QSCI 381 HW 8

Simon-Hans Edasi

6/08/2023

1. (1 point) What is the objective of simple linear regression?

   The objective is to establish linear relationships between an independent, or predictor, and dependant variable, or response. Ultimately, we want to derive an equation that predicts the dependant variable from the independent variable.

2. (1 point) In simple linear regression, we seek to estimate two parameters the equation of a straight-line function: $Y = a + bx$, where a is the intercept and b is the slope. We want to find the estimates for those parameters that provide the "best fit" to our data. What is meant by "best fit"?

   The best fit is the line that minimizes the difference between the observed datapoints and predictions from the linear equation.

3. (3 points) In estimating these parameters in this manner, we apply the calculus to minimize the residual sum of squares. In so doing we get a pair of equations. What are they called? What are the solutions to these two equations?

   The equations are the normal equations and the solutions are:

$$\sum_i y_i = n\beta_0 + \beta_1 \sum_i x_i$$

$$\sum_i x_i y_i = \beta_0 \sum_i x_i + \beta_1 \sum_i x_i^2$$

4. (6 points) What are the statistical assumptions explicitly assumed in simple linear regression?

   (a) **The relationship between independent and dependant variable is linear. The data points will follow a pattern of a straight line when plotted.**
   (b) **Observations used in the regression are independent of each other.**
   (c) **Residuals are normally distributed**
   (d) **Variance of residuals is constant across all independent variables (homoscedacity)**

5. (1 point) With regard to using the regression model what we develop (or fit), what warning must always be remembered?

   Correlation does not imply causation

6. (2 points) The usefulness of an estimated regression equation is based on two important measures. What are these two measures and what do they measure?

   (a) **R^2, coefficient of determination measures the proportion of variation explained by the linear model.**
   (b) **SE, standard error of the estimate measures the mean residual, or difference of prediction and observation**

7. (2 points) It can be shown that the equation for R2 (the coefficient of determination) is identical to that for r 2 (the square of the correlation coefficient). Why do these two measures not measure that same thing?

   The coefficient of determination is a measure of residuals, whereas the correlation coefficient is a measure of how close the data are to a line of best fit. Squaring the correlation coefficient does not measure the same thing as the coefficient of determination because magnitude is not considered in the correlation coefficient.

8. (14 points) A physiologist once proposed that the oxygen consumed by an animal should be proportional to its surface area, and since surface area is proportional to the weight raised to the 2/3 power, then oxygen consumption should be proportional to the 2/3 power of the animal's weight. A random sample of 10 animals from certain specie was collected. The physiologist's theory results in a power function. To test this hypothesis, each animal's oxygen consumption was measured and its weight recorded.

   (a) (4 points) Using the symbols in the problem, clearly and specifically state what the linear model coefficients are estimated to be. Using the values of these coefficients, write the estimated regression equation.

   $$Y_i = \alpha + \beta x_i = 4.405 + 0.783 x_i$$

   Written from the software output, the regression equation states that the natural log of the oxygen consumption of the animal (Y_i) equals a constant ($\alpha = 4.405$) plus a regression coefficient ($\beta = 0.783$) times the natural log of the weight of the animal (x_i).

   (b) (4 points) Do the assumptions of linear regression appear to hold? In your answer, state exactly what output from the computer software you would look at to answer that question, what you would check for in looking at it, and what you conclude after looking at it.

   The boxplot looks normal, maybe with a sligh skew. The scatter plot shows a linear relationship within the data. QQ and and residual plot show that the residuals are normally distributed and no patterns of variance. I think the assumptions have been met.

   (c) (2 points) What percentage of the variability in our data can be explained by the regression model? What do we call this measure? R2 = 0.968 .

   This regression model can explain approximately 96.8% of the variability in the dependant variable. This is called the coefficient of determination.

   (d) (2 points) What is the critical value that is appropriate for the physiologist to use to test the hypothesis (that is, using the appropriate table, give the numerical value of the critical value)? Using the value of the test statistic given in the problem, state if the physiologist should reject or fail to reject the null hypothesis that the slope is approximately 0.66.

   The critical value is 2.306. The given test statistic is 2.46, which is higher than our critical value. The physiologist should reject the null hypothesis that the slope is approximately 0.66.

   (e) (2 points) State your conclusion in (d), above, in the context of the problem.

   The conclusion in d. was to reject the null hypothesis that the slope of the regression coefficient was 0.66. This hypothesis was to test the idea that oxygen consumption was proportional to surface area, and 2/3 weight was taken as an easily measureable proxy for surface area. Because we rejected the null hypothesis of the slope = 0.66, we must then reject the hypothesis that oxygen consumption is proportional to surface area.