Homework 3

Due: Tuesday 2/11/20 by 10:00am

Both problems will require that you continue to work with `broc` data 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.

Fitting Versus Forecasting

1. Consider the model $\texttt{price}_i = \mu + \sum_{j=1}^{d} \phi_j z_{ji} + \sum_{k=2}^{12} \alpha_k (\texttt{month}_i = k) + \epsilon_i$, $\epsilon_i \overset{i.i.d.}{\sim}$ normal$(0, \sigma^2)$, where $\boldsymbol{z}_j$ corresponds to the orthogonal polynomials of degree j, respectively, over the set of points given by `days since start`. You can use the `poly` function to construct the orthogonal polynomials.

(a) Using leave-one-out cross validation on the all but the last 12 months of data. Plot the average MSE of of the predicted training data as a function of d. Which value of d produces the prediction error on the training data?

(b) Perform one-step-ahead cross validation on the all but the last 12 months of data, using time series of length 100 for each training subset. Plot the average MSE of the one-step-ahead forecast as a measure of model performance as a function of d. Which value of d minimizes one-step-ahead forecast error?

(c) Plot AIC, AICc, and BIC/SIC for the model fit to the training data as a function of d. Pick a criterion (AIC, AICc, or BIC), and state which value of d you would choose based on it. In at most one sentence, justify your choice of criterion with reference to the data.

Autocorrelation

For this problem, we will examine the residuals r_i from fitting the model to all but the last 12 months of the data: $\texttt{price}_i = \mu + \beta_1 \texttt{days since start}_i + \sum_{k=2}^{12} \alpha_k (\texttt{month}_i = k) + \epsilon_i$, $\epsilon_i \overset{i.i.d.}{\sim}$ normal$(0, \sigma^2)$.

(a) Plot the autocorrelation function of the residuals for lags $0, 1, \dots, 50$.

(b) Using the parametric bootstrap, simulate $10,000$ bootstrap samples of the data from the model, using the least squares estimates of $\boldsymbol{\beta}$ and σ^2, according to the following two procedures:

- Procedure 1: Simulate bootstrap samples of the data $\boldsymbol{y}^{(k)}$ according to the model;
- Procedure 2: Simulate bootstrap samples of the residuals $\boldsymbol{\epsilon}^{(k)} \sim$ normal $\left(\boldsymbol{0}, \sigma^2 \boldsymbol{I}\right)$, using the estimate of σ^2 from the least squares regression fit. For each simulated dataset, compute the autocorrelation function for lags $0, 1, \dots, 50$. Save each autocorrelation function value. Add two 95% intervals for each autocorrelation function value to your plot - for each autocorrelation value you will have one 95% interval for Procedure 1, and one 95% interval for Procedure 2.

(c) In at most one sentence, what feature(s) of the residuals does Procedure 2 account for, whereas Procedure 1 does not, and does this appear to matter for this data?

(d) Based on the Figure you made in (b), is there evidence for residual correlation across time in the broccoli data after subtracting off a linear time trend? Answer in at most one sentence.

(e) Recall the approximate distribution of each sample autocorrelation value $\hat{\rho}_y(h)$ as $n \to \infty$, for fixed h, for a Gaussian white noise process $\boldsymbol{y}$. If we assume that $\hat{\rho}_y(h)$ and $\hat{\rho}_y(l)$ are independent if $h \neq l$, what is the approximate distribution of $n \sum_{l=1}^{h} \hat{\rho}_y(h)^2$ for a Gaussian white noise process $\boldsymbol{y}$ as $n \to \infty$, for fixed h?

(f) Using your results from (e), test the null hypothesis that the first $h = 50$ autocorrelations sum to exactly zero at level $\alpha = 0.05$. Give the value of the test statistic, the corresponding quantiles of the test statistic under the null.

(g) Using the two parametric bootstrap procedures described in (b), test the null hypothesis that the first $h = 50$ autocorrelations sum to exactly zero at level $\alpha = 0.05$. Give the value of the test statistic, the corresponding quantiles of the test statistic under the null for each procedure.

(h) Based on (f) and (g), does your answer to (d) change? Answer in at most one sentence.