STAT 523 HW 10

Md Muhtasim Billah

4/7/2020

Question 1.

(HW9 question 3 continued) In biofiltration of wastewater, air discharged from a treatment facility is passed through a damp porous membrane that causes contaminants to dissolve in water and be transformed into harmless products. The accompanying data on x= inlet temperature (°C) and y= removal efficiency (%) was the basis for a scatter plot that appeared in the article "Treatment of Mixed Hydrogen Sulfide and Organic Vapors in a Rock Medium Biofilter"(Water Environment Research, 2001: 426–435). The scatter plot and the summary statistics are given.

a. Calculate the fitted regression equation.
b. Interpret the intercept in a.
c. Interpret the slope in a.
d. Obtain a prediction of removal efficiency when temperature=12.

Answer:

a.

The simple linear regression model,

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon$$

Here, β_0 is the intercept, β_1 is the slope and ε is the random noise of the system. The fitted regression equation will be of the form,

$$\hat{y_i} = \hat{\beta_0} + \hat{\beta_1} x_i$$

Here, the hat sign indicates the estimated values.

Now, given that,

$n = 33$
$\sum(xi) = 387$
$\sum(yi) = 3365$
$\sum(x_i - \bar{x})^2 = 514$
$\sum(y_i - \bar{y})^2 = 6.847$
$\sum((x_i - \bar{x})(y_i - \bar{y})) = 46.578$

The intercept and slope can be estimated using these values with the formulae below.

$$\hat{\beta_1} = \frac{S_{xy}}{S_{xx}}$$

$$\hat{\beta_0} = \bar{y} - \hat{\beta_1}\bar{x}$$

Here,

$$S_{xx} = \sum(x_i - \bar{x})^2 = \sum {x_i}^2 - \frac{(\sum x_i)^2}{n} = 514$$

$$S_{xy} = \sum((x_i - \bar{x})(y_i - \bar{y})) = \sum x_i y_i - \frac{\sum x_i \sum y_i}{n} = 46.578$$

$$\bar{y} = \frac{\sum y_i}{n} = \frac{3365}{33} = 101.969697$$

$$\bar{x} = \frac{\sum x_i}{n} = \frac{387}{33} = 11.727273$$

Plugging the values, we get,

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{46.578}{514} = 0.09062$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 101.969697 - 0.09062 \times 11.727273 = 100.90697$$

Thus, the fitted regression equation is,

$$\hat{y}_i = 100.90697 + 0.09062\ x_i$$

b.

The intercept estimate $\hat{\beta}_0$ indicates the value of the predicted/fitted value of the response, y (removal efficiency) for the scenario when the predictor x (inlet temperature) is 0. Here, although the 0 value for the predcitor has a significant meaning, it falls outside the range of the dataset and thus, $\hat{\beta}_0$ is estimated via extrapolation which is not ideal. Thus, for this case, the intercept cannot be interpreted.

c.

The slope estimate $\hat{\beta}_1$ indicates that the average change in the predicted/fitted value of the response, y (removal efficiency) for every unit change in the predictor x (inlet temperature) is 0.09062. The posotive sign of the slope indicates that the relation between x and y is positive. In other words, if the the inlet temperature increases by 1°C, then the removal efficiency will increase by 0.09062%.

d.

If the predictor (inlet temperature) is $x = 12°C$, then the reponse y (removal efficiency) can be predicted using the regression equation as below.

$$\hat{y}_i = 100.90697 + 0.09062\ x_i$$

$$\hat{y}_i = 100.90697 + 0.09062 \times 12$$

$$\hat{y}_i = 101.99441$$

Problem 2.

(From Jay L. Devore, Chapter 12, Section 12.5, Exercise 58 parts a b c). The Turbine Oil Oxidation Test (TOST) and the Rotating Bomb Oxidation Test (RBOT) are two different procedures for evaluating the oxidation stability of steam turbine oils. The article "Dependence of Oxidation Stability of Steam Turbine Oil on Base Oil Composition" (J. of the Society of Tribologists and Lubrication Engrs., Oct. 1997: 19–24) reported the accompanying observations on x = TOST time (hr) and y = RBOT time (min) for 12 oil specimens.

a. Calculate and interpret the value of the sample correlation coefficient (as do the article's authors).
b. How would the value of r be affected if we had let x = RBOT time and y = TOST time?
c. How would the value of r be affected if RBOT time were expressed in hours?

Use the given summary statistics.

Hints:

- Do not forget to interpret the correlation in part (a). Qualify the strength of the linear relationship between x and y.
- For parts (b) and (c), just say if will increase, decrease, or remain the same. No need to explain.

Answer:

a.

Given summary statistics for the dataset,

$n = 12$
$\sum(x_i) = 44,615$
$\sum(x_i)^2 = 170,355,425$
$\sum(y_i) = 3,860$
$\sum(y_i)^2 = 1,284,450$
$\sum(x_i y_i) = 14,755,500$

The formula for calculating the sample correlation coefficient,

$$r = \frac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{yy}}}$$

Here,

$$S_{xx} = \sum(x_i - \bar{x})^2 = \sum {x_i}^2 - \frac{(\sum x_i)^2}{n} = 170,355,425 - \frac{(44,615)^2}{12} = 4,480,572.917$$

$$S_{yy} = \sum(y_i - \bar{y})^2 = \sum {y_i}^2 - \frac{(\sum y_i)^2}{n} = 1,284,450 - \frac{(3,860)^2}{12} = 42,816.66667$$

$$S_{xy} = \sum((x_i - \bar{x})(y_i - \bar{y})) = \sum x_i y_i - \frac{\sum x_i \sum y_i}{n} = 14,755,500 - \frac{44,615 \times 3,860}{12} = 404,341.6667$$

Plugging this values, we get the value for the sample correlation coefficient,

$$r = \frac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{yy}}} = \frac{404,341.6667}{\sqrt{4,480,572.917} \times \sqrt{42,816.66667}} = 0.92316$$

Interpretation:

Since the the sample correlation coefficient, r is very close to 1, it suggests that there is a strong linear relationship between the predictor and the response variable.

b.

If we had x = RBOT time and y = TOST time, the value of r would not be affected and remain the exact same.

This means that despite swapping the predictor and response variable, there will still be strong linear correlation between them. It can be easily noticed from the above equations as well.

c.

If RBOT time were expressed in hours, the value of r would not be affected and remain the exact same.

This is because the standard deviation of the variable y (RBOT time) will be the same since it wouldn't depend on the unit. This attribute in term will keep the r unaffected.