DSCI 6619: Assignment 1 - Due Wednesday, October 9, 2024

Instructions

(a) You may use R or OTHER SOFTWARE. Do not submit your codes.

(b) **A printout of the section of your output showing ONLY the results you used to answer the questions must be submitted if a software was used.**

(c) **Your answers to each of the questions must be handwritten or typed on a sheet separate from your computer printout.**

(d) For clarity, **DO NOT USE A PENCIL TO WRITE YOUR ANSWERS. USE A PEN.**

1. **Crime rate.** A criminologist studying the relationship between level of education and crime rate in medium-sized U.S. counties collected data from a random sample of 84 counties; X is the percentage of individuals in the county having at least a high-school diploma, and Y is the crime rate last year. The data can be found on the course web page. Assume that the simple linear regression model is appropriate for this data.

   (a) Obtain the estimated regression function.

   (b) Construct a scatterplot of the data and a plot of the fitted regression function on the same graph. Does the linear regression function appear to give a good fit here? Explain.

   (c) Suppose you which to test whether the difference in the mean crime rate for two counties whose high-school graduation rates differ by one percentage point is significant,

      i. state the null and alternative hypotheses.

      ii. show that the student t statistic for the test can be written as

      $$t_0 = \frac{SS_{xy}}{s\sqrt{SS_{xx}}},$$

      where, $s^2 = MSE$.

      iii. use the p-value to conduct the test and state your conclusion in plain language.

      iv. Set up the ANOVA table and test whether there is a linear association between crime rate and percentage of high school graduates. What is estimated by the MSR in your ANOVA table? by MSE? Under what condition do MSR and MSE estimate the same quantity?

2. The number of galleys for a manuscript (x) and the dollar cost of correcting typographical errors (y) in a random sample of recent orders handled by a firm specializing in technical manuscripts can be found on the D2L shell for Stat 6619. Assume that the regression model,

   $$y_i = \beta_1 x_i + \epsilon_i,$$

   is appropriate, with normally distributed independent error terms whose variance is $\sigma^2 = 16$.

(a) Write down the normal equation for least squares estimation of the effect of x on y for this data.

(b) Solve the normal equation and use the data to compute an estimate of the effect of x on y.

(c) Construct the ANOVA Table for the manuscript data based on the assumed model. Compute the bias of $\hat{\sigma}^2$.

(d) Test whether $\beta_1 = 0$ against a one-sided alternative using a t-test and the p-value of the test.

3. Five observations on an outcome variable y are to be measured at five (5) values of a predictor variable $x = 4, 8, 12, 16,$ and 20, respectively. The true regression function is $E(y) = 20 + 4x$ and the error terms ϵ_i are independent $N(0, 25)$.

   (a) Generate five normal random numbers, with mean 0 and variance 25. Consider these random numbers as the error terms for the five y observations at $x = 4, 8, 12, 16,$ and 20 and calculate $y_1, y_2, y_3, y_4,$ and y_5. Obtain the least squares estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ when fitting $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, $i = 1, \ldots, 5$ to the five cases.

   (b) Calculate $\hat{y}_h$ when $x_h = 10$. Obtain a 95% confidence interval for $E(y_h)$ when $x_h = 10$.

   (c) Repeat Parts (a) and (b) 500 times, generating new random numbers each time. Calculate the mean and standard deviation of the 500 estimates, $\hat{\beta}_1$. Construct a frequency distribution of the 500 estimates, $\hat{\beta}_1$. Are the results consistent with theoretical expectations? Explain.

   (d) What proportion of the 500 confidence intervals for $E(y_h)$ when $x_h = 10$ include $E(y_h)$? Is the result consistent with theoretical expectations? Explain.

4. In a study on the efficacy of hospital-acquired infection control, the researchers sought to determine whether infection surveilliance and control programs have reduced the rates of hospital-acquired infection in hospitals. Data on 12 variables collected from 113 hospitals surveyed are attached. The average length of stay (Y) is anticipated to be related to infection risk, available facilities and services, and routine chest X-ray ratio. Assume that a simple linear regression model is appropriate for each of the three predictor variables.

   (a) For each geographic region, fit average length of stay to infection risk. State the estimated regression function. Overlay the estimated regression function for each region on a scatter plot of the data. On the graph, use different characters to represent points from different regions.

   (b) For each region, estimate the variation in average length of stay that was not accounted for by the fitted regression model. Based on your results from Parts (a) and (b) comment on the similarities and/or differences in the effect of infection risk on average length of stay between the regions.

   (c) For each of the fitted regression models obtain the residuals and prepare a residual plot against x. What does the residual plot show?

(d) Omit cases 47 and 112 from the data and obtain the estimated regression function between average length of stay and infection risk based on the remaining cases. Compare this estimated regression function to that obtained in Part (a). What can you conclude about the effect of the cases (47 and 112) that were omitted?

(e) Using your fitted models in Part (d), obtain separate 95% prediction intervals for new y observations at $x = 6.5$ and $x = 5.9$, respectively. Do observations y_{47} and y_{112} fall outside these prediction intervals? Discuss the significance of this.

In the data set for Problem 4,

column 2 is Average length of stay
column 3 is Age
column 4 is Infection risk
column 5 is Routine culturing ratio
column 6 is Routine chest X-ray ratio
column 7 is Number of beds
column 8 is Medical school affiliation
column 9 is Region
column 10 is Average daily census
column 11 is Number os nurses
column 12 is available facilities and services.