

Statistics 206

Homework 3

Due : October 16, 2019, In Class

1. Tell true or false of the following statements and provide a brief explanation to your answer.
 - (a) If all observations Y_i fall on one straight line, then the coefficient of determination $R^2 = 1$.
 - (b) A large R^2 means that the fitted regression line is a good fit of the data, while a small R^2 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:
 - (a) Show that the regression sum of squares

$$SSR = \hat{\beta}_1^2 \sum_{i=1}^n (X_i - \bar{X})^2.$$

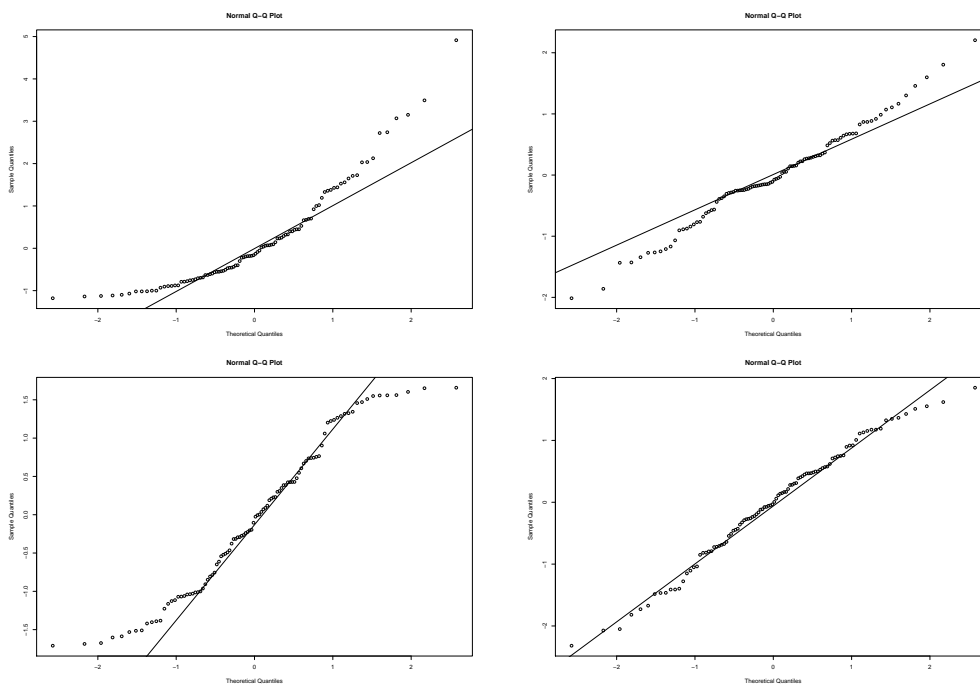
- (b) Derive $E(SSR)$.
3. Under the Normal error model: Show that SSE and SSR are independent.
4. Perform ANOVA on the “crime rate and education” data (from homework 2).
 - (a) Calculate sum of squares: $SSTO$, SSE and SSR . What are their respective degrees of freedom?
 - (b) Calculate the mean squares.
 - (c) Summarize results from parts (a) and (b) into an ANOVA table.
 - (d) Assume Normal error model. Use the F test to test whether or not there is a linear association between crime rate and percentage of high school graduates at significance level 0.01. State the null and alternative hypotheses, the test statistic, its null distribution, the decision rule and the conclusion.
 - (e) Compare your calculation from part (d) with those from part (e) of Problem 6 of homework 2. What do you find?

5. **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 canvas under Files/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 form of alternatives should you 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?
6. **Q-Q plots.** For each of the Q-Q plot in Figure 1, describe the distribution of the data (whether it is Normal or heavy tailed, etc.).

Figure 1: Q-Q plots



7. **Coefficient of determination** . Show that

$$R^2 = r^2, \quad r = \text{sign}\{\hat{\beta}_1\}\sqrt{R^2},$$

where R^2 is the coefficient of determination when regressing Y onto X and r is the sample correlation coefficient between X and Y .