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

Checking Assumptions and transforming data:

Exercises for section 7.4: [Text Book Exercise 7.4– Pg. No. [576 – 584]]

1. In rock blasting, the peak particle velocity (PPV) depends both on the distance from the blast and on the amount of charge. The article "Prediction of Particle Velocity Caused by Blasting for an Infrastructure Excavation Covering Granite Bedrock" (A. Kahriman, Mineral Resources Engineering, 2001:205–218) suggests predicting PPV (y) from the scaled distance (x), which is equal to the distance divided by the square root of the charge. The results for 15 blasts are presented in the following table.

PPV (mm/s)	Scaled Distance (m/kg$^{0.5}$)
1.4	47.33
15.7	9.6
2.54	15.8
1.14	24.3
0.889	23.0
1.65	12.7
1.4	39.3
26.8	8.0
1.02	29.94
4.57	10.9
6.6	8.63
1.02	28.64
3.94	18.21
1.4	33.0
1.4	34.0

a. Plot PPV versus scaled distance. Does the relationship appear to be linear?
b. Compute the least-squares line for the model $ln\ PPV\ =\ \beta_0 + \beta_1$ ln scaled distance $+ \varepsilon$. Plot the residuals versus fitted values. Does this linear model seem appropriate?
c. Use the least-squares line computed in part (b) to predict the PPV when the scaled distance is 20. Find a 95% prediction interval.

[Text Book Exercise – Section 7.4 – Q. No. 4 – Pg. No. 577]

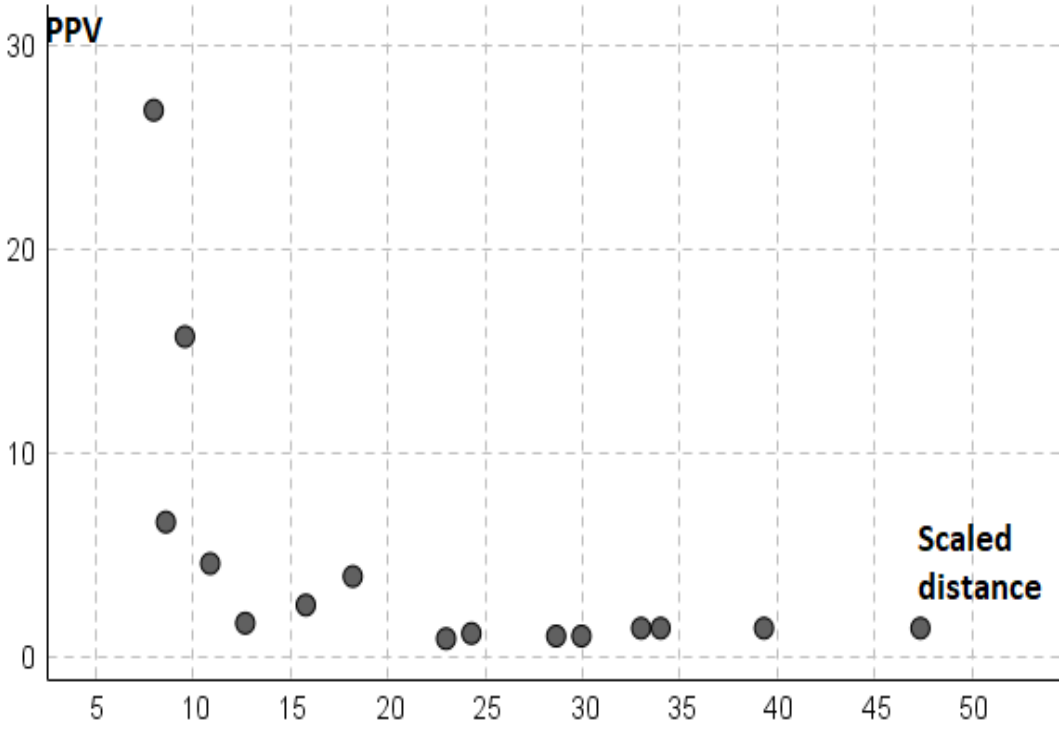
Solution:

Given n = 15

(a) Scatterplot:

Scaled distance is on the horizontal axis and
PPV is on the vertical axis.

There is strong curvature present in the scatterplot and thus the relationship does not appear to be linear.



(b) x represents ln(Scaled Distance) and y represents PPV.

Let us first determine the necessary sums:

$$\sum x_i = 13.6516$$
$$\sum y_i = 44.7200$$
$$\sum x_i y_i = 33.6667$$
$$\sum {x_i}^2 = 27.7507$$

Next, we can determine s_{xx} and s_{xy}

$$s_{xx} = \sum {x_i}^2 - \frac{(\sum x)^2}{n} = 27.7507 - \frac{13.6516^2}{15} = 15.3262$$

$$s_{xy} = \sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n} = 33.667 - \frac{(13.6516)(44.7200)}{15} = -7.0334$$

$$\widehat{\beta_1} = \frac{s_{xy}}{s_{xx}} = \frac{-7.0334}{15.3262} = -0.4589$$

$$\bar{x} = \frac{13.6516}{15} = 0.9101$$

$$\bar{y} = \frac{44.7200}{15} = 2.9813$$

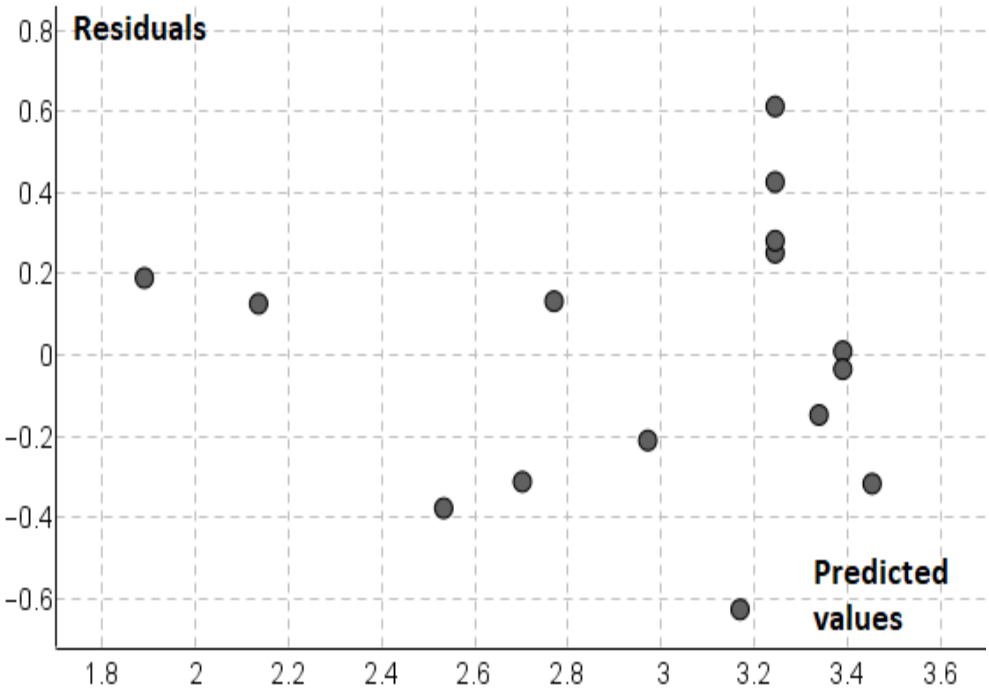
$$\widehat{\beta_0} = \bar{y} - \widehat{\beta_1}\bar{x} = 2.9813 - (-0.4589)(0.9191) = 3.3990$$

General least-square equation: $\hat{y} = \beta_0 + \beta_1 x$. Replace β_0 by $\widehat{\beta_0} = 3.3990$ and β_1 by $\widehat{\beta_1} = -0.4589$ in the general least-squares equation: $\hat{y} = 3.3990 - 0.4589x$

Let us determine the predicted value for each given x-value.
The residual is then the difference between the observed y-value and the predicted y-value.

$\ln x$	$\ln y$	$\ln \hat{y}$	Residual
0.34465	3.8571	3.2446	0.6126
2.7357	2.2618	2.1353	0.1265
0.9322	2.7600	2.9712	-0.2112
0.1310	3.1905	3.3389	-0.1484
-0.1177	3.1355	3.4530	-0.3175
0.5008	2.5416	3.1692	-0.6276
0.3365	3.6712	3.2446	0.4266
3.2284	2.0794	1.8899	0.1895
0.0198	3.3992	3.3899	0.0093
1.5195	2.3888	2.7017	-0.3129
1.8871	2.1552	2.5330	-0.3777
0.0198	3.3548	3.3899	-0.0351
1.3712	2.9020	2.7697	0.1322
0.3365	3.4965	3.2446	0.2519
0.3365	3.5264	3.2446	0.2818

Residual plot
The $\hat{y}$- values are on the horizontal axis and the residuals are on the vertical axis.
There is no strong curvature nor outliers present in the residual plot and thus the model seems appropriate.



(c) $n = 15$

$$\sum (x_i - \bar{x})^2 = 15.3262$$

Total sum of squares = $\sum (y_i - \bar{y})^2 = 4.8004$

$$\bar{x} = 0.9101$$

$$\bar{y} = 2.9813$$

$$\sum (x_i - \bar{x})(\, y_i - \bar{y}) = -7.0334$$

$$c = Confidence\ level = 95\% = 0.95$$

$$x_0 = 20$$

Let us evaluate the equation of the linear model at x=20:

$$\ln \hat{y} = 3.3990 - 0.4589\ ln\ 20 \approx 2.0242$$

Let us first determine the correlation:

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2}\sqrt{\sum(y_i - \bar{y})^2}} = \frac{-7.0334}{\sqrt{15.3262}\sqrt{4.8004}} \approx -0.82$$

Let us next determine the value of s^2:

$$s^2 = \frac{\sum(y_i - \hat{y})^2}{n-2}$$
$$= \frac{(1-r^2)\sum(y_i - \bar{y})^2}{n-2}$$
$$= \frac{(1-(-0.82)^2)(4.8004)}{15-2} \approx 0.1210$$

Determine the critical value using the Students T distribution table in the appendix, which is given in the row df = n-2=15-2=13 and in the column $\alpha/2$ =(1-c)/2 = 0.025

$$t_{\frac{\alpha}{2}} = 2.160$$

Determine the margin of error:

$$E = t_{\frac{\alpha}{2}}\sqrt{s^2\left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum(x_i - \bar{x})^2}\right)}$$
$$= 2.160\sqrt{0.1210\left(1 + \frac{1}{15} + \frac{(\ln 20 - 0.9101)^2}{15.3262}\right)} \approx 0.8732$$

The confidence interval then becomes:

$$1.1510 = 2.0242 - 0.8732 = ln\hat{y} - E < \ln y < ln\hat{y} + E$$
$$= 2.0242 + 0.8732 = 2.8974$$

Finally, we are interested in a prediction interval for y instead of $\ln y$, thus let us take the exponential of each side of the equation:

$$3.1614 = e^{1.1510} < y < e^{2.8974} = 18.1270$$

We are 95% confident that the PPV is between 3.1614 mm/s and 18.1270 mm/s, when the scaled distance is $20\ m/kg^{0.5}$.

2. The article "Mechanistic-Empirical Design of Bituminous Roads: An Indian Perspective" (A. Das and B. Pandey, Journal of Transportation Engineering, 1999:463–471) presents an equation of the form $y = a(1/x_1)^b(1/x_2)^c$ for predicting the number of repetitions for laboratory fatigue failure (y) in terms of the tensile strain at the bottom of the bituminous beam(x_1) and the resilient modulus(x_2). Transform this equation into a linear model, and express the linear model coefficients in terms of a, b, and c.

   [Text Book Exercise – Section 7.4 – Q. No.16 – Pg. No. 583]

   Solution:

   Given

$$y = a\left(\frac{1}{x_1}\right)^b\left(\frac{1}{x_2}\right)^c$$

   Rewrite the model as the product of powers of x_i:

$$y = a(x_1)^{-b}(x_2)^{-c}$$

   Take the natural logarithm on both sides:

$$\ln y = \ln(\, a(x_1)^{-b}(x_2)^{-c})$$

$$\ln y = \ln\ a - b \ln x_1 - c\ \ln x_2$$

Finally, we include a possible error ∈ which can occur in any predicted model:

$$\ln y = \ln\ a - b \ln x_1 - c\ \ln x_2 + \in$$