STATS 415, Homework 3

Due Thursday Oct 1st, 2020

Note: Turn in a pdf scan of your homework on Canvas. Please limit your answer to Q3 to 8 pages, organized into a coherently typed data analysis report. Answers to Q1 and Q2 maybe either typed or handwritten. Please clearly write yourname, your UMID, and your GSI/lab number on the home-work.

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.

   (a) When two random variables X and Y are independent of each other, they are uncorrelated as well. (5 points)

   (b) Under the assumption of the linear model that $\mathrm{E}(\epsilon|\mathbf{x}) = 0$, ϵ is independent of $\mathbf{x}$. (5 points)

   (c) Under the same assumption as in (b), ϵ and $\mathbf{x}$ are uncorrelated. (5 points)

   (d) Given a n-by-p design matrix $\mathbf{X}$ with full column rank, the projection matrix $\mathbf{P}_{\mathbf{X}} := \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$ is symmetric. (5 points)

   (e) Following the definition of $\mathbf{P}_{\mathbf{X}}$ in (d), $(\mathbf{I} - \mathbf{P}_{\mathbf{X}})^{100} = \mathbf{I} - \mathbf{P}_{\mathbf{X}}$. (5 points)

   (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. (5 points)

   (g) Under the linear model, the OLS residual vector $\hat{\boldsymbol{\epsilon}} := (\hat{\epsilon}_1, \ldots, \hat{\epsilon}_n)^\top$ satisfies that $\hat{\boldsymbol{\epsilon}} \perp \mathbf{X}_j$ for all $j = 1, \ldots, p$, where $\mathbf{X}_j$ is the jth column of $\mathbf{X}$. (5 points)

2. In this exercise, we are going to explicitly derive the testing MSE of the ordinary least squares estimator $\hat{\boldsymbol{\beta}}$ under a linear model. Consider the linear model that $Y = \mathbf{x}^\top\boldsymbol{\beta}^* + \epsilon$, where $\mathbb{E}(\epsilon|\mathbf{x}) = 0$ and $\mathrm{Var}(\epsilon|\mathbf{x}) = \sigma^2$. Suppose we have n independent and identically distributed observations $\{(\mathbf{x}_i, y_i)\}_{i\in[n]}$ from this model as our training sample. Write the design matrix as $\mathbf{X} := (\mathbf{x}_1, \ldots, \mathbf{x}_n)^\top$ and the response vector as $\mathbf{y} := (y_1, \ldots, y_n)^\top$. The OLS estimator is defined as $\hat{\boldsymbol{\beta}} := (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$.

(a) Derive $\text{cov}(\hat{\boldsymbol{\beta}})$ explicitly in terms of $\mathbf{X}$ and σ^2. (5 points)

(b) Given a new data point that is independent of the training sample and has $\mathbf{z}$ as its feature vector, what should be the prediction for its response using OLS? Is the prediction unbiased? (5 points)

(c) Derive the variance of the prediction in the previous question and then the testing MSE of $\hat{\boldsymbol{\beta}}$. (5 points)

3. 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. (10 points 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.