Homework 4

Due: Thursday 2/20/20 by 10:00am

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. Throughout this problem, we're just going to work with the residuals from fitting a linear model with a linear time trend and month effects to all but the last 12 months of data. We'll call them `y`, because we'll be thinking of them as our observed time series.

```
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$fdate))
n.sub <- nrow(broc) - 12
linmod <- lm(price~dayssincestart+factor(month), data = broc,
             subset = 1:(n.sub))
pred <- predict(linmod, broc)
y <- broc$price - pred
```

In last week's homework, we saw evidence of substantial dependence across time in the residuals `y`. Now we're going to try modeling that dependence over time, using an autoregressive model of order p:

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

Note: for this assignment, only assume what is written above! Do *not* assume that y_t is a stationary process.

(a) Write down what you can of the likelihood for $y_1, \ldots, y_{n-12}$ Indicate whether or not there are any values y_t which you cannot write down the likelihood for and/or need to condition on.

(b) The likelihood in (a) looks like a linear regression model. Clearly indicate how you would construct the response vector and matrix of covariates for a fixed value of p.

(c) For $p = 1, 2, 4, 8, 16, 32$, compute estimates of μ, $\phi_1, \ldots, \phi_p$ using `lm` or any other approach to computing regression coefficients and residual standard errors for a linear regression model. Make a plot with 6 panels. Using one panel for each value of p, plot the last 24 observations, the fitted values from corresponding fitted model, and the approximate 95% confidence intervals obtained by treating this as a standard regression problem.

(d) Recall that AIC, AICc, and BIC/SIC are only appropriate when all of the models being compared were fit to the same data, i.e. the likelihoods all corresponded to the likelihood of the same set of data points. Are AIC, AICc, or BIC/SIC appropriate for comparing these models fit in (c) with different values of p? If yes, pick a criterion (AIC, AICc, or BIC/SIC), justify your choice, and indicate which model you would choose.

(e) For $p = 1, 2, 4, 8, 16, 32$, compute estimates of μ, $\phi_1, \ldots, \phi_p$ using only $y_{33}, \ldots, y_{282}$ as values of the response. Make a plot with 6 panels. Using one panel for each value of p, plot the last 24 observations, the fitted values from corresponding fitted model, and the approximate 95% confidence intervals obtained by treating this as a standard regression problem.

(f) Are AIC, AICc, or BIC/SIC appropriate for comparing these models fit in (e) with different values of p? If yes, pick a criterion (AIC, AICc, or BIC/SIC), justify your choice, and indicate which model you would choose.

(g) For each value of $p = 1, 2, 4, 8, 16, 32$, simulate 100 synthetic time series $\boldsymbol{y}^{(1)}, \ldots, \boldsymbol{y}^{(100)}$ according to the model given by (1), with $\hat{\mu}$, $\hat{\phi}_i$, and $\hat{\sigma}_w^2$ given by the linear model fits from (e).Condition on or hold constant whatever you need to condition on. Make a plot with six panels, one for each value of p. In each panel, plot the average lag-$h = 0, \ldots, 24$ autocorrelations across all simulations. Comment on how the autocorrelations change with p.

(h) For $p = 1, 2, 4, 8, 16, 32$, perform leave-one-out cross-validation. Leaving out observations $y_{33}, \ldots, y_{282}$ from one at a time and using as much data as possible as the response for each value of p. Record the squared error loss on the left out observations and plot the average it as a function of p. Indicate which model performs best according to this metric. Note any complications you encounter. Hint: does leaving out a single value y_t only affect the response?

(i) For $p = 1, 2, 4, 8, 16, 32$, perform leave-one-out cross-validation. Leaving out observations y_{184}, y_{282} from the response one at a time, fit the model to the previous 150 values of the time series. Record the squared error loss on the left out observation and plot the average it as a function of p. Indicate which model performs best according to this metric.

(j) Make a plot with two panels. In the first panel, plot all of the values of $\boldsymbol{y}$ along wtih the fitted values and approximate 95% confidence intervals using the model identified in (h). In the second panel, plot all of the values of $\boldsymbol{y}$ along wtih the fitted values and approximate 95% confidence intervals using the model identified in (i). Comment on how the fits from the two models compare.