Statistics 206

Homework 2

Due : Monday, October 12, 2015, In Class

1. Please tell true or false of the following statements.

   (a) If all observations Y_i fall on one straight line, then the coefficient of determination $R^2 = 1$.

   (b) A large R^2 always means that the fitted regression line is a good fit of the data, while a small R^2 always means that the predictor and the response are not related.

   (c) The scatter plot is the most effective graph to show a nonlinear relationship between two variables.

   (d) The regression sum of squares SSR tends to be large if the estimated regression slope is large in magnitude or the dispersion of the predictor values is large.

2. Under the simple linear regression model, show that the residuals e_i's are uncorrelated with the LS estimators $\hat{\beta}_0$ and $\hat{\beta}_1$, i.e., $Cov(e_i, \hat{\beta}_0) = 0$, $Cov(e_i, \hat{\beta}_1) = 0$ $(i = 1, \cdots, n)$.

3. Under the Normal error model:

   (a) Show that SSE is independent with the LS estimators $\hat{\beta}_0$ and $\hat{\beta}_1$.

   (b) Show that SSE and SSR are independent. (Hint: Recall 6(a) of homework 1)

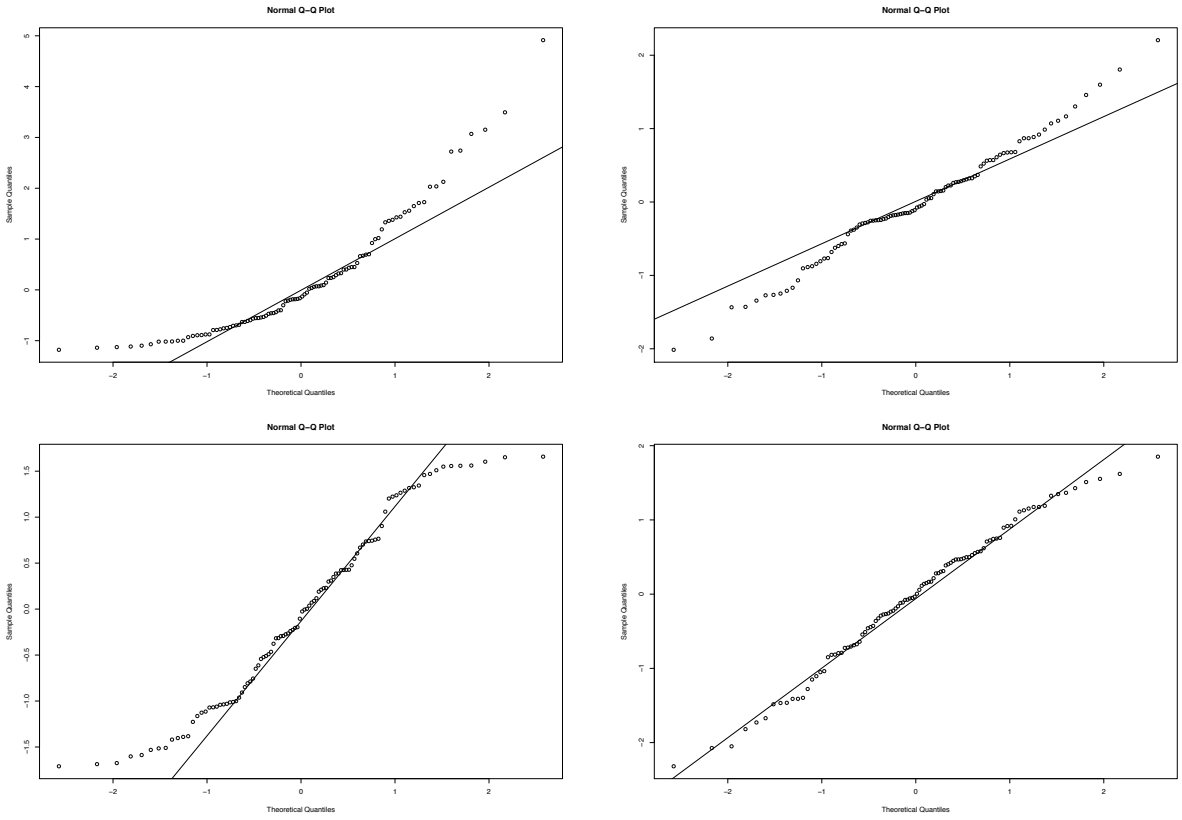
4. Under the simple linear regression model, derive $Var(\widehat{Y}_h)$, where

$$\widehat{Y}_h = \hat{\beta}_0 + \hat{\beta}_1 X_h$$

   is the estimator of the mean response $\beta_0 + \beta_1 X_h$. (Hint: Recall 4 (b) of homework 1: $\bar{Y}$ and $\hat{\beta}_1$ are uncorrelated.)

5. Q-Q plots. For each of the Q-Q plot in Figure 1, describe the distribution of the data (whether it is Normal, heavy tailed, light tailed, skewed to right or skewed to left, etc.).

Figure 1: Q-Q plots



6. **A simple linear regression case study by R**. You must use R and the *lm* function and its associated functions to do this problem. Please also attach your R codes.

   A person's muscle is expected to decrease with age. To explore this relationship in women, a nutritionist randomly selected 15 *women from each of the four* 10*-year age groups, beginning with age* 40 *and ending with age* 79*. Two variables being measured are: age (X) and the amount of muscle mass (Y).* (The data is on smartsite under Resources/Homework/muscle.txt.)

   (a) Read data into R. Draw histogram for muscle mass and age, respectively. Comment on their distributions. Draw the scatter plot of muscle mass versus age. Do you think their relation is linear? Does the data support the anticipation that the amount of muscle mass decreases with age?

   (b) Use the box-cox procedure to decide whether a transformation of the response variable is needed.

   (c) Perform linear regression of the amount of muscle mass on age and obtain a summary. From the summary, obtain the estimated regression coefficients and their standard errors, the mean squared error (MSE) and its degrees of freedom.

   (d) Write down the fitted regression line. Add the fitted regression line to the scatter plot. Does it appear to fit the data well?

   (e) Obtain the fitted values and residuals for the 6th and 16th cases in the data set.

   (f) Draw the residuals vs. fitted values plot and the residuals Normal Q-Q plot. Write down the simple linear regression model with Normal errors and its assumptions. Comment on these assumptions based on the residual plots.

   (g) Construct a 99% confidence interval for the estimated regression intercept. Interpret your confidence interval.

   (h) Conduct a test at level 0.01 to decide whether or not there is a negative linear association between the amount of muscle mass and age. State the null and alternative hypotheses, the test statistic, its null distribution, the decision rule and the conclusion. (Hint: Which alternative you should use?)

   (i) Construct a 95% prediction interval for the muscle mass of a woman aged at 60. Interpret your prediction interval.

   (j) Obtain the ANOVA table for this data. Test whether or not there is a linear association between the amount of muscle mass and age by an F test at level 0.01. State the null and alternative hypotheses, the test statistic, its null distribution, the decision rule and the conclusion.

   (k) What proportion of the total variation in muscle mass is "explained" by age? What is the correlation coefficient between muscle mass and age?