm{Var}\,(x_{t+1} \,|\, \mathcal{F}_t) = 1$. Deduce also from this that

$$\mathrm{E}\, x_t = \rho\, \mathrm{E}\, x_{t-1} \quad \text{and} \quad \mathrm{Var}\, x_t = 1 + \rho^2\, \mathrm{Var}\, x_{t-1}.$$

Show that $\mathrm{E}\, x_t = 0$, for all t, and find a formula for the variance of x_t.

(c) Starting from

$$\begin{aligned} x_2 &= \rho\varepsilon_1 + \varepsilon_2, \\ x_3 &= \rho^2\varepsilon_1 + \rho\varepsilon_2 + \varepsilon_3, \\ x_4 &= \rho^3\varepsilon_1 + \rho^2\varepsilon_2 + \rho\varepsilon_3 + \varepsilon_4, \end{aligned}$$

find a general formula for x_t, expressed in terms of the i.i.d. components $\varepsilon_1, \ldots, \varepsilon_t$. Use this to find and explicit distribution of x_t. Also show

$$\mathrm{Var}\, X_t = 1 + \rho^2 + \rho^4 + \cdots + \rho^{2(t-1)} = \frac{1 - \rho^{2t}}{1 - \rho^2},$$

re-proving what you found in (b).

(d) Find the explicit covariance and correlation between x_i and x_{i-1}.

(e) When the time series has been at work for some time, show that

$$\mathrm{Var}\, x_i \to \frac{1}{1 - \rho^2}, \quad \mathrm{cov}(x_i, x_{i+1}) \to \frac{\rho}{1 - \rho^2}, \quad \mathrm{cov}(x_i, x_{i+2}) \to \frac{\rho^2}{1 - \rho^2},$$

etc.

(f) Show that the real acf (the autocorrelation function) becomes $1, \rho, \rho^2, \rho^3, \ldots$,

(g) Simulate a few time series using the above construction, with a few combinations of n and ρ. Verify that with n moderate-to-large, the empirical `acf(xdata)` becomes close to the real $1, \rho, \rho^2, \rho^3, \ldots$.

3. Using regression modelling for the Johnson & Johnson dataset

Consider the dataset called `jj` in the `astsa` package, giving the quarterly earnings of the J & J company, from quarter 1 1960 to quarter 4 1980. One wishes to study how these $y_1, \ldots, y_n$ evolve over time (with $n = 84$ quarters over 21 years), e.g. to predict earnings for the coming year. The task here is to go through some regression models, so to speak before factoring in correlations and specific time series aspects.

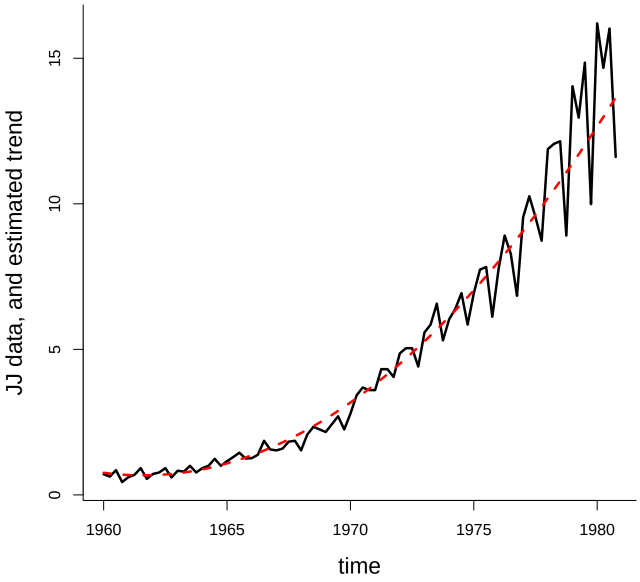

Figure 0.1: The JJ data, with estimated trend, from the five-parameter model.

(a) Write $x_t = t - 1960$, for $t = 1, \ldots, n$. Fit the rather simple classic linear regression model, with $y_t = \beta_0 + \beta_1 x_t + \varepsilon_t$, with the ε_t taken i.i.d. $\mathrm{N}(0, \sigma^2)$. Look at the fitted trend $\widehat{m}_1(t) = \widehat{\beta}_0 + \widehat{\beta}_1 x_t$, alongside data, to check that this model is far too simple. For the practice, check also the residuals $r_{1,t} = y_t - \widehat{m}_1(t)$; these will vary too much, indicating again that this model is too coarse.

(b) A rather better model is to include a quadratic term for the trend. Fit the regression model $y_t = \beta_0 + \beta_1 x_t + \beta_2 x_t^2 + \varepsilon_t$, again with the ε_t taken i.i.d. from the $\mathrm{N}(0, \sigma^2)$. Plot the estimated trend $\widehat{m}_2(t) = \widehat{\beta}_0 + \widehat{\beta}_1 x_t + \widehat{\beta}_2 x_t^2$ alongside data, examine the residuals $r_{2,t} = y_t - \widehat{m}_2(t)$, and comment on what you find.

(c) You learn from the above that the trend function is adequately described by such a parabola, but that that variance of data is not constant; it increases over time. So try the variance heteroscedastic model

$$y_t = \beta_0 + \beta_1 x_t + \beta_2 x_t^2 + \sigma_t \varepsilon_t, \quad \text{for } t = 1, \ldots, n, \quad \text{with } \sigma_t = \exp\{\gamma_0 + \gamma_1 (x_t - \bar{x})\},$$

and with the ε_t now being i.i.d. and standard normal. The model has three parameters for the mean and two for the variance. Show that the log-likelihood function for this five-parameter model can be expressed as

$$\ell(\theta) = \sum_{t=1}^{n} \{-\log \sigma_t - \tfrac{1}{2}(y_t - \beta_0 - \beta_1 x_t - \beta_2 x_t^2)^2/\sigma_t^2 - \tfrac{1}{2}\log(2\pi)\},$$

in terms of the full parameter vector θ.

(d) Find the maximum likelihood (ML) estimates, say $\widehat{\theta}_{\mathrm{ml}}$, by numerically maximising the log-likelihood function. Compute also approximate standard errors, for the five parameter estimates, via the general normal approximation theorem for parametric models,

$$\widehat\theta_{\rm ml} \approx_d {\rm N}_p(\theta,\widehat\Sigma), \quad \text{with } \widehat\Sigma=\widehat J^{-1}. \tag{0.1}$$

Here $\widehat J=-\partial^2\ell(\widehat\theta_{\rm ml})/\partial\theta\partial\theta^{\rm t}$, the Hesse matrix of second order derivatives, computed at the the ML position. Using `nlm` in `R` you get the Hesse matrix for free, along with the numerical optimisation, using something like

```
hello = nlm(minuslogL,starthere,hessian=T)
```

followed, pretty generically and very usefully, by

```
ML = hello$estimate
Jhat = hello$hessian
se = sqrt(diag(solve(Jhat)))
showme = cbind(ML,se,ML/se)
print(round(showme,4))
```

(e) Produce a version of Figure 0.1.

(f) There's at least one more very useful practical thing to learn, following from the general machinery of the Master Theorem (0.1, namely the so-called *delta method.* If one is interested in a a certain parameter, day γ, which is a function $\gamma=g(\theta)$ of the model parameters, then (i) the ML estimator is $\widehat\gamma_{\rm ml}=g(\widehat\theta_{\rm ml})$, i.e. via simple plug-in; and (ii) it is approximately a normal, with

$$\widehat\gamma_{\rm ml}\approx_d {\rm N}(\gamma,\widehat\tau^2),$$

with $\widehat\tau^2=\widehat c^{\rm t}\widehat\Sigma\widehat c$, where $\widehat c=\partial g(\widehat\theta_{\rm ml})/\partial\theta$ is the gradient of g, evaluated at the ML estimate. In `R` language, if we first programme the g as a `function`, we have

```
gammahat = g(ML)
chat = grad(g,ML)
tauhat = sqrt(chat %*% solve(Jhat) %*% chat)
```

I find it practical to include the `numDeriv` package, which has `grad` and `hessian` on board. Now try out such a machinery, by working with γ, the 0.90 quantile of the distribution for the next datapoint, in the JJ estup.

(g) Once you have the basic code up and running it is relatively easy to try out other variations of such models. Try to put in a cyclic term, perhaps $\beta_4\cos(2\pi t/4)$, and again look at both the residuals and the acf.

4. Understanding the empirical acf, under independence

Suppose $x_1,x_2,\ldots$ are really independent, with mean zero and variance one. What happens then, with the `acd(xdata)`? Below, write $\bar x_{a,b}$ for the average of values $x_a,\ldots,x_b$.

(a) Consider first $A_n=(1/n)\sum_{t=1}^{n-1}x_tx_{t+1}$. Show that A_n has mean zero and variance $(n-1)/n^2$, i.e. approximately $1/n$.

(b) Then go to the proper empirical $B_n = (1/n)\sum_{t=1}^{n-1}(x_t - \bar x_{1,n})(x_{t+1} - \bar x_{1,n})$. Show that

$$B_n = A_n - \frac{n-1}{n}\bar x_{1,n}\bar x_{1,n-1} - \frac{n-1}{n}\bar x_{1,n}\bar x_{2,n} + \frac{n-1}{n}\bar x_{1,n}^2 \doteq A_n - \bar x_{1,n}^2,$$

with $\doteq$ meaning 'good approximation, not affecting limits when n grows'.

(c) Show that B_n, like the simpler A_n, has mean zero and variance approximately equal to $1/n$. Show then that $A_n \to_{\rm pr} 0$, $B_n \to_{\rm pr} 0$, with '$\to_{\rm pr}$' denoting convergence in probability: $\Pr(|B_n| \ge \varepsilon) \to 0$ for each small ε.

(d) Since A_n is a sum of variables with the same distribution, with mean zero, and ${\rm Var}\,A_n \doteq 1/n$, it is natural to expect limiting normality, i.e. $\sqrt{n}A_n \to_d {\rm N}(0,1)$. This does *not* follow from the traditional CLTs (central limit theorems), since x_1x_2 is not independent of x_2x_3, etc. Check with the book's Appendix A.2, however, concerning CLTs for m-dependent variables, and verify that indeed $\sqrt{n}A_n \to_d 1$.

(e) From $\sqrt{n}B_n \doteq \sqrt{n}A_n - \sqrt{n}\bar x_{1,n}^2$, show that also $\sqrt{n}B_n \to_d {\rm N}(0,1)$, i.e. the same limit distribution.

(f) Now go from 1-step to 2-step, and work through the details for $A_n = (1/n)\sum_{t=1}^{n-2} x_t x_{t+2}$ and

$$B_n = \widehat\gamma(2) = (1/n)\sum_{t=1}^{n-2}(x_t - \bar x_{1,n})(x_{t+2} - \bar x_{1,n}).$$

The main things are that $\widehat\gamma(2) \to_{\rm pr} 0$, the true value of $\gamma(2)$ under independence, and that $\sqrt{n}\widehat\gamma(2) \to_d {\rm N}(0,1)$.

(g) Generalise properly to the result $\sqrt{n}\widehat\gamma(h) \to_d {\rm N}(0,1)$, for

$$\widehat\gamma(h) = (1/n)\sum_{t=1}^{n-h}(x_t - \bar x_{1,n})(x_{t+h} - \bar x_{1,n}).$$

(h) So far we've assumed variance $\sigma^2 = 1$, for simplicity of presentation and argumentation. For the general case, show that for a sequence of independent variables, with some mean μ and variance σ^2, we have $\sqrt{n}\widehat\gamma(h) \to_d {\rm N}(0,\sigma^4)$. Finally show that for

$$\widehat\rho(h) = \frac{\widehat\gamma(h)}{\widehat\gamma(0)} = (1/n)\sum_{t=1}^{n-h}\frac{(x_t - \bar x_{1,n})}{\widehat\sigma}\frac{(x_{t+h} - \bar x_{1,n})}{\widehat\sigma} = \frac{\sum_{t=1}^{n-h}(x_t - \bar x_{1,n})(x_{t+h} - \bar x_{1,n})}{\sum_{t=1}^{n}(x_t - \bar x_{1,n})^2},$$

our good friend the acf, we do have the clarifying easy good result $\sqrt{n}\widehat\rho(h) \to_d {\rm N}(0,1)$.

(i) For such a sequence of i.i.d. variables, show that when one computes the empirical acf, then

$$\Pr\{\widehat\rho(h) \in [-1.96/\sqrt{n}, 1.96/\sqrt{n}]\} \to 0.95,$$

for each lag h. This is the reason for the 'magical band' $\pm 1.96/\sqrt{n}$ provided in the standard use of `acf`.

5. A simple moving average process

Suppose $w_0, w_{\pm 1}, w_{\pm 2}, \ldots$ are i.i.d., with finite variance σ^2. Then consider the process

$$x_t = aw_{t-1} + (1-2a)w_t + aw_{t+1},$$

with a a tuning parameter. We call this a moving average process, with window length 3.

(a) Compute the variance of x_t, and also the covariance function $\gamma(h)$ and autocorrelation function $\rho(h)$. Plot the acf for a few values of a, including the equal balance case of $a = 1/3$.

(b) Then do a similar analysis for a 5-window moving average process, of the type

$$x_t = aw_{t-2} + aw_{t-1} + (1-4a)w_t + aw_{t+1} + aw_{t+2}.$$

Again, plot the acf for a few values of a, including the balanced case of $a = 1/5$.

(c) Similarly consider the case of

$$x_t = \rho^2 w_{t-2} + \rho w_{t-1} + w_t + \rho w_{t+1} + \rho^2 w_{t+2}.$$

Find the acf, and plot it, for a few values of ρ.

6. A general stationary normal time series model

Suppose $x_1, \ldots, x_n$ is a stationary normal time series, which means that the full vector has a multinormal distribution; this is also equivalent to saying that all linear combinations are normal. Assume it has mean μ, varianec σ^2, and correlation function $\rho(h) = \mathrm{corr}(x_t, x_{t+h})$.

(a) Show that the joint distrisbution of the full series is a $\mathrm{N}_n(\mu\mathbf{1}, \sigma^2 A)$, where $\mathbf{1} = (1, \ldots, 1)^{\mathrm{t}}$ is the vector of 1s, and A the $n \times n$ matrix of $\rho(s-t)$, for $s, t = 1, \ldots, n$; in particular, the diagonal elements are all 1.

(b) Using the basic definition of the multinormal joint density, show that the log-likelihood function can be written

$$\ell(\theta) = -n \log \sigma - \tfrac{1}{2} \log |A| - \tfrac{1}{2}(y - \mu\mathbf{1})^{\mathrm{t}} A^{-1}(y - \mu\mathbf{1})/\sigma^2 - \tfrac{1}{2} n \log(2\pi),$$

wiith θ the parameters involved. If the correlation function is known, then A is known, and θ comprises only μ, σ. For such a case, show that the ML estimators become

$$\widehat{\mu} = \frac{\mathbf{1}^{\mathrm{t}} A^{-1} y}{\mathbf{1}^{\mathrm{t}} A^{-1}\mathbf{1}} \quad \text{and} \quad \widehat{\sigma}^2 = \frac{Q_0}{n}, \quad \text{with } Q_0 = (y - \widehat{\mu}\mathbf{1})^{\mathrm{t}} A^{-1}(y - \widehat{\mu}\mathbf{1}).$$

Check that this leads to familiar formulae in the case of i.i.d. observations, where $A = I_n$, the identity matrix.

(c) If there is a parameter, say λ, in the correlation function, however, we need also $A = A(\lambda)$, and we have

$$\ell(\mu, \sigma, \lambda) = -n \log \sigma - \tfrac{1}{2} \log |A(\lambda)| - \tfrac{1}{2}(y - \mu\mathbf{1})^{\mathrm{t}} A(\lambda)^{-1}(y - \mu\mathbf{1})/\sigma^2 - \tfrac{1}{2} n \log(2\pi).$$

Use the above to find that the log-likelihood profile function, in λ, becomes

$$\ell_{\mathrm{prof}}(\lambda) = -n \log \widehat{\sigma}(\lambda) - \tfrac{1}{2} \log |A(\lambda)| - \tfrac{1}{2} n - \tfrac{1}{2} n \log(2\pi).$$

Here one first compputes

$$\widehat{\mu}(\lambda) = \frac{\mathbf{1}^{\mathrm{t}} A(\lambda)^{-1} y}{\mathbf{1}^{\mathrm{t}} A(\lambda^{-1}\mathbf{1}} \quad \text{and then} \quad \widehat{\sigma}^2(\lambda) = (1/n) Q_0(\lambda),$$

where

$$Q_0(\lambda) = \{y - \widehat{\mu}(\lambda)\mathbf{1}\}^{\mathrm{t}} A(\lambda)^{-1}\{y - \widehat{\mu}(\lambda)\mathbf{1}\}.$$

(d) Take e.g. $n = 100$, generate $x_1, \ldots, x_n$ from the standard normal in your computer, and fit the three-parameter model which has unknown μ, σ, λ, where the correlation function is modelled as $\rho(h) = \exp(-\lambda h) = \rho^h$, i.e. with $\rho = \exp(-\lambda)$ the 1-step correlation. Repeat the experiment a few times, to see how well the ML estimators succeed in coming close to the true values.

7. Conditional multinormal distributions

A vector $X = (X_1, \ldots, X_n)$ has the multinormal distribution, with mean ξ and covariance matrix Σ, if its density takes the form

$$f(x) = (2\pi)^{-n/2}|\Sigma|^{-1/2}\exp\{-\tfrac{1}{2}(x-\xi)^{\rm t}\Sigma^{-1}(x-\xi)\}.$$

We write $X \sim \mathrm{N}_n(\xi, \Sigma)$ to indicate this; not that the distribution is fully specified by giving the ξ and the Σ.

(a) Check that this becomes the classic formula for $\mathrm{N}(\xi, \sigma^2)$ in the one-dimensional case. In the general case, show that $Y = AX$ has distribution $\mathrm{N}_n(A\xi, A\Sigma A^{\rm t})$, if A is a $n \times n$ matrix. Show that f integrates to 1.

(b) Block X into $X_{(1)}$ and $X_{(2)}$, of lengths p, q, with $p + q = n$. Write

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix},$$

with Σ_{11} of size $p \times p$, etc. Try to show that $X_{(1)} \,|\, (X_{(2)} = x_{(2)})$ is multinormal, in dimension p, with these important formulae for conditional mean and conditional variance:

$$\begin{aligned} \mathrm{E}\,(X_{(1)} \,|\, x_{(2)}) &= \xi_{(1)} + \Sigma_{12}\Sigma_{22}^{-1}(x_{(2)} - \xi_{(2)}), \\ \mathrm{Var}\,(X_{(1)} \,|\, x_{(2)})''' &= \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}. \end{aligned}$$

In particular, the conditional mean is a linear function of $x_{(2)}$, and the conditional variance is constant.

(c) For the most simple but still interesting case of a normalised binormal distribution, show that if

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim \mathrm{N}_2(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}),$$

then $X_2 \,|\, (X_1 = x_1)$ is normal $(\rho x_1, 1 - \rho^2)$. Generalise to the case where X_1, X_2 have means ξ_1, ξ_2 and variances σ_1^2, σ_2^2.

8. Predicting x2 after having seen x1

Part of the business of time series modelling and analysis is *to predict*: what happens next? If we see x_1, what can we say about the x_2 of tomorrow? It is useful to learn from the multinormal situation.

(a) Suppose

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim \mathrm{N}_2(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}),$$

with knoen ρ, and that x_1 has been observed. In which sense is $\widehat{x}_2 = \rho x_1$ the best prediction for x_2? Give a 95 percent prediction interval for x_2, and discuss how its length is influenced by ρ.

(b) Suppose $X_1, \ldots, X_n, X_{n+1}$ have a joint multinormal distribution, as for many time series model, and that $x_1, \ldots, x_n$ are observed. Give the distribution for X_{n+1}, given $x_1, \ldots, x_n$. Give also a prediction for x_{n+1}, and a 95 percent prediction interval.

(c) Specialise the above to the case of a stationary Gaussian time series model, with mean μ, variance σ^2, and correlation function $\rho(h)$ for $h = 1, 2, 3, \ldots$. Again give a prediction, and a prediction interval, for x_{n+1}, assuming that $x_1, \ldots, x_n$ have been observed.

(d) Discuss how these formulae hold up outside the multinormal situation.

9. The AIC and the BIC

Suppose there are competing parametric models for the same dataset, of size n (the number of observed data points, or data vectors). One first fits these candidate models, say $M_1, \ldots, M_k$, by maximising their likelihoods. Writing $\ell_j(\theta_j)$ for model M_j, we find the ML estimate $\widehat{\theta}_j$ and the maximised log-likelihood value,

$$\ell_{j,\max} = \ell_j(\widehat{\theta}_j) \quad \text{for } j = 1, \ldots, k.$$

Then we define

$$\mathrm{aic}_j = 2\dim(\theta_j) - 2\ell_{j,\max} \quad \text{and} \quad \mathrm{bic}_j = \dim(\theta_j)\log n - 2\ell_{j,\max}, \tag{0.2}$$

with $\dim(\theta_j)$ the number of parameters estimated in that model. These are *the Akaike Information Criterion* and *the Bayesian Information Criterion*; see Chapters 2, 3 in Claeskens and Hjort (2008) for considerably more information. These two information criteria act as *ranking scores* for the competing models, with small values being preferred over bigger ones. Thus there is an AIC winner and a BIC winner (perhaps the same).

Note that these AIC and BIC recipes are completely general; they may be used with independent data, or for time series models with dependence, we may compare normal with non-normal models, and almost apples with bananas.

(a) Explain, in intuitive terms, why these ranking criteria make sense, balancing complexity with model fit. Explain also that the BIC places a harsher penalty on complexity (well, as long as $n \geq 8$).

(b) Suppose I have two coins, with probabilities p_a and p_b for 'krone'. I flip them 40 times each, and get 17 krone with the first and 23 krone with the second. Model 1 says that $p_a = p_b$; model 2 says that p_a and p_b are different. Which of these two models is best, according to the AIC, and to the BIC? – Note that we get answers, of the type 'model 1 is better than model 2', etc., without using formal null hypothesis tests, and there's no '0.05' business going on (well, at least not directly).

(c) Interestingly, it turns out that I have *three* coins in my skuff. I call their krone probabilities p_a, p_b, p_c, and the number of times I do get a krone, in 40 flips for each, are 17, 23, 26. Carry out AIC and BIC analysis, to rank as many as five candidate models: (i) p_a, p_b, p_c are equal; (ii) $p_a = p_b$ but different from p_c (iii) $p_a = p_c$ but different from p_b; (iv) $p_b = p_c$ but different from p_a; (v) the three are different.

(d) Suppose a certain start model has dimension k and log-likelihood maximum value $\ell_{0,\max}$, and that one contemplates extending this start model to a bigger one, with one more parameter. Assume specifically that the narrow model lies inside the bigger model. Argue that

$$\Delta = \ell_{1,\max} - \ell_{0,\max}$$

must be positive. Show that AIC thinks the extended model is a good idea, provided $\Delta > 1$. The BIC, however, thinks it's only worth the trouble if $\Delta > \frac{1}{2}\log n$. – One may show that if the narrow model holds, then $2\Delta \approx_d \chi^2_1$, so this can be used to see how likely it is to 'incorrectly', or unnecessarily, choose the bigger model, if the narrow model is already ok.

10. The AIC and the BIC for linear regression models

We now apply the general AIC and BIC schemes for comparing and ranking different linear regression models, for the same dataset, perhaps to decide on which covariates to include and which to exclude.

(a) Suppose we have regression data (z_t, x_t), for $t = 1, \ldots, n$, with x_t the main outcome (perhaps a time series) and $z_t = (z_{t,1}, \ldots, z_{t,k})^{\rm t}$ a covariate vector of length k. Consider the classical linear regression model, with

$$x_t = \beta_1 z_{t,1} + \cdots + \beta_k z_{t,k} + \varepsilon_t = z_t^{\rm t}\beta + \varepsilon_t \quad \text{for } t = 1, \ldots, n,$$

with the ε_t being i.i.d. $\mathrm{N}(0, \sigma^2)$. Show that the log-likelihood function can be written

$$\ell_k(\beta, \sigma) = -n\log\sigma - \tfrac{1}{2}Q(\beta)/\sigma^2 - \tfrac{1}{2}n\log(2\pi), \tag{0.3}$$

with subscript k for the number of covariates included in the model. Here

$$Q(\beta) = \sum_{t=1}^n \{x_t - m_t(\beta)\}^2, \quad \text{where } m_t(\beta) = \mathrm{E}\,(x_t \,|\, z_t) = z_t^{\rm t}\beta,$$

the classic sum of squares.

(b) Show that the ML estimator for β is the least sum of squares estimator, with a formula

$$\widehat\beta = \Sigma_n^{-1} n^{-1} \sum_{t=1}^n z_t x_t = \Big(n^{-1}\sum_{t=1}^n z_t z_t^{\rm t}\Big)^{-1} n^{-1}\sum_{t=1}^n z_t x_t,$$

assuming here that there is no linearity between the covariate vectors, so that Σ_n has full rank. Show then that the ML estimator for σ is $\widehat\sigma_k^2 = Q_{\min}/n = Q(\widehat\beta)/n$. Deduce from this that

$$\ell_{k,\max} = \max\{\ell_k(\beta,\sigma)\colon \text{all } \beta, \sigma\} = -n\log\widehat\sigma_k - \tfrac{1}{2}n - \tfrac{1}{2}n\log(2\pi).$$

(c) Deduce that for such a linear regression model, with k covariates on board, we have

$$\begin{aligned}\mathrm{aic}_k &= 2(k+1) + 2n\log\widehat\sigma_k + n + n\log(2\pi),\\ \mathrm{bic}_k &= (k+1)\log n + 2n\log\widehat\sigma_k + n + n\log(2\pi).\end{aligned}$$

By omitting factors not depending on the different models, show then, that doing well for AIC is the same as having a small $k + n\log\widehat\sigma_k$, or $2k + n\log\widehat\sigma_k^2$; and that doing well for BIC is the same as having a small $k\log n + 2n\log\widehat\sigma_k$, or $k\log n + n\log\widehat\sigma_k^2$.

(d) Above we've derived AIC and BIC formulae from their general definitions. Check that 'doing well with AIC' is equivalent to what we find by using the book's AIC formula, and the same with BIC, even though the book's AIC and BIC formulae are not fully identical to the aic_k and bic_k above. – The general AIC and BIC formulae, as laid out in this exercise, are part of the course's active curriculum, and can specifically be used when comparing different time series models for the same dataset.

11. Where are the snows of yesteryear?

Figure 0.2 is a dramatic one, for at least my segment of civilisation. It gives the number of skiing days at the location Bjørnholt in Nordmarka, a skiing hour away from tram stations Voksenkollen and Frognerseteren, with skiing day defined as there being at least 25 cm snow on the ground. The linear trend is the estimated regression line using what we call Model 2 below, drastically indicating that the climate has consequences also for the skiing days of the Oslo people. See Heger (2011) and Cunen, Hermansen, and Hjort (2019) for further discussion and details.

The time series goes from 1897 to 2015, but, crucially, there's a big hole in the series, with no data recorded from 1938 to 1954. This spells trouble for classes of traditional time series models, since there prefer data to be equidistanced. We may still model and analyse the data, using autocorrelation functions, etc., though.

(a) Let for convenience $z_t = \texttt{year} - 1896$, so that these start out like $1, 2, 3, \ldots$, and let x_t be the skiing days number for year t, if recorded. Fit first Model 0 and Model 1, using ordinary linear regression, ignoring time dependence. Model 0 takes $x_t = \beta_0 + \varepsilon_{0,t}$, with the $\varepsilon_{0,t}$ i.i.d. $\mathrm{N}(0, \sigma_0^2)$, i.e. assumes a constant stationary level. Model 1 takes $x_t = \beta_0 + \beta_1 z_t + \varepsilon_{1,t}$, with the $\varepsilon_{1,t}$ i.i.d. $\mathrm{N}(0, \sigma_1^2)$. Give a 95 percent confidence interval for β, and give an interpretation of this negative trend coefficient. Also carry out AIC analysis. You should find log-likelihood maxima $\ell_{0,\max} = -519.479$ and $\ell_{1,\max} = -512.167$.

(b) For Model 1, compute and inspect the estimated residuals, $r_t = \{x_t - \widehat m_1(t)\}/\widehat\sigma_2$, where $\widehat m_2(t)$ is the estimated trend under Model 1. Check in particular the acf, and comment.

(c) Then go to Model 2, which includes autocorrelation. We take

$$x_t = \beta_0 + \beta_1 z_t + \sigma\varepsilon_t \quad \text{for } t = 1, 2, 3, \ldots, \quad \text{with } \mathrm{corr}(\varepsilon_s, \varepsilon_t) = \rho^{|s-t|}.$$

So ρ is the correlation for skiing days numbers for consecutive years; ρ^2 for times two years apart, etc. We also take the ε_t to be jointly multinormal with mean zero and variance one. Show that this entails

$$x \sim \mathrm{N}_n(\xi, \sigma^2 A_\rho),$$

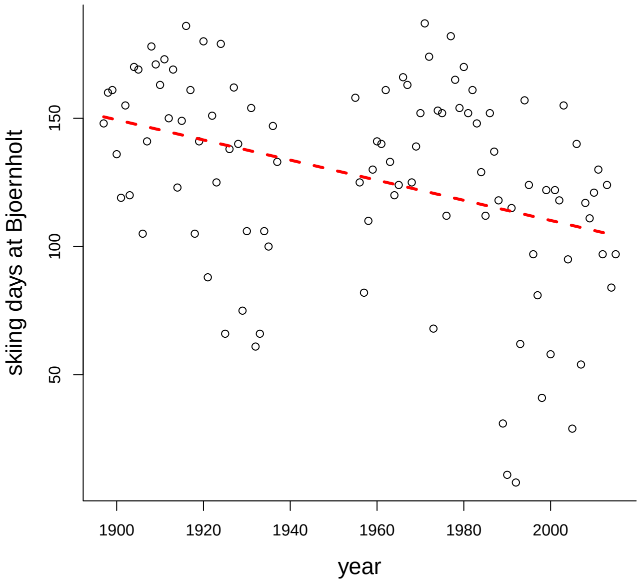

Figure 0.2: The number of skiing days per year, at the location Bjørnholt in Nordmarka, from 1897 to 2015, though with a gap in the series, with no records from 1938 to 1954. The red line is the estimated regression from the four-parameter Model 2.

where $\xi_t = \beta_0 + \beta_1 z_t$, and A_ρ is the $n \times n$ matrix with 1 on the diagonal and $\rho^{d_{i,j}}$ in position (i, j), with $d_{i,j}$ the time difference. Note that this A_ρ is well-defined in spite of the gap in the time series. We have $n = 102$, the number of observations.

(d) Show that the log-likelihood function can be written

$$\ell(\beta_0, \beta_1, \sigma, \rho) = -n \log \sigma - \tfrac{1}{2} \log(\det(A_\rho)) - \tfrac{1}{2}(x - m_t)^{\rm t} A_\rho^{-1}(x - m_t)/\sigma^2 - \tfrac{1}{2} n \log(2\pi),$$

where $m_t = \beta_0 + \beta_1 z_t$. It is numerically a bit troublesome to maximise this here (also since we cannot uitilise simplifying formula for the inverse and determinant of A_ρ, due to the gap in the data, which means data not being equidistant). It is practical to compute and display the log-likelihood profile function instead:

$$\ell_{\rm prof}(\rho) = \max\{\ell(\beta_0, \beta_1, \sigma, \rho)\colon \text{all } \beta_0, \beta_1, \sigma\} = \ell(\widehat\beta_0(\rho), \widehat\beta_1(\rho), \widehat\sigma(\rho), \rho).$$

Try to reproduce Figure 0.3.

(e) In particular, by carrying out these computations, involving maximising over parameters $(\beta_0, \beta_1, \sigma)$ for each ρ, you should find that the ML estimate for ρ is $\widehat\rho = 0.208$, and that $\ell_{2,\max} = -509.983$. Carry out AIC analysis for comparing Models 0, 1, 2.

(f) Given Model 2, predict the numnber of skiing days in 2013, given the data collected up to 2012, and give an approximate 90 percent confidence interval. Do this exercise also trusting Model 1; compare, and discuss.

(g) Try out one or two more models for these data.

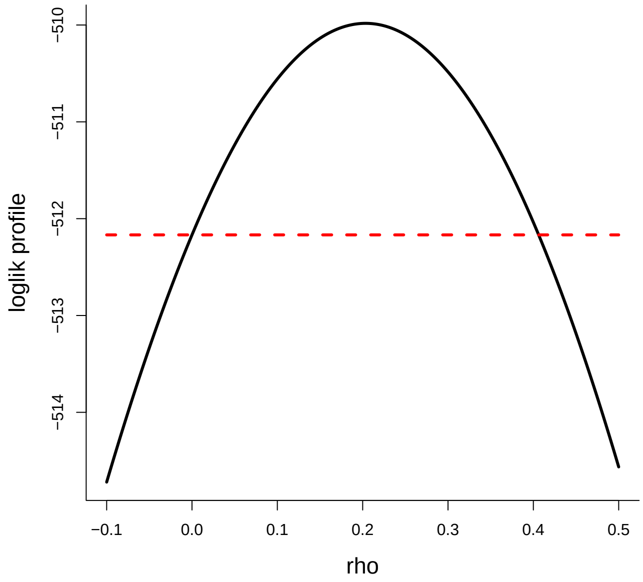

Figure 0.3: The log-likelihood profile funcion $\ell_{\rm prof}(\rho)$, for the Bjørnholt data, for the four-parameter model with linear trend, a constant σ, and correlation function $\rho^{|s-t|}$, for pairs of data with interdistance $|t-s|$. The horizontal dashed line indicates the level $\ell_{1,\max}$ obtained for the submodel of independence, where $\rho = 0$.

12. Estimating the three parameters in stationary AR(1)

Consider the stationary Gaussian AR(1) model, with

$$x_t = \mu + \sigma\varepsilon_t \quad \text{for } t = 1, \ldots, n,$$

where the ε_t are standard normal, but correlated with $\text{corr}(\varepsilon_s, \varepsilon_t) = \rho^{t-s|}$.

(a) Take $n = 100$, $\mu = 0$, $\sigma = 1$, $\rho = 0.555$ in your computer, and simulate a dataset from this model. Use results and insights from Exercise 2 to do this. – There are also other general simulation schemes, for simulating from a general multinormal distribution, which I will briefly come back to in my teaching. You may also use `library(MASS)` and then use `mvrnorm`.

(b) Then estimate (μ, σ, ρ) from the data you've created, using ML, maximum likelihood. You may do this via the log-likelihood profile function $\ell_{\rm prof}(\rho)$; see earlier R scripts from Nils of this type.

(c) Compare your $\widehat\rho_{\rm ml}$ with two other estimators, both of the form

$$\rho^* = \frac{1}{n}\sum_{t=2}^{n} \frac{x_{t-1} - \widehat\mu}{\widehat\sigma}\frac{x_t - \widehat\mu}{\widehat\sigma}.$$

Version (i) uses the simple classic estimates for (μ, σ), trusting independence; version (ii) uses the more elaborate $(\widehat\mu_{\rm ml}, \widehat\sigma_{\rm ml})$, from ML in the three-parameter model.

(d) Construct both an estimator and an (approximate) 90 percent confidence interval for the next point, i.e. x_{n+1}, based on having observed the first n datapoints.

(e) When your code works, for a single simulated dataset, to a loop on top, to simulate the full thing e.g. `sim = 1000` times, to learn how the estimators perform. Whare are the differences in performance, for the three estimators of ρ? Do your 90 percent confidence intervals manage to capture x_{n+1} anout 90 percent of the time?

13. Estimating cycle length

A model used a few places in the book for capturing cyclic behaviour is

$$x_t = a\cos(2\pi t/\omega + \phi) + \varepsilon_t \quad \text{for } t = 1, \ldots, n,$$

with different natural assumptions for the the ε_t. We will do fuller time series versions of this later, but on this occasion we make life simple by taking the ε_t i.i.d. $\mathrm{N}(0, \sigma^2)$. The model has three parameters for the mean, including the crucial cycle length parameter ω, and so far one for the variability.

(a) Simulate such a dataset, for say $n = 200$, and with values you choose yourself for $a_{\rm true}, \phi_{\rm true}, \sigma_{\rm true}$, and take $\omega_{\rm true} = 7$ (think about seven days a week). First take $\omega_{\rm true}$ to be known, and estimate the parameters a, ϕ. You may use the trick of Example 2.10 in the book, to convert the problem to linear regression in $\cos(2\pi t/\omega_{\rm true})$ and $\sin(2\pi t/\omega_{\rm true})$; or why not attack the problem directly, minimsing

$$Q_n(a, \phi) = \sum_{t=1}^{n} \{x_t - a\cos(2\pi t/\omega + \phi)\}^2$$

by throwing it to the clever `nlm` minimisation algorithm. Check that these two computational methods give the same answers.

(b) Then estimate also ω from your simulated dataset, using the profiled log-likelihood function

$$\ell_{\rm prof}(\omega) = \max\{\ell(a, \phi, \sigma, \rho) : \text{over all } a, \phi, \sigma\}.$$

You might find that the cycle length ω is rather sharply estimated, with good precision.

(c) How can you set approximate 90 percent confidence intervals for the parameters? Play with your code a bit, setting diffferent values for $(a, \phi, \omega, \sigma)$, and also n. Check how your estimates work.

(d) Try to extend your model and estimation schemes to the case where there also is an autocorrelation parameter.

References

Claeskens, G. and Hjort, N.L. (2008). *Model Selection and Model Averaging.* Cambridge University Press.

Cunen, C., Hermansen, G., and Hjort, N.L. (2019). Confidence distributions for change-points and regime shifts. *Journal of Statistical Planning and Inference* 195, 14–34.

Heger, A. (2011). Jeg og jordkloden. Dagsavisen.
`aheger.blogspot.com/2007/03/jeg-og-jordkloden.html`

Schweder, T.. and Hjort, N.L. (2008). *Confidence, Likelihood, Probability.* Cambridge University Press.

Shumway, R.H. and Stoffer, D.S. (2016). *Time Series Analysis and Its Applications* (4th ed.). Springer, Berlin.