Tutorial 8

August 27, 2019

1

$$
\begin{aligned}
r_s &= 1 - \frac{6\,(28)}{8\,(8^2 - 1)} \\
&= \frac{2}{3} \\
&= 0.6667
\end{aligned}
$$

There is a high positive linear correlation between the value to the company and preference of the managers.

2

$$
\begin{aligned}
r_s &= 1 - \frac{6\,(48)}{10\,(10^2 - 1)} \\
&= \frac{39}{55} \\
&= 0.7091
\end{aligned}
$$

There is a high positive linear correlation that it represents a measure of agreement between the two panel members.

3

1. From the scatter diagram, there is a moderate negative linear relationship between the weeks of experience and the number of rejects.

2. Product moment correlation coefficient

$$\begin{aligned} r &= \frac{n\sum XY - \left(\sum X\right)\left(\sum Y\right)}{\sqrt{\left[n\sum X^2 - \left(\sum X\right)^2\right]\left[n\sum Y^2 - \left(\sum Y\right)^2\right]}} \\ &= \frac{8\,(1069) - (72)\,(128)}{\sqrt{\left[8\,(732) - (72)^2\right]\left[8\,(2156) - (128)^2\right]}} \\ &= -0.8714 \end{aligned}$$

   (a) There is a very high negative linear correlation between the weeks of experience and the number of rejects.

3. Spearman's rank correlation

$$\begin{aligned} r_s &= 1 - \frac{6\,(160.5)}{8\,(8^2 - 1)} \\ &= -0.9107 \end{aligned}$$

   (a) There is a very high negative linear correlation between the weeks of experience and the number of rejects

4

1. Draw and comment the scatter diagram.

   (a) Interpretation: There is a high positive linear correlation between the output and the cost

2. Product moment coefficient correlation

$$\begin{aligned} r &= \frac{n\sum XY - \left(\sum X\right)\left(\sum Y\right)}{\sqrt{\left[n\sum X^2 - \left(\sum X\right)^2\right]\left[n\sum Y^2 - \left(\sum Y\right)^2\right]}} \\ &= \frac{10\,(5704) - (120)\,(400)}{\sqrt{\left((10)\,(1866) - (120)^2\right)\left((10)\,(18200) - (400)^2\right)}} \\ &= 0.9338 \end{aligned}$$

   (a) Interpretation:

      i. Interpretation: There is a very high positive linear correlation between the output and the cost

3. Spearman's rank coefficient correlation

$$\begin{aligned} r_s &= 1 - \frac{6\,(8.5)}{10\,(10^2 - 1)} \\ &= 0.9485 \end{aligned}$$

(a) Interpretation: There is a very high positive linear correlation between the output and the cost

5

1. Product moment correlation coefficient

$$
\begin{aligned}
r &= \frac{n\sum XY - \left(\sum X\right)\left(\sum Y\right)}{\sqrt{\left[n\sum X^2 - \left(\sum X\right)^2\right]\left[n\sum Y^2 - \left(\sum Y\right)^2\right]}} \\
&= \frac{12\,(7449) - (659)\,(135)}{\sqrt{\left[(12)\,(36313) - (659)^2\right]\left[(12)\,(1561) - (135)^2\right]}} \\
&= 0.4891
\end{aligned}
$$

(a) Interpretation: There is some positive linear correlation between the median income and the house purchase price

2. Spearman's rank correlation coefficient

$$
\begin{aligned}
r_s &= 1 - \frac{6\,(192.5)}{12\,(12^2 - 1)} \\
&= 0.3269
\end{aligned}
$$

6 Answer

1. Least squares regression line of time on year.

$$
\hat{Y} = \hat{a} + \hat{b}X
$$

$$
\begin{aligned}
\hat{b} &= \frac{n\sum XY - \left(\sum X\right)\left(\sum Y\right)}{n\sum X^2 - \left(\sum X\right)^2} \\
&= \frac{(9)\,(7838.4) - (144)\,(496.6)}{(9)\,(3264) - (144)^2} \\
&= -0.112
\end{aligned}
$$

$$
\begin{aligned}
\hat{a} &= \frac{(496.6)}{144} - (-0.112)\,\frac{144}{9} \\
&= 5.241
\end{aligned}
$$

$$
\hat{Y} = 5.241 - 0.112X
$$

2. Predict winning time for year 2012.

$$
\begin{aligned}
\hat{Y} &= -0.1117 * 56.964 * 40 \\
&= 52.50 seconds
\end{aligned}
$$

7 Answer

1. Least square regressions line

$$\hat{Y} = \hat{a} + \hat{b}X$$

$$\begin{aligned}\hat{b} &= \frac{n\sum XY - \left(\sum X\right)\left(\sum Y\right)}{n\sum X^2 - \left(\sum X\right)^2} \\ &= \frac{(7)(2273) - (21)(641)}{(7)(91) - (21)^2} \\ &= 12.5\end{aligned}$$

$$\begin{aligned}\hat{a} &= \frac{\sum Y}{n} - \hat{b}\frac{\sum X}{n} \\ &= \frac{641}{7} - 12.5\frac{21}{7} \\ &= 54.0714\end{aligned}$$

$$\hat{Y} = 54.0714 + 12.5X$$

2. Predict average annual earnings for year 2011 and 2012.

   (a) 2011

      i. $2011 - 2004 = 7$

      ii.

$$\begin{aligned}\hat{Y} &= 54.0714 + 12.5(7) \\ &= 141.5714\end{aligned}$$

   (b) 2012

      i. $2012 - 2004 = 8$

      ii.

$$\begin{aligned}\hat{Y} &= 54.0714 + 12.5(8) \\ &= 154.0714\end{aligned}$$

8 Answer

A random sample of eight drivers insured with a company and having similar car insurance policies was selected. The following table lists their driving experiences and monthly car insurance premiums.

(a) Does the insurance premium depend on the driving experience or does the driving experience depend on the insurance premium? State the dependent and independent variables.

The insurance premium depends on the driving experience.

In statistics: Y depends on X.

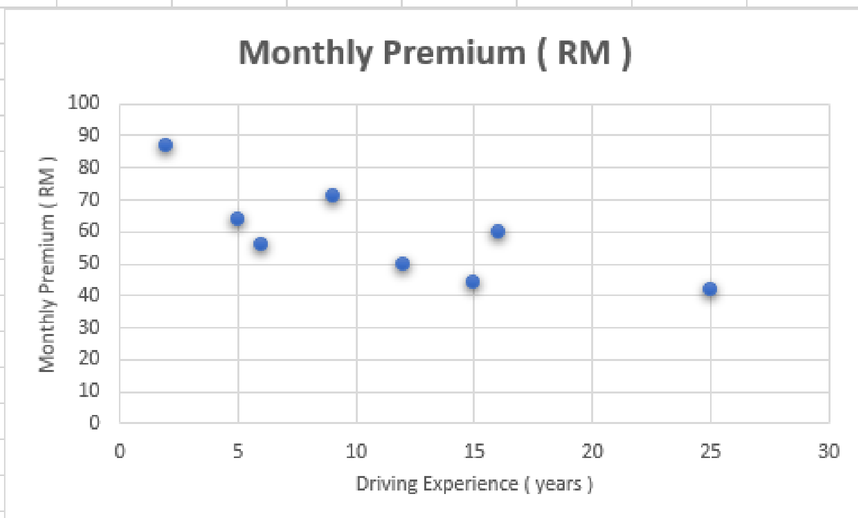
Y: Insurance premium (dependent variable)

X: Driving experience (Independent variable)

Driving Experience	X		90
Insurance Premium	Y		474
	XY		4739
	X^2		1396
	Y^2		29642

(b) Draw and comment the scatter diagram.

Note: Can start from lowest to highest, no need to start from zero



From the scatter diagram, there is a moderate negative linear relationship between driving experience and monthly premium. (Note: The relationship must be 3 parts, intensity + positive/negative + linear, which one come first doesn't matter)

(c) Find the appropriate least squares regression line based on your answer in part (a).

$$S_{XY} = \sum xy - \frac{\sum x \sum y}{n} = 4739 - \frac{90\,(474)}{8}$$
$$=$$

$$S_{XX} = \sum X^2 - \frac{\left(\sum X\right)^2}{n} = 1396 - \frac{90^2}{8}$$
$$= 383.5$$

$$S_{XY} = \sum y^2 - \frac{\left(\sum y\right)^2}{n}$$
$$= 29642 - \frac{474^2}{8}$$
$$= 1557.5$$

$$\hat{b}\left(gradient, \frac{y_a - y_b}{x_a - x_b}\right) = \frac{S_{X\boldsymbol{Y}}}{S_{X\boldsymbol{X}}}$$
$$= -\frac{593.5}{383.5}$$
$$= -1.5476$$

$$\hat{a} = \bar{y} - \hat{b}X$$
$$= \frac{474}{8} + 1.5476\left(\frac{90}{8}\right)$$
$$= 76.6605$$

$$\hat{y} = 76.6605 - 1.5476x (\text{When X=0, } \hat{y} = \hat{a})$$

(d) Interpret the meaning of the values of regression coefficients a and b obtained in part (c).

For a driver with no experience, the predicted monthly premium is RM76.66.

(General format: When X increases by 1 unit, Y is expected to change by $\hat{b}$ units, on the average)

When the driving experience increases by 1 year, the monthly premium is expected to decrease by RM1.55 on average)

(e) Calculate and interpret the correlation coefficient between the two variables.

$$r = \frac{S_{XY}}{\sqrt{S_{XX}}\sqrt{S_{YY}}}$$
$$= \frac{-593.5}{\sqrt{383.5}\sqrt{1557.5}}$$
$$= -0.7679$$

(Note: Look at notes page 81 table, notes, and decide it is from where to where)

There is a high negative linear correlation(proper)/relationship between driving experience and monthly premium

(f) Calculate and interpret the coefficient of determination.

$$\begin{aligned} r^2 &= (-0.7659) \\ &= 0.5879 \end{aligned}$$

58.97% of the total variance in monthly premium is explained by the driving experience (Generally, the higher the better).

(g) Predict the monthly car insurance premium for a driver with 10 years of driving experience.

When X = 10,

$$\begin{aligned} \hat{y} &= 76.6605 - 1.5476\,(10) \\ &= 61.1845 \end{aligned}$$

The predicted monthly premium monthly premium is RM61.18.

(h) Predict the monthly car insurance premium for a driver with 30 years of driving experience.

When X = 30,

$$\begin{aligned} \hat{y} &= 76.6605 - 1.5476\,(30) \\ &= 30.2325 \end{aligned}$$

The predicted monthly premium monthly premium is RM30.2325.

(i) Comment on the accuracy of your estimates obtained in parts (g) and (h).

The accuracy of estimates in part g is more accurate than h. This is because part g is interpolation and part h is extrapolation. g is inside the range of the samples, while h is outside the range of samples, which is less accurate.

9 Answer (Check with teacher)

1. State the dependent and independent variables.
   (a) Dependent: Cholesterol level
   (b) Independent: Age
2. Find the least squares regression line of cholesterol level on age.

$$\hat{Y} = \hat{a} + \hat{b}X$$

$$\begin{aligned} \hat{b} &= \frac{n\sum XY - \left(\sum X\right)\left(\sum Y\right)}{n\sum X^2 - \left(\sum X\right)^2} \\ &= \frac{(10)\,(39531) - (202)\,(1896)}{(10)\,(5976) - (202)^2} \\ &= 0.6498 \end{aligned}$$

$$
\begin{aligned}
\hat{a} &= \frac{\sum Y}{n} - \hat{b}\frac{\sum X}{n} \\
&= \frac{1896}{10} - 0.6498\frac{202}{10} \\
&= 176.474
\end{aligned}
$$

$$
\hat{Y} = 176.474 + 0.6498X
$$

(Note: the answer for 9b is wrong, as you cannot get 9e answer using 9b's equation given in the answer)

3. Interpret the meaning of the values of regression coefficients a and b obtained in part (b).

   (a) For a men with an age of 58, the cholesterol level is 176.47

   (b) As the age of the men increase by 1, the cholesterol level increases by 0.6498.

4. Calculate and interpret the correlation coefficient between the two variables.

$$
\begin{aligned}
r &= \frac{S_{XY}}{\sqrt{S_{X^2}}\sqrt{S_{Y^2}}} \\
&= \frac{(395310 - 382992)}{\sqrt{[59760 - 40804]\,[3642800 - 3594816]}} \\
&= 0.4084
\end{aligned}
$$

   (a) Interpretation

      i. There is a some positive linear correlation between the age and the cholesterol level of men.

5. Calculate and interpret the coefficient of determination.

$$
\begin{aligned}
r^2 &= (0.8473)^2 \\
&= 0.7179
\end{aligned}
$$

   (a) Interpretation

      i. 71.79% of the total variance is explained by the age.

6. Predict the cholesterol level of a 60-year-old man.

   (a) When $X = 60 - 31 = 29$

$$
\begin{aligned}
\hat{Y} &= 176.474 + 0.6498\,(29) \\
&= 195.3182
\end{aligned}
$$