Homework 7

Due: Friday 4/3/20 by 5:00pm

This week, we're going to wrap things up with ARMA models and the broccoli data. We'll try modeling dependence over time using an autoregressive moving average model of orders p and q.

Again, we're going to keep working with the `broc` data for a little while longer. Again, it is posted on the course website, which contains the average price of one pound of broccoli in urban areas each month, from July 1995 through December 2019. In this problem, we're going to consider two different ways of modeling the `broc` data. Letting z_t refer to the observed broccoli prices, the two approaches we will consider are:

i. First regressing out a linear time trend and month effects obtained from fitting a linear regression model to all but the last 12 months of data. Letting y_t refer to the corresponding residuals, we will model the residuals as an ARMA(p, q) process:

$$y_t = \mu + \sum_{i=1}^{p} \phi_i \left(y_{t-i} - \mu\right) + \sum_{j=1}^{q} \theta_j w_{t-j} + w_t, \quad w_t \overset{i.i.d.}{\sim} \text{normal}\left(0, \sigma^2_w\right), \tag{1}$$

where $y_t - \mu$ is a stationary process. We will treat the estimated regression coefficients from the initial linear regression model as fixed and known throughout.

ii. Directly incorporating the linear time trend and month effects:

$$z_t = \boldsymbol{x}'_t\boldsymbol{\beta} + \sum_{i=1}^{p} \phi_i \left(z_{t-i} - \boldsymbol{x}'_{t-i}\boldsymbol{\beta}\right) + \sum_{j=1}^{q} \theta_j w_{t-j} + w_t, \quad w_t \overset{i.i.d.}{\sim} \text{normal}\left(0, \sigma^2_w\right), \tag{2}$$

where $z_t - \boldsymbol{x}'_t\boldsymbol{\beta}$ is a stationary process.

```r
load("~/Dropbox/Teaching/TimeSeries2020/stat697/content/data/broc.RData")
set.seed(1)
broc$fdate <- as.Date(broc$date, "%Y-%m-%d")
broc$month <- format(broc$fdate, "%m")
broc$dayssincestart <- as.numeric(broc$fdate) - min(as.numeric(broc$fdate))
n <- nrow(broc)
m <- nrow(broc) - 12
linmod <- lm(price~dayssincestart+factor(month), data = broc,
             subset = 1:m)
pred <- predict(linmod, broc)
y <- broc$price - pred
z <- broc$price
```

Order Selection

In this part, our goal is to choose the order of the ARMA(p, q) model for approaches (i) and (ii). Consider $p = 0, 1, 2, 3$ and $q = 0, 1, 2$, and the following measures of model performance:

- Leave-one-out moving window style cross-validation, using the mean squared error of the forecasts
- Leave-twelve-out moving window style cross-validation, using the mean squared error of the forecasts
- AIC

- BIC
- Box-Pierce Test Result
  - This is the test you implemented on Homework 3, see parts (e) and (f) of the second problem

(a) Describe any choices you need to make when computing these measures of model performance, e.g. number of moving windows, and justify your choices.

(b) Summarize your results in a table which shows the best choice of p and q for each measure of of model performance.

(c) For this problem, does the best choice of p and q depend much on the approach used?

(d) Choose one measure of model performance to base your choice of p and q on, and justify your choice.

2. Repeat 1, but directly incorporating the regression function into the model.

Forecasting

Using the best ARMA(p, q) model based on your choice in part (d) of the previous problem, compute predicted values of the rest of the time series using the observed data using both of the two approaches. You can use the `predict` function to do get these predictions, with `n.ahead` set to the number of remaining observations.

(b) Make a table that shows estimates of the coefficient of the linear trend term, the autoregressive and/or moving average parameters, and noise variance fit using both approaches. Does the approach used affect the estimates much?

(c) Compare the predicted values obtained using each approach by plotting them on the same plot as the last 24 observations. Does the approach used affect the predicted values much?

(d) Use `predict` to obtain the variances of the predictions, and add 95% prediction intervals to your plot for each method. Does the approach used affect variances of the predictions much?

(e) Using the parametric bootstrap, obtain approximate standard errors of the predicted values. Remember - in approach i. we are treating the fitted values from the regression as fixed, so just simulate new residuals and continue to use the same fitted values. Add 95% confidence intervals to your plot for each method. Does the approach used affect variances of the confidence intervals much? How do the confidence intervals compare to the prediction intervals?

(f) The prediction intervals obtained in (c) account for variability due to uncertainty regarding future values of the time series but do not account for uncertainty regarding the values of the parameters of the model. We can obtain a variance estimate that accounts for both sources of uncertainty by taking the sum of the variances obtained in (c) and (d). Prediction intervals that account for both sources of uncertainty can be obtained using this variance new estimate. Add these prediction intervals to your plot for each approach. Does the approach used affect variances of the confidence intervals much? Which source of variability appears to dominate?

Final Project

If youb have not yet submitted a proposed data set for the final project, please do so as part of this assignment. At minimum, provide a link to the data set, and plot of the time series you would like to analyze, and the number of observations. You may work with your classmates to find a dataset, and you multiple students can use the same data set for their final project.