Statistics 206

Homework 6

Due : Nov. 16, 2015, In Class

1. *(Homework 4, Problem 4 Continued)* **Partial coefficients and added-variable 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 smartsite under Resources/Homework/property.txt; The first column is Y, followed by X_1, X_2, X_3, X_4.)

   (a) Perform regression of the rental rates Y on the four predictors X_1, X_2, X_3, X_4 (Model 1). (Hint: To help answer the subsequent questions, the predictors should enter the model in the order X_1, X_2, X_4, X_3.)

   (b) Based on the R output of Model 1, obtain the fitted regression coefficient of X_3 and calculate the coefficient of partial determination $R^2_{Y3|124}$ and partial correlation $r_{Y3|124}$. Explain what $R^2_{Y3|124}$ measures and interpret the result.

   (c) Draw the added-variable plot for X_3 and make comments based on this plot.

   (d) Regressing the residuals $e(Y|X_1, X_2, X_4)$ to the residuals $e(X_3|X_1, X_2, X_4)$. Compare the fitted regression slope from this regression with the fitted regression coefficient of X_3 from part (b). What do you find?

   (e) Obtain the regression sum of squares from part (d) and compare it with the extra sum of squares $SSR(X_3|X_1, X_2, X_4)$ from the R output of Model 1. What do you find?

   (f) Calculate the correlation coefficient r between the two sets of residuals $e(Y|X_1, X_2, X_4)$ and $e(X_3|X_1, X_2, X_4)$. Compare it with $r_{Y3|124}$. What do you find? What is r^2?

   (g) Regressing Y to the residuals $e(X_3|X_1, X_2, X_4)$. Compare the fitted regression slope from this regression with the fitted regression coefficient of X_3 from part (b). What do you find? Can you provide an explanation?

2. **Bias-variance trade-off**. Consider the following simulation study. You can modify the codes in bias-variance-trade-off-simulation.R under the smartsite/Resources/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.

3. **Exploratory data analysis and preliminary investigation**. Read the data in "Car.csv" on smartsite/Resources/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. (Hint: Check out Lab5 handout)

(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?