PES University, Bangalore

(Established under Karnataka Act No. 16 of 2013)

UE19CS203 – STATISTICS FOR DATA SCIENCE

Unit-5 - Power of Test and Simple Linear Regression

QUESTION BANK - SOLVED

Simple Linear Regression: Introduction to correlation

Exercises for section 7.1: [Text Book Exercise 7.1– Pg. No. [520 – 523]]

1. In a study of ground motion caused by earthquakes, the peak velocity (in m/s) and peak acceleration (in m/s^2) were recorded for five earthquakes. The results are presented in the following table.

Velocity	1.54	1.60	0.95	1.30	2.92
Acceleration	7.64	8.04	8.04	6.37	5.00

   a. Compute the correlation coefficient between peak velocity and peak acceleration.
   b. Construct a scatterplot for these data.
   c. Is the correlation coefficient an appropriate summary for these data? Explain why or why not.
   d. Someone suggests converting the units from meters to centimeters and from seconds to minutes. What effect would this have on the correlation?

[Text Book Exercise – Section 7.1 – Q. No.6 – Pg. No. 521]

Solution:

(a). Correlation coefficient :

$$r = \frac{\sum xy - n\bar{x}\bar{y}}{\sqrt{x_i^2 - n\bar{x}^2}\sqrt{y_i^2 - n\bar{y}^2}}$$

$$r = \frac{\sum xy - \frac{(\sum x)(\sum y)}{n}}{\sqrt{x_i^2 - \frac{(\sum x)^2}{n}}\sqrt{y_i^2 - \frac{(\sum y)^2}{n}}}$$

X	Y	XY	X^2	Y^2
1.54	7.64	11.7656	2.3716	58.3696
1.6	8.04	12.864	2.56	64.6416
0.95	8.04	7.638	0.9025	64.6416
1.3	6.37	8.281	1.69	40.5769
2.92	5	14.6	8.5264	25
8.31	35.09	55.1486	16.0505	253.2297

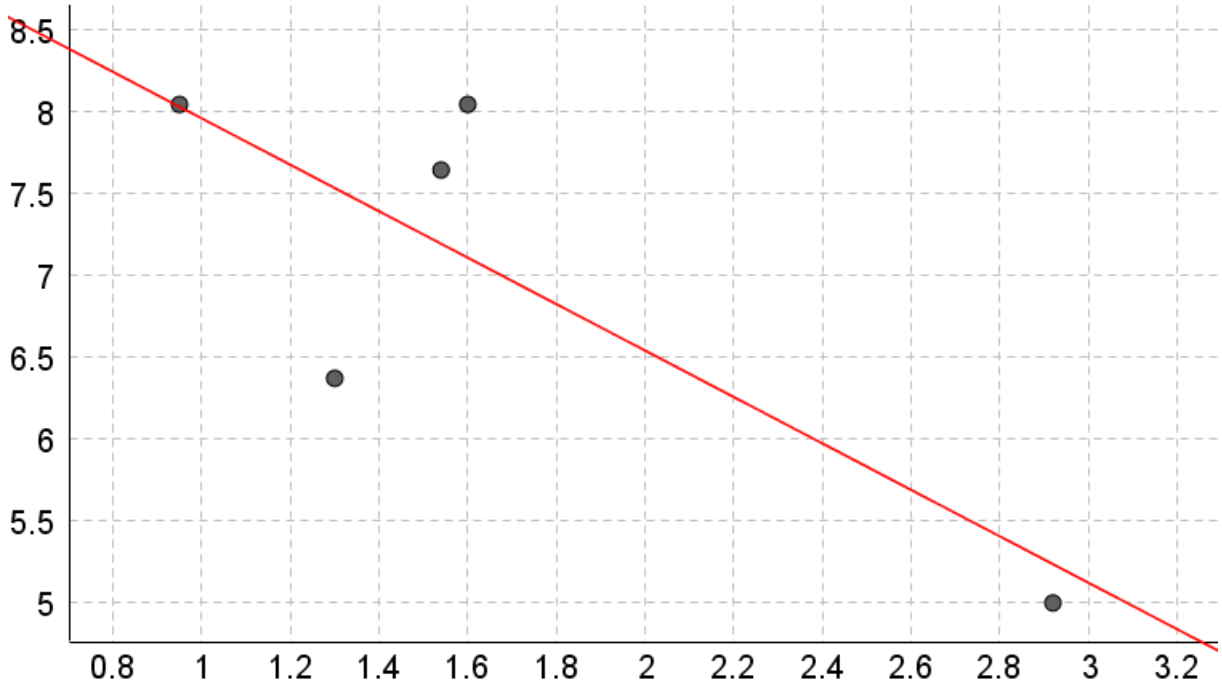
From the table we observe that

$$\sum x = 8.31$$
$$\sum y = 35.09$$
$$\sum xy = 55.1486$$
$$\sum x^2 = 16.0505$$
$$\sum y^2 = 253.2297$$

Therefore

$$r = \frac{55.1486 - \frac{(8.31)(35.09)}{5}}{\sqrt{16.0505 - \frac{(8.31)^2}{5}}\sqrt{253.2297 - \frac{(35.09)^2}{5}}} \approx -0.8028$$

(b) Velocity is on the horizontal axis and Acceleration is on the vertical axis.



(c) Based on the scatterplot in part(b), the correlation coefficient is an appropriate summary for these data, because the scatterplot does not contain strong curvature.

(d) The correlation coefficient will remain unchanged, because the correlation coefficient is independent of the units of the variables.

2. Phonics is an instructional method in which children are taught to connect sounds with letters or groups of letters. The article "Predictive Accuracy of Nonsense Word Fluency for English Language Learners" (M. Vanderwood, D. Linklater, and K. Healy, School Psychology Review 2008:5–17) reports that in a sample of 134 English-learning students, the correlation between the score on a phonics test given in first grade and a reading comprehension given in third grade was $r = 0.25$. Can you conclude that there is a positive correlation between phonics test score and the reading comprehension score?

[Text Book Exercise – Section 7.1 – Q. No.12 – Pg. No. 522]

Solution:

Given n = 134, r = 0.25

Let us assume α = 0.05

$$H_0: \rho \leq 0 \ (\textit{There is no positive linear correlation})$$
$$H_1: \rho \leq 0 \ (\textit{There is a positive linear correlation})$$

The value of the test statistic is

$$U = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0.25\sqrt{134-2}}{\sqrt{1-0.25^2}} \approx 2.966$$

The P-value is the probability of obtaining the value of the test statistic, or a value more extreme, assuming that the null hypothesis is true.

Determine the corresponding probability(P-value) using Students T table with df= n-2 = 134 -2 = 132 > 120.

$$0.001 < P < 0.005$$

If the P-value is smaller than the significance level, reject the null hypothesis

$$P < 0.05 \Rightarrow \text{Reject } H_0$$

There is sufficient evidence to support the claim of a positive linear relationship.
