Homework 3

Due Thursday 2/24

Solutions for applied solutions should include R code and output – try to use `R markdown`. Homework must be submitted on Blackboard as one file.

Jiawei Liu, Chenyang Lu, Dexiu Ma and Yongxiao Ma will lead HW discussion.

1. (50 points) *Comparing Linear Regression, Regularization and Dimension Reduction for Log COVID-19 Case Rate Models*

   You will continue modeling COVID-19 case rates in the U.S. as a function of characteristics of the geographic area. This homework uses a new version of the HW2 data with a reduced number of variables. The data for this HW is *covid_data_pdb_v2.csv* and is posted on Blackboard under HW3.

   County-level COVID-19 case counts (as of 1/25/22) are obtained from `https://usafacts.org/visualizations/coronavirus-covid-19-spread-map/`. The target (a.k.a. dependent, output) variable is the log covid case rate for the county

   $$\log\left[100 * (covid\_count/Tot\_Population\_ACS\_14\_18)\right]$$

   where the denominator is the total population of the county estimated from the 5-year (2014-2018) American Community Survey from the U.S. Census Bureau. Geographic and demographic characteristics of each county are obtained from the Census Bureau's 2020 Planning Database (PDB).

   For this question use standardized versions of all of the $pct\_*ACS\_14\_18$ variables as predictors. These $pct\_*ACS\_14\_18$ variables are the percentage of the county that can be categorized into each of the particular geographic and demographic variables from the American Community Survey. We are interested in models that can help us sort through the 70 or so predictors.

   (a) Fit the following models*

   - least squares linear regression,
   - least squares linear regression with backward or forward selection (specify your selected procedure and selection criteria),
   - ridge regression,
   - lasso regression,
   - principal components regression and
   - partial least squares regression,

and obtain an estimate of test error assuming squared error loss [using any of the CV or bootstrap algorithms for estimating test error]. For each method with tuning parameters – e.g. λ in ridge regression, t in lasso, M in PCR – select the appropriate tuning parameter value based on the lowest estimated test error; include and discuss the plot of your test error curve for each.

(b) Discuss how the test error estimates compare across the six methods. Why do you think the methods with the lowest test error has the lowest test error? Why do you think the model with the largest test error has the largest test error?

(c) Compare the coefficients and predictors selected into the model for the different models (as in Table 3.3 of EoSL). Do they all suggest the inclusion of the same variables? Why or why not? Include and discuss the coefficient path plot for ridge regression and lasso.

(d) Which method would you recommend for this prediction task and why?

2. (50 points) *Comparing Linear Regression, Polynomial Regression and Splines for Log COVID-19 Case Rate vs. Household Technology*

For this problem use *pct_Pop_NoCompDevic_ACS*_14_18 as the only predictor. You are to investigate if there is evidence of a nonlinear relationship between log of COVID-19 case rate and percent of the population without a computing device of any kind.

(a) Fit the following models*

- least squares linear regression,
- polynomial regression
- piecewise polynomial regression,
- regression spline,
- natural cubic spline and
- smoothing spline,

and obtain an estimate of test error using squared error loss. For each method with tuning parameters – d in polynomial regression, d in piecewise polynomial regression, M for regression splines, λ in smoothing splines – select the appropriate tuning parameter value based on the lowest estimated test error; include and discuss the plot or table of your test error for varying values of the tuning parameters. For models that require knots, explain your choice of number and location of knots.

(b) Create a scatterplot of log COVID-19 case rate versus percent without computing devices. Add the fitted line/curve for each of the six models on the same graph. Comment on the similarities and differences in the fitted functions.

(c) Discuss how the test error estimates compare across the models. Why do you think the model with the lowest test error has the lowest test error? Why do you think the model with the largest test error has the largest test error?

(d) For polynomial regression, piecewise polynomial regression, regression spline, and natural spline, state and explain exactly what each row in the `lm()` coefficient summary output is. If you used functions such as `poly()` with `lm()`, the row names are automatically created in R and I want to make sure you know exactly what transforms of *pct_No_Health_Ins_ACS*_14_18 are being used.

(e) Which model would you recommend and why?

* You may use any pre-packaged functions in R to fit these models. That is, you do NOT need to program anything from scratch. You may, however, want to use your own CV or bootstrap functions that you hopefully created in HW2. If you use a pre-packaged function to do cross-validation or bootstrap, make sure to you know exactly what it is doing!

3. (0 points) *Group Request for Special Topic Lecture*

If you have a group preference for the group special topic lecture, please submit the list of students that you would like in your group in the "comments" section when you submit your HW on Blackboard. Only one member of each group needs to submit the request. Groups will be two to three students. If you submit a group of two, it is possible you will be randomly assigned another group member depending on how the numbers work out. If you do not have a group preference and would like to be randomly assigned, then you do not need to submit anything.