hw1

1. The least squares estimator of the slope coefficient in a simple linear regression model is

$$\hat{\beta}_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$$

and the sample correlation coefficient has the formula

$$r_{XY} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Starting with the above expression for the slope coefficient, show that it can be expressed as

$$\hat{\beta}_1 = r_{XY}\frac{s_y}{s_x}$$

where s_y is the standard deviation of Y and s_x is the standard deviation of X.

Answer:

$$S_x = \sqrt{\frac{1}{n-1}\sum(x_i - \bar{x})^2}$$

$$S_y = \sqrt{\frac{1}{n-1}\sum(y_i - \bar{y})^2}$$

$$r_{xy}\frac{S_y}{S_x} = \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum(x_i-\bar{x})^2}\sqrt{\sum(y_i-\bar{y})^2}} \times \frac{\sqrt{\frac{1}{n-1}\sum(y_i-\bar{y})^2}}{\sqrt{\frac{1}{n-1}\sum(x_i-\bar{x})^2}}$$

$$r_{xy}\frac{S_y}{S_x} = \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sum(x_i-\bar{x})^2}$$

$$r_{xy}\frac{S_y}{S_x} = \hat{\beta}_1$$

2. Data were collected from 120 young adult patients. Let x be age (in years) and Y be satisfaction with health care provider (scale of 0-4, higher score indicates more satisfaction). Suppose you have the following summary statistics:

$\bar{x} = 24.7$
$\bar{Y} = 3.07$
$\sum_{i=1}^{n}(x_i - \bar{x})^2 = 2379.93$
$\sum_{i=1}^{n}(x_i - \bar{x})(Y_i - \bar{Y}) = 92.41$
$\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = 49.405$
$\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = 45.82$

a. Obtain the sample correlation coefficient, r_{XY}.
b. Obtain the least squares estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ for the simple linear regression model $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, ε_i independent $N(0, \sigma^2)$.
c. Provide an interpretation of $\hat{\beta}_0$ and $\hat{\beta}_1$, including their units.
d. Suppose that you were working with the centered-predictor model, which we will write as $Y_i = \alpha_0 + \alpha_1(x_i - \bar{x}) + \varepsilon_i$, ε_i independent $N(0, \sigma^2)$. What are the least squares estimators of α_0 and α_1?
e. Provide an interpretation of $\hat{\alpha}_0$ and $\hat{\alpha}_1$, including their units.

(a) Answer:

$$r_{xy} = \frac{\sum(x_i - \bar{x})(Y_i - \bar{Y})}{\sqrt{\sum(x_i - \bar{x})^2}\sqrt{\sum(Y_i - \bar{Y})^2}}$$

$$r_{xy} = \frac{92.41}{\sqrt{2379.93} \times \sqrt{49.905}}$$

$$r_{xy} \approx 0.268$$

(b) Answer:

$$Q(\beta_0, \beta_1) = \sum(Y_i - \beta_0 - \beta_1 x_i)^2$$

$$\frac{\partial Q}{\partial \beta_1} = -2\sum(Y_i - \beta_0 - \beta_1 x_i)x_i$$

$$\frac{\partial Q}{\partial \beta_0} = -2\sum(Y_i - \beta_0 - \beta_1 x_i)$$

Let $\frac{\partial Q}{\partial \beta_1} = 0$ and $\frac{\partial Q}{\partial \beta_0} = 0$

We derive:

$$\hat{\beta}_1 = \frac{\sum(x_i - \bar{x})(Y_i - \bar{Y})}{\sum(x_i - \bar{x})^2}$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x}$$

Hence:

$$\hat{\beta}_1 = \frac{92.41}{2379.93} \approx 0.0388$$

$$\hat{\beta}_0 = 3.07 - 0.0388 * 24.7 \approx 2.11$$

(c) Answer:

$\hat{\beta_1}$: We would expect an increase of 0.0388 scores for satisfaction with healthcare provider for each additional year of age in patient. (Unit:Score/Year)

$\hat{\beta_0}$: We would expect a satisfaction with healthcare provider score of 2.11 for a 0 year of age in patient. (Unit:Score)

(d) Answer:

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

$\hat{Y_i} = \hat{\beta_0} + \hat{\beta_1}(x_i + \bar{x} - \bar{x})$

$\hat{Y_i} = (\hat{\beta_0} + \hat{\beta_1}\bar{x}) + \hat{\beta_1}(x_i - \bar{x})$

$\hat{Y_i} = \hat{\alpha_0} + \hat{\alpha_1}(x_i - \bar{x})$

Then:

$\hat{\alpha_1} = \hat{\beta_1}$

$\hat{\alpha_0} = \hat{\beta_0} + \hat{\beta_1}\bar{x}$

Hence:

$\hat{\alpha_1} = 0.0388$

$\hat{\alpha_0} = 2.11 + 0.0388 * 24.7 \approx 3.07$

(e) Answer:

$\hat{\alpha_1}$: We would expect an increase of 0.0388 scores for satisfaction with healthcare provider for each additional year of age in patient. (Unit: Score/Year)

$\hat{\alpha_0}$: We would expect a satisfaction with healthcare provider score of 3.07 for a 24.7 year of age in patient.(Unit: Score)

3. Consider the model $Y_i = \beta_0 + \varepsilon_i$. Using the method of least squares, show that the value of β_0 that minimizes the sum of the squared residuals is $\hat{\beta}_0 = \bar{Y}$.

Answer:

$Q(\beta_0) = \sum (Y_i - \beta_0)^2$

$\frac{\partial Q}{\partial \beta_0} = -2 \sum (Y_i - \beta_0)$

Let $\frac{\partial Q}{\partial \beta_0} = 0$

We derive:

$\hat{\beta_0} = \bar{Y}$

Since Q is a convex function, $\hat{\beta_0}$ is the global minimum of Q.

4. Data analysis exercise: The SENIC data set consists of a sample of 113 United States hospitals. The data were collected as part of the SENIC (Study on the Efficacy of Nosocomial Infection Control) project, in 1975-1976. The variables in the data set are:

risk	infection risk; average estimated probability of acquiring infection in hospital × 100
length	average length of stay of patients in hospital (in days)
age	average age of patients (in years)
culture	routine culturing ratio; ratio of number of cultures performed to number of patients without signs or symptoms of hospital-acquired infection × 100
xray	routine chest x-ray ratio; ratio of number of chest x-rays performed to number of patients without signs or symptoms of pneumonia × 100
beds	average number of beds in hospital during study period
census	average number of patients in hospital per day during study period
region	geographic region (1=North East, 2=North Central, 3=South, 4=West)
nurses	average number of full-time equivalent nurses during study period
msch	medical school affiliation; 1=yes, 2=no
svcs	available facilities and services; percent of 35 potential facilities and services that are provided by the hospital

Fit each of the following simple linear regression models and provide an interpretation of the regression coefficient associated with the x variable.

a. Y = risk, x = beds
b. Y = risk, x = svcs
c. Y = nurses, x = age
d. For medical school affiliation, create a variable called med that equals 1 for yes and 0 for no and fit the model with Y = nurse and x = med.

(a) Answer:

$Y = 3.724044 + 0.0025x$

For every one number increase of average number of beds in hospital during study period, we expect 0.0025 increase of average estimated probability of acquiring infection in hospital.

(b) Answer:

$Y = 2.78402 + 0.03640x$

For every percent increase of available facilities and services of 35 potential facility and services, we expect 0.03640 increase of average estimated probability of acquiring infection in hospital.

(c) Answer:

$Y = 311.06760 - 2.58905x$

For every year increase of average age of patients, we expect 2.58905 number of decrease of average number of full-time equivalent nursses during stay period.

(d) Answer:

$Y = 138.92708 + 228.13174x$

On average, we expect 228.13174 more average number of full-time equivalent nurses during study period nurses in medical affiliation than in non-medical affiliation.