STK 4060-9060 Spring 2022
Time Series: the Oblig

This is The Oblig, the mandatory assignment, for STK 4060-9060, Time Series, Spring 2022. It is made available at the course website Tuesday April 19, and the submission deadline is Tuesday May 2, 13:58, *via the Canvas system*. Reports may be written in nynorsk, bokmål, riksmål, English, or Latin, should preferably be text-processed (for instance with TeX or LaTeX), and must be submitted as a single pdf file. The submission must contain your name, the course, and assignment number.

The Oblig set contains three exercises and comprises five pages (in addition to the present introduction page, 'page 0', and the last page is a brief Appendix).

It is expected that you give a clear presentation with all necessary explanations, but write concisely (in der Beschränkung zeigt sich erst der Meister; brevity is the soul of wit; краткость – сестра таланта). Remember to include all relevant plots and figures. These should preferably be placed inside the text, close to the relevant subquestion.

For a few of the questions setting up an appropriate computer programme might be part of your solution. The code ought to be handed in along with the rest of the written assignment; you might place the code in an appendix.

All aids, including collaboration, are allowed, but the submission must be written by you and reflect your understanding of the subject. If we doubt that you have understood the content you have handed in, we may request that you give an oral account.

Application for postponed delivery: If you need to apply for a postponement of the submission deadline due to illness or other reasons, you have to contact the Student Administration at the Department of Mathematics (email: `studieinfo@math.uio.no`) well before the deadline.

The mandatory assignment in this course must be approved, in the same semester, before you are allowed to take the final examination.

Complete guidelines about delivery of mandatory assignments, along with a 'log on to Canvas', can be found here:

`www.uio.no/english/studies/admin/compulsory-activities/mn-math-mandatory.html`

Enjoy [imperative pluralis].

Nils Lid Hjort

1. The ACF

THE AUTOCORRELATION AND COVARIANCE FUNCTIONS are fundamental tools for stationary time series. This exercise points to a certain type of danger when the empirical ACF is applied to a series which is not stationary.

(a) Explain what is meant by a stationary time series.

(b) Let $x_1, \ldots, x_n$ be a stationary time series, with finite variance $\text{Var}\, x_t = \sigma^2$. Define the covariance function $\gamma(h)$ and autocorrelation function $\rho(h)$, for values $h = 0, 1, 2, \ldots$. For the case of independence, when the x_t are i.i.d., what are the values of $\gamma(h)$ and $\rho(h)$?

(c) Then define the empirical ACF, $\widehat{\rho}(h)$, and the empirical covariance function, $\widehat{\gamma}(h)$. Carry out a very simple illustration, where you simulate $x_1, \ldots, x_n$ i.i.d. $\text{N}(0, \sigma^2)$, with $\sigma = 0.77$, for $n = 100$. You do not need to show me a figure of this time series, but report values you find, for $\widehat{\gamma}(h)$ and for $\widehat{\rho}(h)$, for $h = 0, 1, 2, 3$. These can of course be found from scratch, so to to speak, but it's simpler to use versions of `acf(x)$acf[1:5]` and `acf(xx,"covariance",lag=10)$acf`.

(d) Then change the setup above slightly, to include a modest linear trend function; specifically, consider the model $x_t = \beta t + w_t$, for $t = 1, \ldots, n$, with the w_t being i.i.d. $\text{N}(0, \sigma^2)$ with $\sigma = 0.77$, and with $\beta = 0.05$. Here βt is seen as a deterministic trend, so the randomness lies with the w_t. Find the variance of x_t, and also the correlations $\rho(1), \rho(2), \rho(3)$ for observations at positions 1 and 2 and 3 apart. Then illustrate this with a simulation: generate $x_1, \ldots, x_n$, with $n = 100$, and include figures in your report, with the x_t series, and its ACF.

(e) The empirical ACF for your x_t with trend does not at all look like the real underlying $\rho(1), \rho(2), \ldots$. Explain what has happened.

(f) Then apply some mathematics, to explain this phenomenon in some detail. Suppose $x_t = m_t + w_t$, with m_t deterministic and the w_t being i.i.d. $\text{N}(0, \sigma^2)$, for some σ. Find expressions for the expected values of $\widehat{\gamma}(0), \widehat{\gamma}(1), \widehat{\gamma}(2)$. For the specific case of $x_t = \beta t + w_t$, as above, perhaps with a small β, and $n = 100$, see what these expectation formulae amount to, and comment on your findings.

(g) This little illustration demonstrates that matters may be incorrectly interpreted if one applies ACF and similar tools to series which are not stationary. A typical trick is to 'detrend first', and then work with the detrended series, say $x_t^* = x_t - \widehat{m}_t$. Do this with your simulated data from point (c): estimate the trend βt (assuming here, for simplicity, that you as the statistician in charge know that the trend is linear), and give $\widehat{\rho}^*(h)$ for $h = 0, 1, 2, 3$, for the detrended time series. Comment on what you find.

2. The MA(2) model

CONSIDER THE MOVING AVERAGE OF ORDER 2 model, the MA(2), for simplicity here taken to have mean zero, and then defined by

$$x_t = w_t + \theta_1 w_{t-1} + \theta_2 w_{t-2}, \qquad (*)$$

in terms of i.i.d. zero-mean variables w_t, with variance σ_w^2. Here θ_1, θ_2 are parameters of the model, and since it's useful for formulae to come, we also let $\theta_0 = 1$. To have a clear stationary structure, we take equation (*) to hold for all $t = 1, \ldots, n$, with x_1 and x_2 then involving not only w_1 but also appropriate w_0, w_{-1}.

(a) For the variance $\gamma(0)$ and covariances $\gamma(1), \gamma(2), \ldots$, show that

$$\gamma(0) = (\theta_0^2 + \theta_1^2 + \theta_2^2)\sigma_w^2, \quad \gamma(1) = (\theta_0\theta_1 + \theta_1\theta_2)\sigma_w^2, \quad \gamma(2) = \theta_0\theta_2\sigma_w^2,$$

with $\gamma(h) = 0$ for $h = 3, 4, \ldots$. Show also that the correlations become

$$\rho(1) = \frac{\theta_0\theta_1 + \theta_1\theta_2}{1 + \theta_1^2 + \theta_2^2}, \quad \rho(2) = \frac{\theta_0\theta_2}{1 + \theta_1^2 + \theta_2^2}.$$

(b) Simulate a reasonable number of (θ_1, θ_2), from any distribution around zero, and then show a figure of the resulting $(\rho(1), \rho(2))$. Comment on what you find.

(c) Simulate $x_1, \ldots, x_n$ from such an MA(2) process, with zero mean, taking $(\theta_1, \theta_2) = (0.66, 0.55)$, $\sigma_w = 0.88$, and $n = 250$. Compute $\widehat{\rho}(1), \widehat{\rho}(2)$, e.g. via the ACF. Then match these empirical values with the theoretical ones, and solve the equations, to provide estimates of (θ_1, θ_2). Also estimate the σ_w. – If this has gone well, your estimates should be reasonably close to the real values used in the simulation.

(d) Using the same simulated dataset as for (c), attempt to estimate the parameters using maximum log-likelihood.

(e) The general formula for the spectral density, in terms of the sequence of covariances $\gamma(h)$, has the form

$$f(\omega) = \sum_{h=-\infty}^{\infty} \gamma(h) \exp(-2\pi i h\omega).$$

Show that this can be expressed as

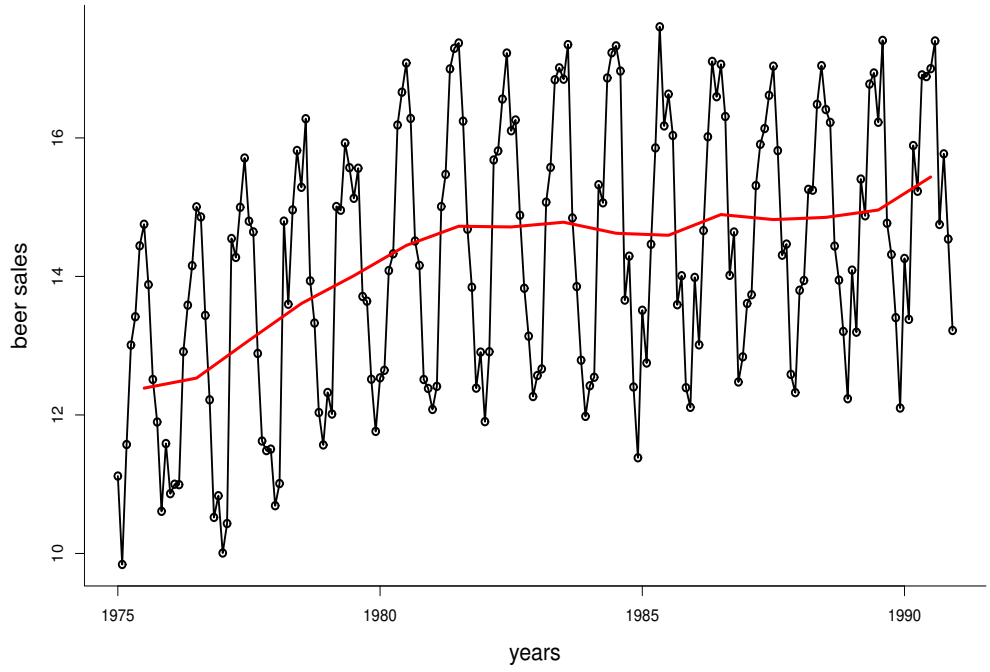
$$f(\omega) = \gamma(0) + 2\sum_{h=1}^{\infty} \gamma(h) \cos(2\pi h\omega).$$

Use this to work out a formula for the spectral density of an MA(2) process. Display this $f(\omega)$, on the $[-\frac{1}{2}, \frac{1}{2}]$ window, again for the case of $(\theta_1, \theta_2) = (0.66, 0.55)$ and $\sigma_2 = 0.88$.

(f) From the same simulated dataset of (c), attempt to estimate the spectral density non-parametrically. Use this to formulate ideas for have you can estimate the parameters of the model in a third way.

3. Beer sales in Sambandsstatene

SOME PEOPLE DRINK BEER. The datafile `beerdata`, available at the course website, has two columns, say `time` and `xx`, with the latter providing monthly beer sales, in millions of barrels, in Sambandsstatene, with time running from January 1975 to December 1990. Read the data into your computer (see the Appendix here, for a few R things), so that you can work with various aspects of the time series x_t, with $t = 1, 2, \ldots, n$, over $n = 16 \cdot 12 = 192$ months.



Monthly beer sales in Sambandsstatene, in millions of barrels, from January 1975 to December 1990; the red line in the middle represents yearly averages.

(a) Use the data to compute $y_1, \ldots, y_{16}$, the yearly averages, i.e. averaged over the year's twelve months, from 1975 to 1990. Construct a version of the figure here.

(b) We start out working with these averages, before we return to the full beer sales time series. Clearly these annual averages ave_i increase over time i. Fit two simple models, (i) with a linear trend, and (ii) with a quadratic trend. So far we do not worry about dependence between the averages. In detail, model (i) takes $\text{ave}_i = \beta_0 + \beta_1 i + \varepsilon_{1,i}$, with the $\varepsilon_{1,i}$ being i.i.d. $\text{N}(0, \sigma_1^2)$, and model (ii) uses $\text{ave}_i = \beta_0 + \beta_1 i + \beta_2 i^2 + \varepsilon_{2,i}$, with the $\varepsilon_{2,i}$ being i.i.d. $\text{N}(0, \sigma_2^2)$. These two models can be fitted from scratch, but since we so far ignore dependence, they are familiar linear regression models, and may be fitted using `lm`; see the brief Appendix. Argue that the quadratic trend model is best here.

(c) For the detrended series of averages, say $\text{ave}_i^* = \text{ave}_i - \widehat{m}_i$, with $\widehat{m}_i = \widehat{\beta}_0 + \widehat{\beta}_1 i + \widehat{\beta}_2 i^2$, give a plot of both ave_i^* and its ACF. Comment on what you find here.

(d) These findings and some diagnostic plots might indicate that a model with quadratic trend and AR(1) type residuals would do a good job. This model uses

$$\text{ave}_i = \beta_0 + \beta_1 i + \beta_2 i^2 + \sigma\varepsilon_i \quad \text{for } i = 1, \ldots, 16,$$

with the ε_i being standard normal, but with AR(1) dependence structure $\text{cov}(\varepsilon_i, \varepsilon_j) = \rho^{|j-i|}$. Try to fit this model, either by fitting the $\text{ave}_i^* = \text{ave}_i - \widehat{m}_i$ to a zero-mean AR(1), or by using a full log-likelihood. Comment on what you find.

(e) We now return to the full time series x_t, but keeping in mind what we've learned about the trend over years. Let $s_t = \text{time}_t - 1975$, where time_t is the first column in the `beerdata`; it starts at 0 (for Jan 1970) and ends at 15.917 (for Dec 1990). Argue why the model

$$x_t = \gamma_0 + \gamma_1 s_t + \gamma_2 s_t^2 + A\cos(2\pi t/12 + \phi) + w_t \quad \text{for } t = 1, \ldots, n$$

might be a good one, with an amplitude A and phase ϕ, and error terms w_t with zero mean. Show that this model can alternatively be represented as

$$x_t = \gamma_0 + \gamma_1 s_t + \gamma_2 s_t^2 + a\cos(2\pi t/12) + b\sin(2\pi t/12) + w_t \quad \text{for } t = 1, \ldots, n.$$

(f) Assume first that the w_t above are simple i.i.d. $\text{N}(0, \sigma_w^2)$. Then the model is a simple linear regression model. Fit the model, perhaps using `lm` in R. Plot the fitted expectation curve

$$\widehat{m}_t = \widehat{\gamma}_0 + \widehat{\gamma}_1 s_t + \widehat{\gamma}_2 s_t^2 + \widehat{a}\cos(2\pi t/12) + \widehat{b}\sin(2\pi t/12)$$

on top of the data, and comment on the fit.

(g) Using this model, predict the January and July beer sales, for the next year, i.e. 1991. Give also 90 percent prediction intervals.

(h) Do some diagnostic checks and plots for the detrended series $x_t^* = x_t - \widehat{m}_t$. Use this to suggest perhaps better time series models than the one of point (f). If you have time, use any models you invent to again predict beer sales for January and July 1991.

Appendix: a few R things

The `beerdata` can be read into your computer as follows:

```
beer <- matrix(scan("beerdata",skip=4),byrow=T,ncol=2)
time <- beer[ ,1]
xx <- beer[ ,2]
```

followed by `plot(time,xx,type="o")` etc.

To fit a linear model, of the type $x = \beta_0 + \beta_1 z_1 + \beta_2 z_2 + \varepsilon$, one may use `lm(x ~ z1 + z2)`. You may also use `summary(lm(x ~ z1 + z2))`, etc.

To find the annual averages, for the beer sales data, there are several tricks, of course. I used something like `ind3 <- 3 + 12*(0:15)` followed by `xx[ind3]`; this gives me the beer sales figures for the month March. Similar tricks give me other subsets of the `xx` series, from which I then can compute averages and other quantities.

Solving three equations with three unknowns (or seven equations with seven unknowns, for that matter): Suppose g_1, g_2, g_3 are functions of (u_1, u_2, u_3), and that we need to solve

$$g_1(u_1, u_2, u_3) = a_1, \quad g_2(u_1, u_2, u_3) = a_2, \quad g_3(u_1, u_2, u_3) = a_3.$$

My default way of doing this in R, if the equations are hard to solve with pen on paper, is to define the `g1`, `g2`, `g3` functions, and then programme the function

$$Q(u) = (g_1(u) - a_1)^2 + (g_2(u) - a_2) + (g_3(u) - a_3)^2.$$

I then use `mlm(Q,starthere)` to minimise Q. If the minimum is zero, I've solved my three equations.