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

The Least squares line:

Exercises for section 7.2: [Text Book Exercise 7.2 – Pg. No. [536 – 539]]

1. Each month for several months, the average temperature in $°C$ (x) and the number of pounds of steam (y) consumed by a certain chemical plant were measured. The least-squares line computed from the resulting data is $y = 245.82 + 1.13x$.

    a. Predict the number of pounds of steam consumed in a month where the average temperature is $65°C$.
    b. If two months differ in their average temperatures by $5°C$, by how much do you predict the number of pounds of steam consumed to differ?

2. In a study of the relationship between the Brinell hardness (x) and tensile strength in ksi (y) of specimens of cold drawn copper, the least-squares line was $y = -196.32 + 2.42x$.

    a. Predict the tensile strength of a specimen whose Brinell hardness is 102.7.
    b. If two specimens differ in their Brinell hardness by 3, by how much do you predict their tensile strengths to differ?

3. A least-squares line is fit to a set of points. If the total sum of squares is $\sum(y_i - \bar{y})^2 = 9615$, and the error sum of squares is $\sum(y_i - \hat{y})^2 = 1450$, compute the coefficient of determination r^2.

4. A least-squares line is fit to a set of points. If the total sum of squares is $\sum(y_i - \bar{y})^2 = 181.2$, and the error sum of squares is $\sum(y_i - \hat{y})^2 = 33.9$, compute the coefficient of determination r^2.

5. In Galton's height data (Figure 7.1, in Section 7.1), the least-squares line for predicting forearm length (y) from height(x) $is\ y = -0.2967 + 0.2738x$.

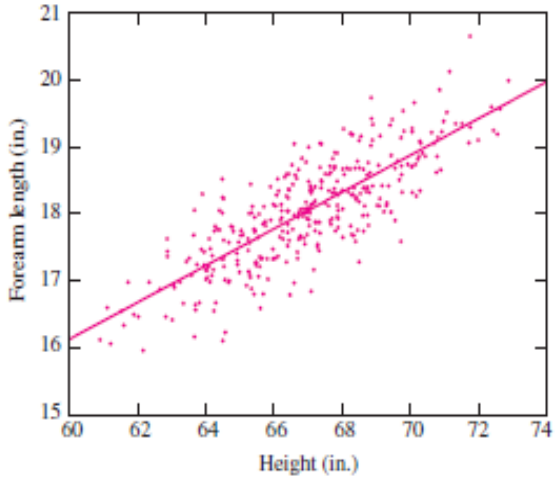

FIGURE 7.1 Heights and forearm lengths of 348 men.

   a. Predict the forearm length of a man whose height is 70 in.
   b. How tall must a man be so that we would predict his forearm length to be 19 in.?
   c. All the men in a certain group have heights greater than the height computed in part (b). Can you conclude that all their forearms will be at least 19 in. long? Explain.

6. In a study relating the degree of warping, in mm, of a copper plate (y) to temperature in $°C\ (x)$, the following summary statistics were calculated: $n = 40, \sum_{i=1}^{n}(x_i - \overline{x})^2 = 98{,}775, \sum_{i=1}^{n}(y_i - \overline{y})^2 = 19.10, \bar{x} = 26.36, \bar{y} = 0.5188, \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \overline{y}) = 826.94.$

   a. Compute the correlation r between the degree of warping and the temperature.
   b. Compute the error sum of squares, the regression sum of squares, and the total sum of squares.

c. Compute the least-squares line for predicting warping from temperature.
d. Predict the warping at a temperature of 40◦C.
e. At what temperature will we predict the warping to be 0.5 mm?
f. Assume it is important that the warping not exceed 0.5 mm. An engineer suggests that if the temperature is kept below the level computed in part (e), we can be sure that the warping will not exceed 0.5 mm. Is this a correct conclusion? Explain.

7. Moisture content in percent by volume (x) and conductivity in mS/m (y) were measured for 50 soil specimens. The means and standard deviations were $\bar{x} = 8.1, s_x = 1.2, \bar{y} = 30.4, s_y = 1.9$. The correlation between conductivity and moisture was computed to be $r = 0.85$. Find the equation of the least-squares line for predicting soil conductivity from moisture content.
8. The following table presents shear strengths (in kN/mm) and weld diameters (in mm) for a sample of spot welds.

Diameter	Strength
4.2	51
4.4	54
4.6	69
4.8	81
5.0	75
5.2	79
5.4	89
5.6	101
5.8	98
6.0	102

a. Construct a scatterplot of strength (y) versus diameter (x). Verify that a linear model is appropriate.
b. Compute the least-squares line for predicting strength from diameter.
c. Compute the fitted value and the residual for each point.
d. If the diameter is increased by 0.3 mm, by how much would you predict the strength to increase or decrease?
e. Predict the strength for a diameter of 5.5 mm.
f. Can the least-squares line be used to predict the strength for a diameter of 8 mm? If so, predict the strength. If not, explain why not.
g. For what diameter would you predict a strength of 95 kN/mm?

9. The article "Testing the Influence of Climate, Human Impact and Fire on the Holocene Population Expansion of Fagus sylvatica in the Southern Prealps (Italy)" (V. Valsecchi, W. Flnsinger, et al., The Holocene 2008:603–614) presents calculations of the ages (in calendar years before 1950) of several sediment samples taken at various depths (in cm) in Lago di Fimon, a lake in Italy. The results are presented in the following table.

Depth	Age
284.5	1255
407.5	3390
512.0	5560
551.0	6670
578.5	7160
697.0	9820
746.5	11,030

   a. Construct a scatterplot of age (y) versus depth (x). Verify that a linear model is appropriate.
   b. Compute the least-squares line for predicting age from depth.
   c. If two samples differ by 100 cm in depth, by how much would you predict their ages to differ? d. Predict the age for a specimen whose depth is 600 cm.
   d. Should the least-squares line be used to predict the age for a depth of 50 cm? If so, predict the age. If not, explain why not.
   e. For what depth would you predict an age of 5000?

10. The processing of raw coal involves "washing," in which coal ash (nonorganic, incombustible material) is removed. The article "Quantifying Sampling Precision for Coal Ash Using Gy's Discrete Model of the Fundamental Error" (Journal of Coal Quality, 1989:33–39) provides data relating the percentage of ash to the density of a coal particle. The average percentage ash for five densities of coal particles was measured. The data are presented in the following table:

Density (g/cm^3)	Percent ash
1.25	1.93
1.325	4.63
1.375	8.95
1.45	15.05
1.55	23.31

   a. Construct a scatterplot of percent ash (y) versus density (x). Verify that a linear model is appropriate.

b. Compute the least-squares line for predicting percent ash from density.
c. If two coal particles differed in density by 0.1 g/cm3 , by how much would you predict their percent ash to differ?
d. Predict the percent ash for particles with density 1.40 g/cm3 .
e. Compute the fitted values.
f. Compute the residuals. Which point has the residual with the largest magnitude?
g. Compute the correlation between density and percent ash.
h. Compute the regression sum of squares, the error sum of squares, and the total sum of squares.
i. Divide the regression sum of squares by the total sum of squares. What is the relationship between this quantity and the correlation coefficient?

11. An agricultural scientist planted alfalfa on several plots of land, identical except for the soil pH. Following are the dry matter yields (in pounds per acre) for each plot.

pH	Yield
4.6	1056
4.8	1833
5.2	1629
5.4	1852
5.6	1783
5.8	2647
6.0	2131

a. Construct a scatterplot of yield (y) versus pH (x). Verify that a linear model is appropriate.
b. Compute the least-squares line for predicting yield from pH.
c. Compute the fitted value and the residual for each point.
d. If the pH is increased by 0.1, by how much would you predict the yield to increase or decrease? e. Predict the yield for a pH of 5.5.
e. Can the least-squares line be used to predict the yield for a pH of 7? If so, predict the yield. If not, explain why not.
f. For what pH would you predict a yield of 1500 pounds per acre?

12. Curing times in days (x) and compressive strengths in MPa (y) were recorded for several concrete specimens. The means and standard deviations of the x and y values were $\bar{x} = 5, s_x = 2, \bar{y} = 1350, s_y = 100$. The correlation between curing time and compressive strength was

computed to be $r = 0.7$. Find the equation of the least-squares line to predict compressive strength from curing time.

13. Varying amounts of pectin were added to canned jellies, to study the relationship between pectin concentration in % (x) and a firmness index (y). The means and standard deviations of the x and y values were $\bar{x} = 3, s_x = 0.5, \bar{y} = 50, s_y = 10$. The correlation between curing time and firmness was computed to be $r = 0.5$. Find the equation of the least-squares line to predict firmness from pectin concentration.

14. An engineer wants to predict the value for y when $x = 4.5$, using the following data set.

x	y	z = ln y	x	y	z = ln y
1	0.2	−1.61	6	2.3	0.83
2	0.3	−1.20	7	2.9	1.06
3	0.5	−0.69	8	4.5	1.50
4	0.5	−0.69	9	8.7	2.16
5	1.3	0.26	10	12.0	2.48

a. Construct a scatterplot of the points (x, y).
b. Should the least-squares line be used to predict the value of y when $x = 4.5$? If so, compute the leastsquares line and the predicted value. If not, explain why not.
c. Construct a scatterplot of the points (x, z), where $z = ln\, y$.
d. Use the least-squares line to predict the value of z when $x = 4.5$. Is this an appropriate method of prediction? Explain why or why not.
e. Let $\hat{z}$ denote the predicted value of z computed in a part (d). Let $\hat{y} = e^{\hat{z}}$ .Explain why $\hat{y}$ is a reasonable predictor of the value of y when $x = 4.5$.

15. A simple random sample of 100 men aged 25–34 averaged 70 inches in height, and had a standard deviation of 3 inches. Their incomes averaged $34,900 and had a standard deviation of $17,200. Fill in the blank: From the least-squares line, we would predict that the income of a man 70 inches tall would be ----------

i. less than $34,900.
ii. greater than $34,900.
iii. equal to $34,900.

iv. We cannot tell unless we know the correlation.

16. A mixture of sucrose and water was heated on a hot plate, and the temperature $(in\ °C)$ was recorded each minute for 20 minutes by three thermocouples. The results are shown in the following table.

Time	T_1	T_2	T_3
0	20	18	21
1	18	22	11
2	29	22	26
3	32	25	35
4	37	37	33
5	36	46	35
6	46	45	44
7	46	44	43
8	56	54	63
9	58	64	68
10	64	69	62
11	72	65	65
12	79	80	80
13	84	74	75
14	82	87	78
15	87	93	88
16	98	90	91
17	103	100	103
18	101	98	109
19	103	103	107
20	102	103	104

a. Compute the least-squares line for estimating the temperature as a function of time, using $T1$ as the value for temperature.
b. Compute the least-squares line for estimating the temperature as a function of time, using $T2$ as the value for temperature.
c. Compute the least-squares line for estimating the temperature as a function of time, using $T3$ as the value for temperature.
d. It is desired to compute a single line to estimate temperature as a function of time. One person gests averaging the three slope estimates to obtain a single slope estimate, and averaging the three intercept estimates to obtain a single intercept estimate. Find the equation of the line that results from this method.
e. Someone else suggests averaging the three temperature measurements at each time to obtain $T = (T1 + T2 +$

$T3)/3$. Compute the least-squares line using T as the value for temperature.

f. Are the results of parts (d) and (e) different?