Linear Regression
HW 5
Due 10/5 at 11:59pm

Directions: Submit a .pdf file containing your responses. The .pdf can be converted from a Latex file, pictures of your handwritten solutions, word files, markdown files, etc. If there are coding problems, upload a separate notebook for Python code.

Written Questions

1. A student says "If extremely influential outlying cases are detected in a data set, simply delete all those cases from the data set." Would you agree? If not, what would you do?
2. Express the OLS solution for $\hat{\beta}$ in terms of the singular value decomposition of the design matrix $X = UDV^T$. In the case of extreme multicollinearity, the singular values of the design are very close to zero. Explain how this creates instability in the OLS estimator.
3. Say whether the following statements are true or false and explain why.
   (a) For *any* set of predictor variables, the larger the number of predictor variables in the model, the larger the R^2.
   (b) For model of the same size (fixed p), their C_p, AIC_p, BIC_p values are monotonically increasing in terms of SSE_p.
   (c) Compared with AIC, BIC criterion tends to select smaller models because it puts higher penalties on model size.
   (d) The best subsets procedure is guaranteed to find the "best" model under a given criterion.

Coding Questions

1. For the dataset `KelleyBlueBookData.csv`, consider using price as the response and regressing against the following predictors: mileage, type, cylinder, liter, cruise, sound, and leather. In this exercise, treat Leather (0 for not-leather, 1 for leather), Type and Cylinder as categorical variables.
   (a) Report the estimated coefficient of "leather" and interpret the t test result for testing whether or not there is a leather effect. Interpret in the context of the problem to comment on the impact of "leather" to price.
   (b) Look at the coefficients associated with the "Type" variable. Which type was used as the reference level? Which type seems to have the highest average price?
   (c) What conclusion you can make about the price when Cylinder=6 compared to other cylinder levels?
   (d) Run a partial ANOVA, interpret the F test result for Cylinder. Combine the results of t-test and F-test for Cylinder, should we conclude that it's a significant predictor?
2. Download the data set `IceCreamConsumption.csv` and regress cons against income, price, and temp.
   (a) Obtain the variance inflation factors. What do these suggest about the effects of multicollinearity in this model?
   (b) Explain how the VIF for income is calculated step by step.

(c) Draw an influence plot of this model where x-axis is the leverage, y-axis is the (externalaly) studentized residuals, and the size of the points are Cook's distance. Which observation has the highest studentized residual? Which observation has the highest leverage? Which observation has highest Cook's distance?

3. For the data set IceCreamConsumption.csv and consider y=cons with predictors income, price, and temp.

   (a) List all the possible models from this data set (without interactions or higher powers).

   (b) Calculate the adjusted R^2 and C_p for all the models, make a summary table with four columns: Number of predictors, R_a^2 values, C_p values, Predictors in the model.

   (c) Based on the table above, which model is selected by R_a^2? By C_p?

   (d) For the two models in part c., calculate the AIC and BIC values. Based on AIC and BIC, what's your final choice of model?

   (e) Is there a difference in the size of the model selected by AIC and BIC? If yes, state which is more parsimonious and explain why this difference exists.

4. Consider the data set `BrandPreference.csv` and a model in which we regress BrandLiking (scale 0-100, 100 being most preferred) against MoistureContent (scale 1-10, 10 being most moist) and Sweetness (scale 1-5, 5 being sweetest). Treat both predictors as numerical variables.

   (a) Perform the regression in python and write down the fitted model.

   (b) Find the fitted value $\hat{y}_1$ for the first observation of the data. Hint: Don't forget that Python indices start at 0.

   (c) Calculate the hat matrix H, and show that

$$\hat{y}_1 = \sum_{i=1}^{n} h_{1i} y_i.$$