STATS 415, Homework 3

Due Thursday Sep 26, 2019

Turn in a printout of your homework in class on Sep 26.

1. Check whether each of the following claims is true or false. If the claim is true, prove it; otherwise, give a counter example to show that it is indeed false. (5pt per claim)

   (a) When two random variables X and Y are independent of each other, they can be correlated sometimes.

   (b) Under the assumption of the linear model that $E(\epsilon|\mathbf{x}) = 0$, ϵ is independent of $\mathbf{x}$.

   (c) Under the same assumption as in (b), ϵ and $\mathbf{x}$ are uncorrelated.

   (d) Given a n-by-p design matrix $\mathbf{X}$ with full column rank, the projection matrix $\mathbf{P_X} := \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$ is symmetric.

   (e) Following the definition of $\mathbf{P_X}$ in (d), $(\mathbf{I} - \mathbf{P_X})^{100} = \mathbf{I} - \mathbf{P_X}$.

   (f) Under the linear model, when the observations are i.i.d., the residuals of the OLS $\{y_i - \widehat{y}_i\}_{i=1}^n$ are i.i.d. too.

   (g) Under the linear model, the OLS residual vector $\widehat{\boldsymbol{\epsilon}} := (\hat{\epsilon}_1, \ldots, \hat{\epsilon}_n)^\top$ satisfies that $\widehat{\boldsymbol{\epsilon}} \perp \mathbf{X}_j$ for all $j = 1, \ldots, p$, where $\mathbf{X}_j$ is the jth column of $\mathbf{X}$.

2. This exercise relates to the `Carseats` data set in the `ISLR` package, the same dataset you used for Homework 2. Before you proceed, divide the data into training and test sets, using the first 80% of the observations as training data, and the remaining 20% as test data. (10pt for each question)

   (a) Fit a multiple regression model to predict `Sales` using all other variables (model 1), and a reduced model with everything except for `Population, Education, Urban, and US` (model 2), using only the training data to estimate the coefficients. For both models, report their training and test errors. Comment on how they differ.

(b) Suppose we fit KNN regression to predict `Sales` from the variables used in model 2, except for `ShelveLoc`. Without computing anything, can you tell whether the training error will be lower for $K = 1$ or for $K = 20$? How about the test error? Explain your answer (without computing anything).

(c) Fitting a KNN regression requires computing distances between data points. Would you standardize the variables in this dataset first? Why or why not? However you answer, provide some data-based supporting evidence to justify your choice.

(d) Fit the KNN regression to predict `Sales` from the variables used in model 2, except for `ShelveLoc`. Plot the training and test errors as a function of K. Report the value of K that achieves the lowest training error and the lowest test error. Comment on the shape of the plots.

(e) Make a plot of residuals against fitted values for both model 2 and KNN regression with K of your choice, for the *test data*. Make sure the scale of the axes is the same in both plots. Comment on any similarities or differences.

Please limit your answer to Q2 to 8 pages, organized into a coherent typed data analysis report. Answers to Q1 may be either typed or handwritten. Please staple everything together and clearly write your name, your UMID, and your GSI/lab number on the homework.