

Homework 8

Due: Wednesday 4/15/20 by 5:00pm

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

We're going to keep working with the `broc` data for a little while longer. It is boring to keep doing so, but it's nice to have a constant baseline as we learn. 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'll continue 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 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 $\text{ARMA}(p, q)$ process:

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

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 month effects:

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

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

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

Differencing

First, let's implement the augmented Dickey-Fuller and Phillips-Perron tests to perform a level $\alpha = 0.05$ test the null hypothesis that the undifferenced data is non-stationary. Implement each test with and without a linear time trend using the `ndiffs` function for from the `forecast` package for R. Make sure that you specify the type of test correctly.

- (a) Make a 2×2 table that provides the results of each test, with and without a linear time trend applied to residuals y_t .
- (b) Based on (a), is it necessary to difference the residuals y_t if a linear time trend is not included?
- (c) Based on (a), is it necessary to difference the residuals y_t if a linear time trend is included?
- (d) Make a 2×2 table that provides the results of each test, with and without a linear time trend applied to observed data z_t .
- (e) Does it make sense to apply either test directly to the observed time series data, z_t , given that we believe month effects may be present? Why or why not?
- (f) Based on (a) and your answer in (d), is it necessary to difference the observed data z_t if a linear time trend is not included?
- (g) Based on (a) and your answer in (d), is it necessary to difference the observed data z_t if a linear time trend is included?

Order Selection for Differenced Data

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

- One-step-ahead moving window style cross-validation, using the mean squared error of the forecasts
- Twelve-steps-ahead 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. If there are any choices you need to make when computing these measures of model performance, e.g. number of moving windows, make the same choices that you made on HW 7.
- (b) Summarize your results in a table which shows the best choice of p and q for each measure 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.

Forecasting

Using the best $\text{ARIMA}(p, 1, q)$ model based on your choice in part (d) of the previous problem, compute predicted values of the rest of the time series (just the last 12 observations) 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.

- (a) Make a table that shows estimates of the month effects, the autoregressive and/or moving average parameters, and noise variance fit using both approaches. Does the approach used affect the estimates much?
- (b) Compare the predicted values for the last 12 observations 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?

- (c) Use `predict` to obtain the variances of the predictions of the last 12 observations, and add 95% prediction intervals to your plot for each method. Does the approach used affect variances of the predictions much?
- (d) Using the parametric bootstrap, obtain approximate standard errors of the predicted values of the last 12 observations. 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?
- (e) The prediction intervals for the last 12 observations 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 for the last 12 observations to your plot for each approach. Does the approach used affect variances of the confidence intervals much? Which source of variability appears to dominate?

Putting It All Together

In this problem, let's just consider the approach (ii). Recall the following measures of model performance:

- One-step-ahead moving window style cross-validation, using the mean squared error of the forecasts
 - Twelve-steps-ahead moving window style cross-validation, using the mean squared error of the forecasts
 - AIC
 - BIC
 - Box-Pierce Test Result
- (a) Which measure(s) of model performance would be appropriate to use if you wanted to compare the performance of undifferenced and differenced $\text{ARIMA}(p, 0, q)$ and $\text{ARIMA}(p, 1, q)$ models?
 - (b) Pick one measure from the ones you identified in (a). You will use this measure to compare the performance of best $\text{ARIMA}(p, 1, q)$ model identified in this homework with the best $\text{ARIMA}(p, 0, q)$ identified in Homework 7. Justify your choice.
 - (c) Provide the values of the measure you chose in (b) for the best $\text{ARIMA}(p, 1, q)$ model identified in this homework and the best $\text{ARIMA}(p, 0, q)$ identified in Homework 7.
 - (d) Based on (c), indicate whether or not you think it is best to difference the data in this particular application.