Statistics 206

Homework 5

Due : Nov. 9, 2015, In Class

1. Tell true or false of the following statements.

   (a) If the response variable is uncorrelated with all X variables in the model, then the least-squares estimated regression coefficients of the X variables are all zero.

   (b) Even when the X variables are perfectly correlated, we might still get a good fit of the data.

   (c) Taking correlation transformation of the variables will not change coefficients of multiple determination.

   (d) If all the X variables are uncorrelated, then the magnitude and the sign of a standardized regression coefficient reflect the comparative importance and direction of effect, respectively, of the corresponding X variable, in terms of explaining the response variable.

   (e) In a regression model, it is possible that none of the X variables is statistically significant when being tested individually, while there is a significant regression relation between the response variable and the set of X variables.

   (f) If an X variable is uncorrelated with the rest of the X variables, then in the standardized model, the variance of its least-squares estimated regression coefficient equals to the error variance.

   (g) If an X variable is uncorrelated with the response variable, then its least-squares estimated regression coefficient must be zero.

   (h) If an X variable is uncorrelated with the response variable and also is uncorrelated with the rest of the X variables, then its least-squares estimated regression coefficient must be zero.

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

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

   (j) 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.

   (k) 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.

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

2. **(Homework 4, Problem 4 Continued). Standardized Regression model.** 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 smartsite under Resources/Homework/property.txt; The first column is Y, followed by X_1, X_2, X_3, X_4.)

   (a) Draw histogram and boxplot and obtain a summary for each variable. Comment on the distributions of these variables. Do the X variables appear to be in the same scale?

   (b) Calculate the sample mean and sample standard deviation of each variable. Perform the correlation transformation. What are sample means and sample standard deviations of the transformed variables?

   (c) Write down the model equation for the the standardized first-order regression model with all four transformed X variables and fit this model. What is the fitted regression intercept? Transform the fitted standardized regression coefficients back to the fitted regression coefficients of the original model. Do you get the same results as those from Homework 4, Problem 4?

   (d) Obtain the standard errors of the fitted regression coefficients of the original model using the standard errors of the fitted standardized regression coefficients. Compare the results with those from the R output of Problem 4 of Homework 4.

   (e) Obtain SSTO, SSE and SSR under the standardized model and compare them with those from the original model (Problem 4 of Homework 4). How are they related to each other?

   (f) Calculate R^2, R_a^2 under the standardized model and compare them with R^2, R_a^2 under the original model (Problem 4 of Homework 4). What do you find?

3. **(Homework 4, Problem 4 Continued). Multicollinearity.**

   (a) Obtain the correlation matrices $\mathbf{r}_{XX}$ of the four X variables and $\mathbf{r}_{XY}$ between the response variable and the four X variables. Confirm that for the standardized model, $\mathbf{X'X} = \mathbf{r}_{XX}$ and $\mathbf{X'Y} = \mathbf{r}_{XY}$.

   (b) Obtain $\mathbf{r}_{XX}^{-1}$ and get the variance inflator factors VIF_k $(k = 1, 2, 3, 4)$. Obtain R_k^2 by regressing X_k to $\{X_j : 1 \le j \ne k \le 4\}$ $(k = 1, 2, 3, 4)$. Confirm that

$$VIF_k = \frac{1}{1 - R_k^2}, \quad k = 1, 2, 3, 4.$$

   Comment on the degree of multicollinearity in this data.

(c) Fit the regression model for relating Y to X_4 and fit the regression model for relating Y to X_3, X_4. Compare the estimated regression coefficients of X_4 in these two models. What do you find? Calculate $SSR(_4)$ and $SSR(X_4|X_3)$. What do you find? Provide an interpretation for your observations.

(d) Fit the regression model for relating Y to X_2 and fit the regression model for relating Y to X_2, X_4. Compare the estimated regression coefficients of X_2 in these two models. What do you find? Calculate $SSR(_2)$ and $SSR(X_2|X_4)$. What do you find? Provide an interpretation for your observations.

4. **(Homework 4, Problem 4 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.

   Based on the analysis from Homework 4, Problem 4, 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 4, Problem 4 ($Y \sim X_1 + X_2 + X_4$). 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 4, Problem 4.

5. **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.

6. **Uncorrelated X variables**. When $X_1, \cdots, X_{p-1}$ are uncorrelated, show the following results. (Hint: Show these results under the standardized regression model and then transform them back to the original model.)

(a) The fitted regression coefficients of regressing Y on $(X_1, \cdots, X_{p-1})$ equal to the fitted regression coefficients of regressing Y on each individual X_j $(j = 1, \cdots, p-1)$ alone.

(b) Let $\mathcal{I} := \{k : 1 \leq k \leq p-1, k \neq j\}$. Show that

$$SSR(X_j|X_{\mathcal{I}}) = SSR(X_j),$$

where $SSR(X_j)$ denotes the regression sum of squares when regressing Y on X_j alone.

7. **Variance Inflation Factor for models with 2 X variables.** Show that for a model with two X variables, X_1 and X_2, the variance inflation factors are

$$VIF_1 = VIF_2 = \frac{1}{1 - r_{12}^2},$$

where r_{12} is the sample correlation coefficient between X_1 and X_2. (Hint: In this case, $r_{12}^2 = R_1^2 = R_2^2$.)

8. **(Optional Problem) Variance Inflation Factor.** Use the formula for the inverse of a partitioned matrix to show:

$$r_{XX}^{-1}(k,k) = \frac{1}{1 - R_k^2},$$

i.e., the kth diagonal element of the inverse correlation matrix equals to $\frac{1}{1-R_k^2}$, where R_k^2 is the coefficient of multiple determination by regressing X_k to the rest of the X variables.

Hints: (i) Assume all X variables are standardized by the correlation transformation; (ii) You only need to prove this for $k = 1$ because you can permute the rows and columns of r_{XX} and r_{XY} to get the result for other k; (iii) Apply the inverse formula below with $A = r_{XX}$ and $A_{11} = r_{11}$, i.e., the first diagonal element of r_{XX}.

Inverse of a partitioned matrix. Suppose A is a $(p+q) \times (p+q)$ square matrix $(p, q \geq 1)$:

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix},$$

where A_{11} is a $p \times p$ square matrix and A_{22} is a $q \times q$ square matrix. Suppose A_{11} and A_{22} are invertible. Then A is invertible and

$$A^{-1} = \begin{bmatrix} \left(A_{11} - A_{12}A_{22}^{-1}A_{21}\right)^{-1} & -\left(A_{11} - A_{12}A_{22}^{-1}A_{21}\right)^{-1} A_{12}A_{22}^{-1} \\ -\left(A_{22} - A_{21}A_{11}^{-1}A_{12}\right)^{-1} A_{21}A_{11}^{-1} & \left(A_{22} - A_{21}A_{11}^{-1}A_{12}\right)^{-1} \end{bmatrix}$$