Homework 2

Due: Tuesday 2/4/20 by 10:00am

Continued Regression and `R` Review

1. This problem will require that you continue to work with `broc` data 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. We will consider the following three regression models:
   1. $\texttt{price}_i = \mu + \beta_1 \texttt{days since start}_i + \sum_{k=2}^{12} \alpha_k (\texttt{month}_i = k) + \epsilon_i$, $\epsilon_i \overset{i.i.d.}{\sim} \text{normal}(0, \sigma^2)$
   2. $\texttt{price}_i = \mu + \beta_1 \texttt{days since start}_i + \beta_2 \texttt{days since start}_i^2 + \sum_{k=2}^{12} \alpha_k (\texttt{month}_i = k) + \epsilon_i$, $\epsilon_i \overset{i.i.d.}{\sim}$ $\text{normal}(0, \sigma^2)$
   3. $\texttt{price}_i = \mu + \phi_1 \texttt{z}_{1i} + \phi_2 \texttt{z}_{2i}^2 + \sum_{k=2}^{12} \alpha_k (\texttt{month}_i = k) + \epsilon_i$, $\epsilon_i \overset{i.i.d.}{\sim} \text{normal}(0, \sigma^2)$, where $\boldsymbol{z}_1$ and $\boldsymbol{z}_2$ correspond to the orthogonal polynomials of degree 1 and 2, respectively, over the set of points given by `days since start`. You can use the `poly` function to construct the orthogonal polynomials.

   (a) Using the parametric bootstrap, choose between Model 1 and Model 2. Justify your choice in at most one sentence, and provide any relevant numerical evidence.

   (b) In at most one sentance, explain whether or not it is appropriate to choose between Models 1 and 3 using a t/z-test, an F-test, or parametric bootstrap versions of either.

   (c) Make one plot with three panels. In each panel, plot the prices in date order, from first to last. In the first panel, add parametric bootstrap fitted values from Model 1 along with 95% parametric bootstrap confidence intervals for each fitted value. In the first panel, add parametric bootstrap fitted values from Model 2 along with 95% parametric bootstrap confidence intervals for each fitted value. In the third and last panel, add parametric bootstrap fitted values from Model 3 along with 95% parametric bootstrap confidence intervals for each fitted value.

   (d) For Models 1, 2 and 3, compute predictions and standard errors for $\mathbb{E}[y_{295}]$.

   (e) For Models 1, 2, and 3, compute a parametric bootstrap mean and standard error for $\mathbb{E}[y_{295}]$. In at most one sentence, justify your choice of the number of bootstrap samples. Compare to the results from (d), and comment on the differences in at most one sentence.

   (f) Choose between Models 1, 2, and 3 using leave-one-out cross validation, using out-of-sample prediction errors to assess model fit. Justify your choice in at most one sentence.

   (g) Choose between Models 1, 2, and 3 using leave-five-out cross validation, using out-of-sample prediction errors to assess model fit. Justify your choice in at most one sentence. In at most one additional sentence, describe any practical challenges you encounter.

   (h) Choose between Models 1, 2, and 3 using leave-ten-out cross validation, using out-of-sample prediction errors to assess model fit. Justify your choice in at most one sentence. In at most one additional sentence, describe any challenges you encounter.

   (i) In at most one sentence, explain which procedure you would use to choose a model from those described in (f), (g), and (h) and justify your choice.