Statistics 206

Homework 7

Due : Dec. 3, 2018, In Class

1. Tell true or false of the following statements.

   (a) To quantify a qualitative variable with three classes C_1, C_2, C_3, we need the following indicator 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-order powers (e.g., higher than the third power) 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.

   (e) With many potential X variables, we can first fit a model with all these variables, and then drop those having non-significant regression coefficients by t-tests.

   (f) A correct model must be a good model.

   (g) With too many nuisance X variables, the model tends to have a large model bias.

2. **(Homework 5 Continued). Polynomial Regression.** You should use R and the *lm()* function and its associated functions (e.g., *summary(), anova(), confint(), predict.lm()*) to do this problem. Please also attach your R codes and plots.

   A commercial real estate company evaluates age (X_1), operating expenses (X_2, in thousand dollar), vacancy rate (X_3), total square footage (X_4) and rental rates (Y, in thousand dollar) for commercial properties in a large metropolitan area in order to provide clients with quantitative information upon which to make rental decisions. The data are taken from 81 *suburban commercial properties.* (The data is on canvas under Files/Homework/property.txt; The first column is Y, followed by X_1, X_2, X_3, X_4.)

   Based on the analysis from Homework 5 the vacancy rate (X_3) is not important in explaining the rental rates (Y) when age (X_1), operating expenses (X_2) and square footage (X_4) are included in the model. So here we will use the latter three variables to build a regression for rental rates.

   (a) Plot rental rates (Y) against the age of property (X_1) and comment on the shape of their relationship.

(b) Fit a polynomial regression model with linear terms for centered age of property ($\tilde{X}_1$), operating expenses (X_2), and square footage (X_4), and a quadratic term for centered age of property ($\tilde{X}_1$). Write down the model equation. Obtain the fitted regression function and also express it in terms of the original age of property X_1. Draw the observations Y against the fitted values $\widehat{Y}$ plot. Does the model provide a good fit?

(c) Compare R^2, R_a^2 of the above model with those of Model 2 from Homework 5 ($Y \sim X_1 + X_2 + X_4$). What do you find?

(d) Test whether or not the quadratic term for centered age of property ($\tilde{X}_1$) may be dropped from the model at level 0.05. State the null and alternative hypotheses, the test statistic, its null distribution, the decision rule and the conclusion.

(e) Predict the rental rates for a property with $X_1 = 4, X_2 = 10, X_4 = 80,000$. Construct a 99% prediction interval and compare it with the prediction interval from Model 2 of Homework 5.

3. **Exploratory data analysis and preliminary investigation**. Read the data in "Car.csv" on canvas /Files/Homework into R:

```
cars = read.csv('Cars.csv', header=TRUE)
```

Consider building a model for "mpg" .

(a) Draw the scatterplot matrix of this data. Do you observe something unusual?

(b) Check the variable type for each variable. Do you observe something unusual? Which variables do you think should be treated as quantitative and which ones should be treated as qualitative/categorical?

(c) Fix the problems that you have identified (if any) before proceeding to the next question.

(d) Draw histogram for each quantitative variable. Do you think any transformation is needed ? If so, make the transformation before proceeding to the next question.

(e) Draw the scatter plot matrix among quantitative variables (possibly transformed). Do you observe any nonlinear relationship with the response variable? If so, what should you do?

(f) Draw pie chart for each categorical variable. Draw side-by-side box plots for the response variable with respect to each categorical variable. What do you observe?

(g) Decide on a model for further investigation. Fit this model and draw residual plots. Does the model seem to be adequate? If not, try to make adjustments and fit an updated model. Repeat this process until you think you have found an adequate model. What would be your next step then?

4. **Polynomial Regression**. Write down model equations for the following models.

(a) A third-order polynomial regression model with one predictor.

(b) A second-order polynomial regression model with K predictors.

5. **Bias-variance trade-off**. Consider the following simulation study. You can modify the codes in bias-variance-trade-off-simulation2.R under canvas /Files/Homework. You should read the codes carefully and use R help whenever necessary to understand the codes.

   - The true regression function is
   $$f(x) = sin(x) + sin(2x).$$
   - The sample size is $n = 30$ and the design points X_i are equally spaced on $[-3, 3]$.
   - The models to be considered are polynomial regression models with order $l = 1, 2, 3, 5, 7, 9$.
   - The observed data are generated according to:
   $$Y_i = f(X_i) + \epsilon_i, \quad i = 1, \cdots, n, \quad \epsilon_i \sim_{i.i.d.} N(0, \sigma^2).$$
   - Consider three different noise levels with $\sigma = 0.5, \quad 2, \quad 5$.
   - Generate 1000 independent sets (replicates) of observations under each noise level.

   Answer the following questions and include relevant plots along with your answers.

   (a) Is there a correct model among the models being considered?Explain your answer.

   (b) What is the (in-sample) model variance for each of these models? Does the model variance change with the error variance?

   (c) Comment on the (in-sample) model bias for each of these models. Does the model bias change with the error variance?

   (d) Which one, the model variance or the model bias, is the dominant component in the (in-sample) mean-squared-estimation-error? Does the answer depend on the error variance and why?

   (e) Which model is the best model according to the mean-squared-estimation-error? Does the answer depend on the error variance and why?

   (f) Comment on $E(SSE)$. Do you observe different patterns under different noise levels? Given an explanation.

6. **(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 $C \geq 0$, $|x^T\beta - x^T\hat{\beta}| \leq \sqrt{Cx^T(X^TX)^{-1}x}$ for all $x \in \mathbb{R}^p$ if and only if $(\hat{\beta}-\beta)^T(X^TX)(\hat{\beta}-\beta) \leq C$.

(c) Show that the $(1-\alpha)$ simultaneous confidence bands for the regression function, $x^T\beta, x \in \mathbb{R}^p$, are:

$$x^T\hat{\beta} \pm \sqrt{pF(1-\alpha;p,n-p))}\sqrt{MSEx^T(X^TX)^{-1}x}, \quad x \in \mathbb{R}^p,$$

i.e.,

$$P(x^T\beta \in x^T\hat{\beta} \pm \sqrt{pF(1-\alpha;p,n-p))}\sqrt{MSEx^T(X^TX)^{-1}x}, \text{ for all } x \in \mathbb{R}^p) = 1-\alpha.$$

7. **(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

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

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

$$\hat{\beta}^{(2)} = (X_2^TX_2)^{-1}X_2^TY, \quad if, \quad X_1 \perp X_2,$$

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