Statistics 206

Homework 7

Due : Friday, Nov. 18, 2022, 11:59PM

Instructions:

- You should upload homeworkX files on canvas (under "Assignments/hwX") before its due date.
- Your homework may be prepared by a word processor (e.g., Latex) or through handwriting.
- For handwritten homework, you should either scan or take photos of your homework: Please make sure the pages are clearly numbered and are in order and the scans/photos are complete and clear; Check before submitting.
- Please name the files following the format: "FirstName-LastName-HwX". If there are several files, you can use "-Questions1-5", "-Questions6", etc., to distinguish them. E.g., "Jie-Peng-Hw1-Questions1-5.pdf", "Jie-Peng-Hw1-Questions6.html".
- Your name should be clearly shown on the submitted files: By putting on your name, you also acknowledge that you are the person who did and prepared the submitted homework.
- **Optional Problems** are more advanced and are not counted towards the grade.
- Showing/sharing/uploading homework or solutions outside of this class is prohibited.

1. Tell true or false of the following statements. Provide a brief justification for you answer.

   (a) To quantify a qualitative variable with three classes C_1, C_2, C_3, we need the following dummy variables:

   $$X_1 = \begin{cases} 1 & if \quad C_1 \\ 0 & if \quad otherwise \end{cases} \quad X_2 = \begin{cases} 1 & if \quad C_2 \\ 0 & if \quad otherwise \end{cases} \quad X_3 = \begin{cases} 1 & if \quad C_3 \\ 0 & if \quad otherwise \end{cases}$$

   (b) Polynomial regression models with higher than the third power terms are preferred since they provide better approximations to the regression relation.

   (c) In interaction regression models, the effect of one variable depends on the value of another variable with which it appears together in a cross-product term.

   (d) With a qualitative variable, the best way is to fit separate regression models under each of its classes.

2. **(Cars) Exploratory Data Analysis**. You need to submit your codes alongside with the answers, plots, outputs, etc. You are required to use R Markdown: Please submit a .rmd file **and** its corresponding .html file.

   (a) Conduct a visual inspection of the data in "Cars.csv" and then read the data into R.

   (b) Are there missing values? If so, replace missing values by "NA".

   (c) Check the variable types. Which variables do you think should be treated as quantitative and which should be treated as qualitative/categorical? Fix the problems that you have identified (if any).

   (d) Draw histogram for each quantitative variable. Comment on their distributions.

   (e) Draw scatter plot matrix among quantitative variables with the lower panel showing correlation coefficients. Comment on their relationships.

   (f) Draw pie chart (with class percentage) for each categorical variable.

   (g) Draw side-by-side box plots for "mpg" with respect to each categorical variable. What do you observe?

3. **(Cars Cont'd) Regression with Categorical Variables**. In this question, we consider models for "mpg" using "cylinders", "horsepower", and "weight" as predictors, where "cylinders" should be treated as a categorical variable. You need to submit your codes alongside with the answers, plots, outputs, etc. You are required to use R Markdown: Please submit a .rmd file **and** its corresponding .html file.

   (a) Decide on whether you'd like to make any transformation of the "mpg".

   (b) Fit a first-order model with the (transformed) variables. Conduct model diagnostics. Does this model appear to be adequate?

   (c) Derive the regression function for cars with 4 cylinders.

   (d) Fit a model including interactions between "cylinders" and "horsepower", and, "cylinders" and "weight". Derive the regression function for cars with 4 cylinders.

   (e) Compare the two models using the function `anova()`. What do you find?

   (f) Construct a 95% prediction interval of "mpg" for a car with 4 cylinders, 100 horsepower and 3000 pounds under these two models. What do you observe?

4. **(Optional problem). Regression coefficients as partial coefficients.** Let $X = (X_1, X_2)$ where $X_1 \in \mathbb{R}^{n\times s}, X_2 \in \mathbb{R}^{n\times t}$. Write the LS fitted regression coefficients as $\hat{\beta} = \begin{pmatrix} \hat{\beta}^{(1)} \\ \hat{\beta}^{(2)} \end{pmatrix}$. Show that:

   (a) The LS fitted regression coefficients of X_2 is

$$\hat{\beta}^{(2)} = (\tilde{X}_2^T \tilde{X}_2)^{-1} \tilde{X}_2^T Y = (\tilde{X}_2^T \tilde{X}_2)^{-1} \tilde{X}_2^T (Y - \hat{Y}(X_1)), \quad \tilde{X}_2 = X_2 - \hat{X}_2(X_1),$$

   i.e., $\hat{\beta}^{(2)}$ *is the LS fitted regression coefficients by regressing* Y *(or* $Y - \hat{Y}(X_1)$*) onto* $X_2 - \hat{X}_2(X_1)$. Such coefficients are called **partial coefficients.**

(b) If $X_1 \perp X_2$ (i.e., the columns of X_1 and the columns of X_2 are orthogonal), then

$$\hat{\beta}^{(2)} = (X_2^T X_2)^{-1} X_2^T Y, \quad if, \quad X_1 \perp X_2,$$

i.e., the LS fitted regression coefficients by regressing Y onto X_2 alone.

5. **(Optional problem). Simultaneous confidence bands of the regression function.** Under the Normal error model, derive the simultaneous confidence bands of the regression function by the following steps.

   (a) Show that
   $$\frac{(\hat{\beta} - \beta)^T (X^T X)(\hat{\beta} - \beta)}{MSE} \sim pF_{p,n-p}.$$

   (b) Show that for a constant $C \geq 0$, $|x^T \beta - x^T \hat{\beta}| \leq \sqrt{C x^T (X^T X)^{-1} x}$ for all $x \in \mathbb{R}^p$ if and only if $(\hat{\beta} - \beta)^T (X^T X)(\hat{\beta} - \beta) \leq C$.

   (c) Show that the $(1-\alpha)100\%$ simultaneous confidence bands for the regression function, $x^T \beta, x \in \mathbb{R}^p$, are:
   $$x^T \hat{\beta} \pm \sqrt{pF(1-\alpha; p, n-p))} \sqrt{MSE x^T (X^T X)^{-1} x}, \quad x \in \mathbb{R}^p,$$
   i.e.,
   $$P(x^T \beta \in x^T \hat{\beta} \pm \sqrt{pF(1-\alpha; p, n-p))} \sqrt{MSE x^T (X^T X)^{-1} x}, \text{ for all } x \in \mathbb{R}^p) = 1-\alpha.$$