Homework 4
UMN STAT 5511
Charles R. Doss
(Fall 2024)
Assigned: Mon, October 21
Due: Mon, November 4

The usual formatting rules:

- Your homework (HW) should be formatted to be easily readable by the grader.
- You may use knitr or Sweave in general to produce the code portions of the HW. However, the output from knitr/Sweave that you include should *be only what is necessary to answer the question*, rather than just any automatic output that R produces. (You may thus need to avoid using default R functions if they output too much unnecessary material, and/or should make use of `invisible()` or `capture.output()`.)
  - For example: for output from regression, the main things we would want to see are the estimates for each coefficient (with appropriate labels of course) together with the computed OLS/linear regression standard errors and p-values. If other output is not needed to answer the question, it should be suppressed!
- Code snippets that directly answer the questions can be included in your main homework document; ideally these should be preceded by comments or text at least explaining what question they are answering. Extra code can be placed in an appendix.
- All plots produced in R should have appropriate labels on the axes as well as titles. Any plot should have explanation of what is being plotted given clearly in the accompanying text.
- Plots and figures should be *appropriately sized,* meaning they should not be too large, so that the page length is not too long. (The arguments fig.height and fig.width to knitr chunks can achieve this.)
- **Directions for "by-hand" problems:** In general, credit is given for (correct) shown work, not for final answers; so show **all** work for each problem and explain your answer fully.

For some questions on this homework you will need to do parameter estimation from a series, which we have not yet discussed at the time of the assignment of this homework (but we will discuss this shortly). You may use the arima() function for estimation. You do not need to worry about the 'method' argument to arima() (any choice will be OK).

1. Find (on Canvas) the file "hw4dat.rsav" (which can be loaded into R using load("hw4dat.rsav")). It contains a time series ("xx"). The series is a "demeaned" monthly revenue stream (in millions of dollars) for a company. There are $n = 96$ observations.

   The series has been "demeaned"; usually that would mean we subtract off $\bar{X}$ from every data point, but pretend for now we know the mean μ exactly so we have subtracted off μ from every data point, so the new series is exactly (theoretically) mean 0. (But thus its sample mean is not precisely 0.)

   We will consider possible ARMA models for the series X_t. We assume that the corresponding white noise is Gaussian (so X_t is Gaussian).

   We will consider first an AR(2) model. We assume we know the true model exactly: it is

   $$\text{Model 1:} \qquad X_t = 1.34X_{t-1} - .48X_{t-2} + W_t, \quad W_t \overset{\text{iid}}{\sim} N(0,1).$$

   (a) Compute forecasts and backcasts using Model 1, up to 25 time steps in the future and into the past. Write code to do the prediction by hand (i.e., I just/only mean that you should not use the predict() function). Plot the data, forecast, backcast, and 95% prediction intervals based on assuming gaussianity (all on one plot). (Note: you do not need to do a multiplicity correction for the prediction intervals.)

   (b) Give a constant (nonrandom) number that the 100-step-ahead forecast, X_{196}^{100}, will be approximately equal to.

   (c) If you were to do the one-step-ahead prediction but based on no data, what would your prediction be? (Based on no data, it is the same for predicting at time 97 or at any other time.) What would the mean-squared prediction error (call it E) be? Compare P_{97}^{96} to E.

   (d) Now say that we know the true mean of the company's revenue series is .3 (million dollars). Provide

i. a plot of the company's (not-demeaned) revenue series (let's call it Y_t),

ii. and the prediction equation for Y^n_{n+1} (I do not mean for you to write anything having to do with setting any expectations to 0 here; I simply mean for you to write the formula for predicting Y at time $n+1$ based on the model (including the mean), and past data).

(e) The series Y_t is a monthly revenue stream for a company. The company needs to decide, before the current month is up (i.e., before seeing the one-month-ahead revenue, Y_{97}), whether to make an important investment in equipment which will cost 1.1 million dollars. If their revenue next month cannot cover the cost (is less than the cost of the investment) they will go bankrupt (they have exactly 0 cash on hand and cannot take out loans). Explain why they should or should not make the investment.

(f) The second model we will consider (Model 2) is an MA(2) model. Estimate an MA(2) model on the X_t (demeaned) series using the arima() function (there is an include.mean variable; set it to false). Then plot (on one plot): the data, forecasts up to 25 months ahead, and 95% prediction intervals [assuming gaussianity]. You may use the predict() function.

(g) Compare the forecasts for Model 1 and Model 2: specifically, discuss how quickly the two forecasts revert to the long run average of the series. Provide an explanation of this. [Note: I am not intending for you to discuss the fact that in one case the coefficients were estimated and in the other they were not estimated.]

(h) Now we consider changing the frequency of observation. Imagine someone outside the company observes the company revenue but only on a quarterly basis, by which I mean every 3 months (thus they observe March's revenue, then June's revenue, ...). Let this series of de-meaned observations be Z_t. If the true model for X_t is AR(1), $X_t = \phi_1 X_{t-1} + W_t$ where $W_t \overset{\text{iid}}{\sim} N(0, \sigma^2)$, then what is the (true) model for Z_t (including the distribution of the white noise series)?

2. Shumway and Stoffer (4th ed.), question 3.9

3. Shumway and Stoffer (4th ed.), question 3.15

4. Shumway and Stoffer (4th ed.), question 3.20

5. A data analyst is analyzing a time series with 1000 observations. She fits an ARMA(2,1) model (mean 0 i.e., with no intercept), yielding the following estimate output:

```
Coefficients:
         ar1      ar2     ma1
      -0.062    0.817   0.971
```

To check the robustness of the fit, the analyst removes the last 100 observations (leaving 900) and fits the ARMA(2,1) model again. The analyst is surprised to see the following very different estimates output:

```
Coefficients:
         ar1      ar2     ma1
       0.752    0.112   0.100
```

(a) Provide an explanation for why these different results were output.

(b) Provide a diagnostic tool or mechanism or method to assess the answer you gave in the previous part, and explain how you would use it or what you would look for.