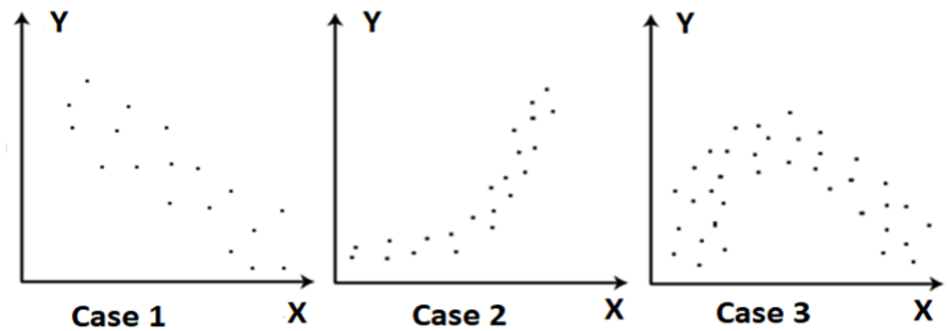
Tutorial Worksheet 4 - Regression Analysis (with Solutions)

Objective Questions

1. Which of the following three cases depicts a 'non-monotonic relationship' between the two variables X and Y?



   (a) Case 1 (Plot in the left) (b) Case 2 (Plot in the centre) (c) Case 3 (Plot in the right) (d) None of the plots depicts 'non-monotonic relationship'

   Solution: Case 3 (Plot in the right)

2. Look into the following three statements carefully:

   (I) Karl Pearson's Correlation Analysis method can be used to find correlation coefficient between two numerical attributes.

   (II) Charles Spearman's Correlation Analysis method can be used to find correlation coefficient between two ordinal attributes

   (III) Chi-square Coefficient of Correlation Analysis method can be used to find correlation coefficient between two nominal attributes

   Which of the following options is TRUE?

   (a) Only Statement I is correct, other two statements are incorrect
   (b) Only Statement II is correct, other two statements are incorrect
   (c) Only Statement III is correct, other two statements are incorrect
   (d) All three statements, I, II, and III are correct

   Solution: All three statements, I, II, and III are correct

3. In order to compute the Charles Spearman's Coefficient of Correlation between two variables, a rank is assigned to each data. Further, the ranks are modified/corrected if two data are of same value. The observations of such two variables X and Y are given in below table.

X	5	10	15	20	25
Y	200	300	300	300	500

   What will be the modified (final) rank of the observations of variable Y = 200, 300, 300, 300, 500?

   (a) 5, 4, 3, 2, 1 (b) 3, 2, 2, 2, 1 (c) 5, 3, 3, 3, 1 (d) 5, 4.5, 3.5, 2.5, 1.

   Solution: 5, 3, 3, 3, 1

4. In regression analysis, the variable that is being predicted is called as?

(a) Response (b) Regressor (c) Independent variable (d) Dependent variable

Solution: Response

5. In the least square method of regression analysis, what is the relationship between the sum of squares of the errors (SSE), total corrected sum of squares (SST) and the coefficient of determination (R^2)?

   (a) $R^2 = 1 - \frac{SSE}{SST}$ (b) $R^2 = 1 + \frac{SSE}{SST}$ (c) $R^2 = 1 - \frac{SST}{SSE}$ (d) $R^2 = 1 + \frac{SST}{SSE}$

   Solution: $R^2 = 1 - \frac{SSE}{SST}$

6. If the coefficient of determination is a positive value, then the regression equation?

   (a) must have a positive slope (b) must have a negative slope (c) could have either a positive or a negative slope (d) must have a positive y intercept

   Solution: could have either a positive or a negative slope

7. What is (are) the required assumption(s) for the auto-regression analysis?

   (a) The time series under consideration is non-stationary.
   (b) The time series under consideration is non-uniform.
   (c) The time series under consideration is stationary, but not uniform.
   (d) The time series under consideration is both stationary and uniform.

   Solution: The time series under consideration is both stationary and uniform.

8. For a given dataset, to compute the relationship between the variables x and y, following two regression models are obtained.

   (I) Model 1: $\mathrm{Y} = 2\mathrm{X}_2 + \mathrm{X}_1$, with R^2 score $= 0.68$
   (II) Model 2: $\mathrm{Y} = 3\mathrm{X}_3 + 2\mathrm{X}_2 + \mathrm{X}_1$ with R^2 score $= 0.87$

   Which model is more acceptable?

   (a) a) Model 1 (b) Model 2 (c) Both models are equally acceptable (d) None of the model is acceptable

   Solution: Model 2

9. In a regression analysis if the coefficient of determination $R^2 = 1$, then sum of squares of the errors (SSE) must be equal to?

   (a) 1 (b) 0 (c) Any positive value (d) Infinity

   Solution: 0

10. A study was conducted to investigate the relationship between a student's pocket-money in AED per year and his/her family income in AED per year. Data of the same were collected for 100 students, and a simple linear regression line of the form $Y = \alpha + \beta X$ was fitted. Now suppose that the data of both the family income and pocket money is converted from AED to USD. The impact of this conversion on the regression line is

    (a) The sign of the slope will change, but the magnitude of the slope will remain unchanged
    (b) The magnitude of the slope will change, but the sign of the slope will remain unchanged
    (c) Both the sign and magnitude of the slope will change
    (d) None of the sign and magnitude of the slope will change

    Solution: None of the sign and magnitude of the slope will change

Subjective Questions

Problem 1.

The ranks for 10 students on foundation year examination (A) and L3 examination (B) are given in the table below. What is the Spearman's rank correlation coefficient between A and B?

Rank in A	5	3	10	8	2	6	9	4	1	7
Rank in B	6	2	10	1	4.5	4.5	8	3	9	7

Solution:

Rank in A	5	3	10	8	2	6	9	4	1	7
Rank in B	6	2	10	1	4.5	4.5	8	3	9	7
Difference in Ranks (d_i)	-1	1	0	7	-2.5	1.5	1	1	-8	0
d_i^2	1	1	0	49	6.25	2.25	1	1	64	0

Thus, $\sum d_i^2 = 125.5$ hence, the rank correlation coefficient

$$r_s = 1 - \frac{6\Sigma d_i^2}{n\left(n^2-1\right)} = 1 - \frac{6 \times 125.5}{10 \times 99} = 0.24$$

Problem 2.

In order to find out the correlation between an independent variable X and a dependent variable Y, following information is available.

$$\sum\left(Y_i-\bar{Y}\right)\left(X_i-\bar{X}\right) = 498, \sum\left(X_i-\bar{X}\right)^2 = 338, \sum\left(Y_i-\bar{Y}\right)^2 = 1212$$

What is the value of Karl Pearson's coefficient of Correlation between X and Y ?

Solution:

$r^2 = \frac{\Sigma\left(X_i-\bar{X}\right)\left(Y_{\bar{i}}-\bar{Y}\right)}{\sqrt{\Sigma\left(X_i-\bar{X}\right)^2}\sqrt{\Sigma\left(Y_T-\bar{Y}\right)^2}} = \frac{498}{\sqrt{338\times1212}} = 0.778$

Problem 3.

The sigmoid function takes the form $f(x) = \frac{1}{1+e^{-x}}$. What is the first order differentiation of $f(x)$ w.r.t to $\mathbf{x}$?

Solution:

The sigmoid function is $f(x) = \frac{1}{1+e^{-x}}$. Differentiating w.r.t to x,

$$f(x) = \left(\frac{1}{1+e^{-x}}\right)^2 \frac{d}{dx}\left(1+e^{-x}\right) = \left(\frac{1}{1+e^{-x}}\right)^2\left(-e^{-x}\right) = f(x)(1-f(x))$$

Problem 4.

A simple linear regression model of the form $Y = a + bX$ is used to compute the relationship between the variables X and Y. Suppose there are n sample points, $\left(x_i, y_i\right), i = 1, 2, \ldots, n$, and $\bar{x}$ and $\bar{y}$ are their corresponding means. The value of linear regression model coefficient b is given by?

Solution:

The value of linear regression model coefficient b is given by $b = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})}$

Problem 5. What are the maximum and minimum values of the function $p(x) = \frac{1}{1+e^{-\frac{-(x-\mu)}{s}}}$, for $-\infty < x < \infty$, and $\mathbf{s}, \boldsymbol{\mu}$ are constants?

Solution:

$$\begin{aligned}&\text{At } x = \infty, p(x) = 0\\&\text{At } x = -\infty, p(x) = 1\end{aligned}$$

Problem 6. Studies have shown that people who suffer sudden cardiac arrest have a better chance of survival if a defibrillator shock is administered very soon after cardiac arrest. How is survival rate related to the time between when cardiac arrest occurs and when the defibrillator shock is delivered? The accompanying data give $y =$ survival rate (percent) and $x = 5$ mean call-to shock time (minutes) for a cardiac rehabilitation center (in which cardiac arrests occurred while victims were hospitalized and so the call-to-shock time tended to be short) and for four communities of different sizes:

Mean call-to-shock-time, x :	2	6	7	9	12
Survival rate, y :	90	45	30	5	2

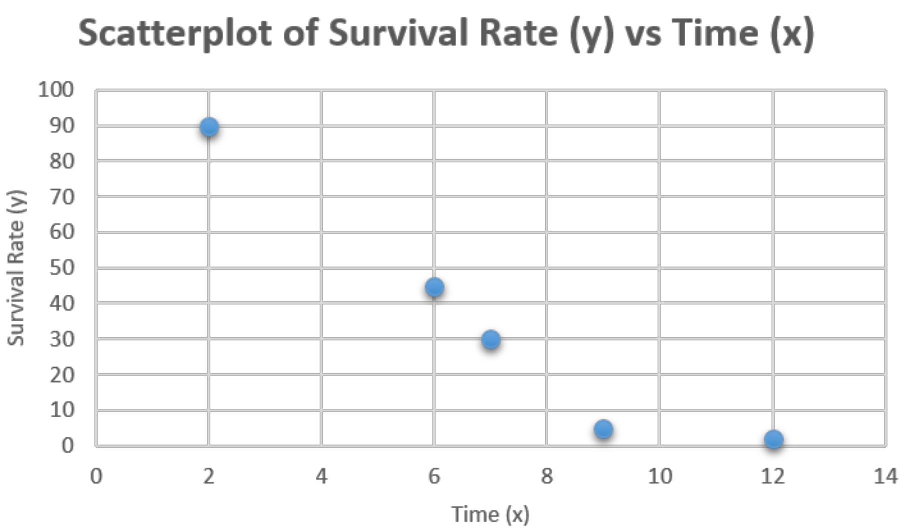
(a) Construct a scatterplot for these data. How would you describe the relationship between mean call-to shock time and survival rate?

(b) Find the equation of the least-squares line.

(c) Use the least-squares line to predict survival rate for a community with a mean call-to-shock time of 10 minutes.

Solution:

(a) The relationship between mean call-to shock time and survival rate is roughly linear.



(b) The equation for the regression line can be computed as follows:-

	Time x	Survival Rate y	X^2	Y^2	XY
	2	90	4	8100	180
	6	45	36	2025	270
	7	30	49	900	210
	9	5	81	25	45
	12	2	144	4	24
Total:	36	172	314	11054	729

$$\bar{x} = 36/5 = 7.2, \bar{y} = 172/5 = 34.4, SS_{nx} = \Sigma x^2 - \frac{(\Sigma x)^2}{n} = 314 - \frac{36^2}{5} = 54.8$$

$$SS_{xy} = \sum xy - \frac{\Sigma x \Sigma y}{n} = 729 - \frac{36^{*}172}{5} = 1238.4, SS_{yy} = \Sigma y^2 - \frac{(\Sigma y)^2}{n} = 11054 - \frac{172^2}{5} = 5137.2$$

$$\text{Slope b } = \frac{S_{xy}}{SS_{xx}} = \frac{-509.4}{54.8} = -9.29562, \text{ Intercept } a = \bar{y} - b\bar{x} = 7.2 - (-9.29562) * 34.4 = 101.3285$$

Therefore the regression line of Y on X is $Y = -9.29562X + 101.3285$.

(c) Given $x = 10; Y = (-9.29562 * 10) + 101.3285 = 8.3723$

Problem 7. Let x be the size of a house (in square feet) and y be the amount of natural gas used (therms) during a specified period. Suppose that for a particular community, x and y are related according to the simple linear regression model with

$\beta =$ slope of population regression line $= 0.017$, $\alpha = y$ intercept of population regression line $= -5.0$

Houses in this community range in size from 1000 to 3000 square feet.

(a) What is the equation of the population regression line?

(b) Graph the population regression line by first finding the point on the line corresponding to $x = 1000$ and then the point corresponding to $x = 2000$, and drawing a line through these points.

(c) What is the mean value of gas usage for houses with 2100 sq. ft. of space?

(d) What is the average change in usage associated with a 1 sq. ft. increase in size?

(e) What is the average change in usage associated with a 100 sq. ft. increase in size?

Solution: Let x be the size of a house and y be the amount of natural gas used.

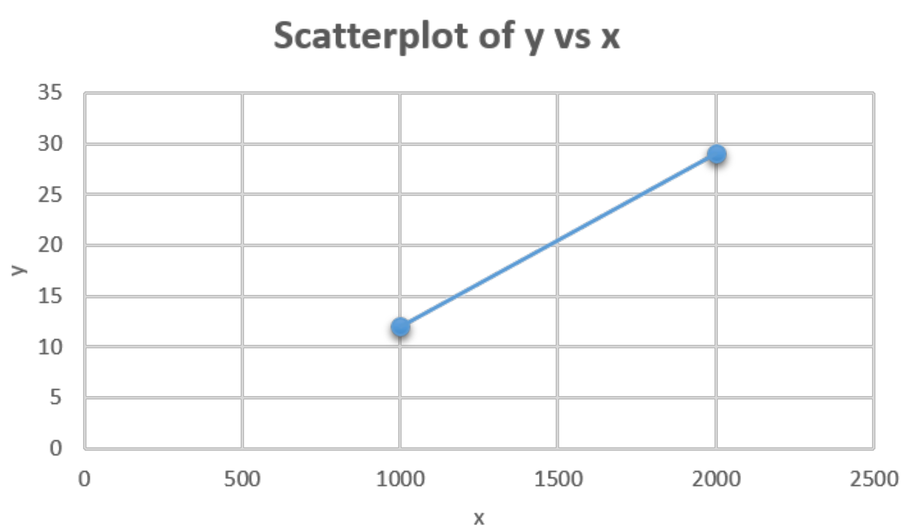
(a) The equation of the population regression line can be derived from the given information i.e., slope of the regression line is 0.017 and the intercept of the regression line is -5.0. So, the equation of the population regression line is given by,

$$y = \alpha + \beta x = -5.0 + 0.017x$$

(b) To find the y values corresponding to the x values 1000 and 2000. We substitute 1000 and 2000 for x in the regression equation $y = -5.0 + 0.017x$ i.e.,

$$y = -5.0 + 0.017(1000) = 12 \text{ and } y = -5.0 + 0.017(2000) = 29$$

So, the points are (1000, 12) and (2000, 29). Using these points we construct the plot as shown below:



(c) To compute the mean value of gas usage for houses with 2100 sq. ft. of space, substitute 2100 for x

$$y = -5.0 + 0.017(2100) = 30.7$$

The mean value of gas usage for houses with 2100 sq. ft. of space is 30.7 therms.

(d) The slope of the regression equation is 0.017. From this value it can be said that a 1 sq. ft. increase in the house size will lead to an increase of 0.017 units in natural gas usage.

(e) The slope of the regression equation is 0.017. From this value it can be said that a unit increase in the house size will lead to increase in 0.017 units in natural gas usage. From this value it can be said that a 100 sq. ft. increase in the house size will lead to an increase of $0.017 \times 100 = 1.7$ units in natural gas usage.

Problem 8. The data in the accompanying table is from the paper "Six-Minute Walk Test in Children and Adolescents" (The Journal of Pediatrics [2007]: 395-399). Two hundred and eighty boys completed a test that measures the distance that the subject can walk on a flat, hard surface in 6 minutes. For each age group shown in the table, the median distance walked by the boys in that age group is also given.

Age Group	Representative Age (Mid point)	Median Distance (Meters)
$3-5$	4	$544 \cdot 3$
$6-8$	7	584.0
$9-11$	10	$667 \cdot 3$
$12-15$	13.5	701.1
$16-18$	17	727.6

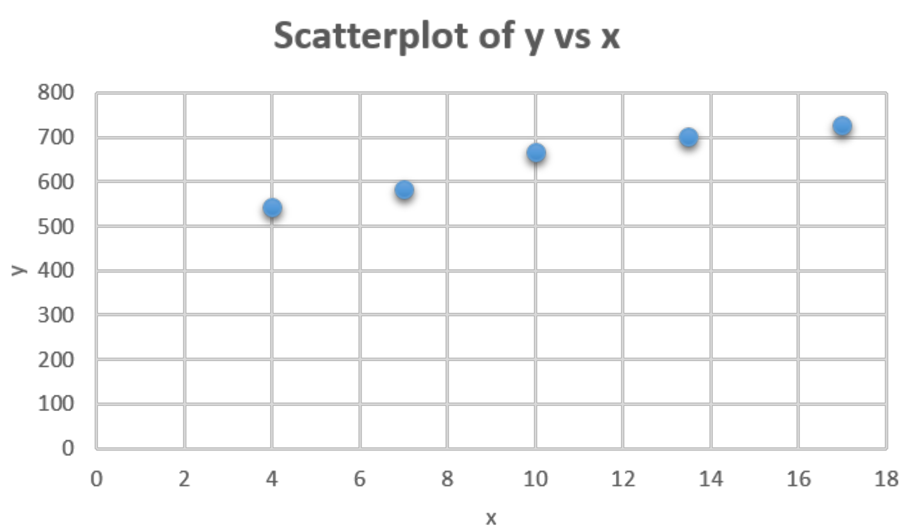
(a) With $x =$ representative age and $y =$ median distance walked in 6 minutes, construct a scatterplot. Does the pattern in the scatterplot look linear?

(b) Find the equation of the least-squares regression line that describes the relationship between median distance walked in 6 minutes and representative age.

(c) Compute the five residuals and construct a residual plot. Are there any unusual features in the plot?

Solution:

(a) From the scatterplot we can infer that the relationship between representative age and median distance walked in 6 minutes looks linear.



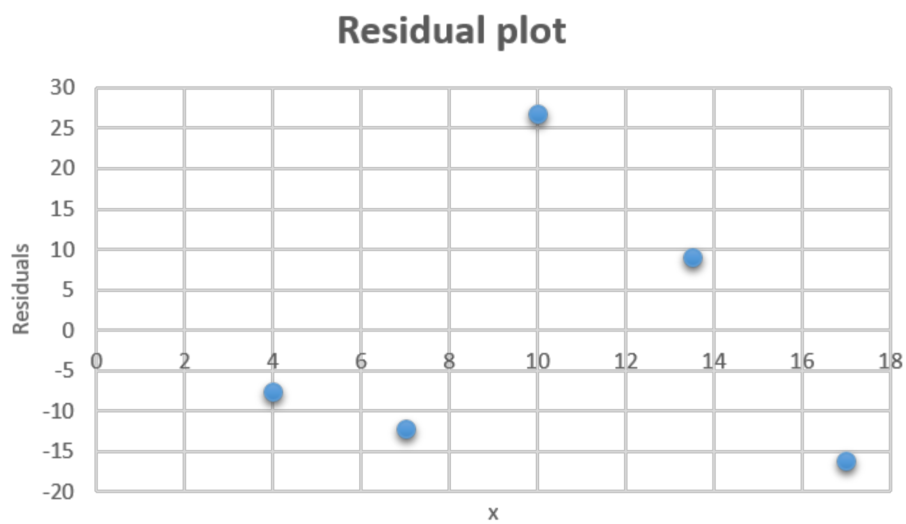
(b) $x =$ representative age, $y =$ median distance walked in 6 minutes, $n = 5$, $\sum x = 51.5$, $\sum y = 3224.3$, $\sum x^2 = 636.25$, $\sum y^2 = 2103550.75$, and $\sum xy = 34772.25$.

Hence, $\bar{x} = 10.3, \bar{y} = 644.86$, $S_{xy} = \sum xy - \frac{(\sum x)(\sum y)}{n} = 1561.96$, $S_{xx} = \sum x^2 - \frac{(\sum x)^2}{n} = 105.8$.

Thus, $b = \frac{S_{xy}}{s_{xx}} = 14.76$ and $a = \bar{y} - b\bar{x} = 492.832$. Hence the required least square regression equation is: $\hat{y} = 492.832 + 14.76x$

(c) From the residual plot we can conclude that there are no unusual patterns in the plot.

x	**y**	Predicted **y**	Residuals
4	544.3	551.872	-7.572
7	584	596.152	-12.152
10	667.3	640.432	26.868
13.5	701.1	692.092	9.008
17	727.6	743.752	-16.152



Problem 9. A simple linear regression model was used to describe the relationship between y = hardness of molded plastic and x = amount of time elapsed since the end of the molding process. Summary quantities included $n = 15$, SSResid = 1235.470, and SSTo = 25,321.368.

(a) Calculate a point estimate of σ. On how many degrees of freedom is the estimate based?

(b) What percentage of observed variation in hardness can be explained by the simple linear regression model relationship between hardness and elapsed time?

Solution:

(a) A point estimate of σ is the estimated standard deviation (s_e) which is given by

$$s_e = \frac{\text{SSResid}}{n-2} = \frac{1235.470}{15-2} = 95.0362$$

Therefore, the point estimate of σ is 95.0362.

The number of degrees of freedom associated with estimating σ in simple linear regression is

$$\text{df } = n - 2 = 15 - 2 = 13$$

Therefore, the number of degrees of freedom associated with estimating σ in simple linear regression is 13.

(b) The percentage of observed variation in dependent variable can be explained by the simple linear regression model is the coefficient of determination $\left(R^2\right)$ which is defined as follows:

$$R^2 = 1 - \frac{\text{SSResid}}{\text{SSTo}}$$

Hence, the percentage of observed variation in hardness that can be explained by the simple linear regression model between hardness and elapsed time is given by

$$R^2 = 1 - \frac{\text{SSResid}}{\text{SSTo}} = 1 - \frac{1235.470}{25321.368} = 1 - 0.0488 = 0.9512$$

Therefore, 95.12% of the observed variation in hardness can be explained by the simple linear regression model between hardness and elapsed time.

Problem 10. An experiment to study the relationship between x = time spent exercising (minutes) and y = amount of oxygen consumed during the exercise period resulted in the following summary statistics.

$$n = 20, \quad \sum x = 50, \quad \sum y = 16705,$$
$$\sum x^2 = 150, \quad \sum y^2 = 14194231, \quad \sum xy = 44194$$

(a) Estimate the slope and y intercept of the population regression line.

(b) One sample observation on oxygen usage was 757 for a 2-minute exercise period. What amount of oxygen consumption would you predict for this exercise period, and what is the corresponding residual?

(c) Compute a 99% confidence interval for the average change in oxygen consumption associated with a 1-minute increase in exercise time.

Solution:

(a) x = time spent exercising (min), y = amount of oxygen consumed during the exercise period
$n = 20, \quad \sum x = 50, \quad \sum y = 16705, \quad \sum x^2 = 150, \quad \sum y^2 = 14194231, \quad \sum xy = 44194, \quad \bar{x} = 2.5, \quad \bar{y} = 835.25$
$S_{xy} = \sum xy - \frac{(\sum x)(\sum y)}{n} = 2431.5, \quad S_{xx} = \sum x^2 - \frac{(\sum x)^2}{n} = 25$
$b = \frac{s_{xy}}{s_{xx}} = 97.26, \quad a = \bar{y} - b\bar{x} = 592.1$

(b) $\hat{y} = 592.1 + 97.26x$, hence for $x = 2, \hat{y} = 786.62$. The corresponding residual is $757 - 786.62 = -29.62$

(c)
$$SSR = \sum y^2 - a\sum y - b\sum xy = 14194231 - 592.1 \times 16705 - 97.26 \times 44194 = 4892.06$$
$$S_e = \sqrt{\frac{SSR}{n-2}} = \sqrt{\frac{4892.06}{20-2}} = 16.48$$
$$S_b = \frac{S_e}{\sqrt{S_{xx}}} = \frac{16.48}{\sqrt{25}} = 3.29$$

99% confidence interval for slope requires a t critical value based on df = n − 2 = 20 − 2 = 18, which is 2.88 (from t-distribution tables)

The resulting interval is then $= b\pm$ critical value $\times S_b = 97.26 \pm 2.88 \times 3.29 = (87.76, 106.76)$

Problem 11. A simple linear regression model was used to describe the relationship between sales revenue y (in thousands of dollars) and advertising expenditure x (also in thousands of dollars) for fast-food outlets during a 3-month period. A sample of 15 outlets yielded the accompanying summary quantities.
$\sum x = 14.10, \quad \sum y = 1438.50, \quad \sum x^2 = 13.92, \quad \sum y^2 = 140354$
$\sum xy = 1387.20, \quad \sum (y - \bar{y})^2 = 2401.85, \quad \sum (y - \hat{y})^2 = 561.46$

(a) What proportion of observed variation in sales revenue can be attributed to the linear relationship between revenue and advertising expenditure?

(b) Calculate s_e and s_b.

(c) Obtain a 90% confidence interval for β, the average change in revenue associated with a \$1000 (that is, 1-unit) increase in advertising expenditure.

(d) Test the hypothesis $H_0 : \beta = 0$ versus $H_a : \beta \neq 0$ using a significance level of 0.05. What does your conclusion say about the nature of the relationship between x and y?

Solution:

(a) The coefficient of determination, $R^2 = 1 - \frac{\text{SSR}}{\text{SST}}$

Here,

$$\text{SSR} = \sum (y - \hat{y})^2 = 561.46, \qquad \text{SST} = \sum (y - \bar{y})^2 = 2401.85$$

Therefore, the value of coefficient of determination can be calculated as follows:

$$R^2 = 1 - \frac{\text{SSR}}{\text{SST}} = 1 - \frac{561.46}{2401.85} = 1 - 0.2338 = 0.7662$$

Hence, the amount of observed variation in sales revenue that can be attributed to the linear relationship between revenue and advertising expenditure is 0.7662, that is, 76.62% of variation in sales revenue can be attributed to the linear relationship between revenue and advertising expenditure.

(b) A point estimate of σ is the estimated standard deviation (s_e) which is given by $s_e^2 = \frac{\text{SSR}}{n-2}$

For the given data, the point estimate of σ is calculated as follows:

$$s_e^2 = \frac{\text{SSR}}{n-2} = \frac{561.46}{15-2} = 43.18923, \qquad s_e = \sqrt{s_e^2} = \sqrt{43.18923} = 6.5719$$

Hence, the estimated standard deviation, $s_e = 6.5719$

$$S_{xx} = \sum (x - \bar{x})^2 = \sum x^2 - \frac{\left(\sum x\right)^2}{n} = 13.92 - \frac{(14.10)^2}{15} = 0.666$$
$$s_b = \frac{s_e}{\sqrt{S_{xx}}} = \frac{6.5719}{\sqrt{0.666}} = 8.0529$$

Therefore, the values of s_e and s_b are 6.5719 and 8.0529 respectively.

(c) The 90% confidence interval for β, the slope of the population regression line is given by $b \pm (t \text{ critical value }) \times s_b$

The value of slope b can be calculated as follows: $b = \frac{S_{xy}}{S_{xx}}$

Here,

$$S_{xy} = \sum xy - \frac{\left(\sum x\right)\left(\sum y\right)}{n} = 1387.20 - \frac{(14.10)(1438.5)}{15} = 1387.2 - 1352.19 = 35.01$$

Therefore, the values of slope and the corresponding standard error of the slope are,

$$b = \frac{S_{xy}}{S_{xx}} = \frac{35.01}{0.666} = 52.568 \quad s_b = 8.0529$$

The number of degrees of freedom associated with estimating σ in simple linear regression is, $n - 2 = 15 - 2 = 13$.

Then, the critical value of t corresponding to 13 degrees of freedom at 90% confidence level is 1.771.

Now, the 90% confidence interval for slope is,

$$CI = b \pm (t \text{ critical value }) \times s_b = 52.568 \pm 1.771 \times 8.0529 = 35.01 \pm 14.2611 = (38.3065, 66.8287)$$

Hence, the 90% confidence interval for the slope of the regression line is $(38.3065, 66.8287)$.

(d) To test the hypothesis $H_0 : \beta = 0$ versus $H_a : \beta \neq 0$ at $\alpha = 0.05$ level of significance.

Test statistics,

$$t = \frac{b - \text{hypothesized value}}{S_b} = \frac{b - 0}{S_b} = \frac{52.568}{8.0529} = 6.5278$$

The critical value, $t_{13,0.025} = 2.1604$. Since, $|t| > t_{13,0.025}$, we reject $H_0 : \beta = 0$

Thus we might conclude that there is a useful linear relationship between x and y, and knowledge of x is useful for predicting y.

Problem 12. The article "Effect of Temperature on the pH of Skim Milk" (Journal of Dairy Research [1988]: 277-280) reported on a study involving x = temperature (°C) under specified experimental conditions and y = milk pH. The accompanying data (read from a graph) are a representative subset of that which appeared in the article:

x : 4	4	24	24	25	38	38	40	45	50	55	56	60	67	70	78
y : 6.85	6.79	6.63	6.65	6.72	6.62	6.57	6.52	6.50	6.48	6.42	6.41	6.38	6.34	6.32	6.34

$\sum x = 678 \sum y = 104.54 \sum x^2 = 36,056 \sum y^2 = 683.4470 \sum xy = 4376.36$

(a) Do these data strongly suggest that there is a negative linear relationship between temperature and pH? State and test the relevant hypotheses using a significance level of 0.01.

(b) Obtain a 95% confidence interval for $\alpha + \beta(40)$ the mean milk pH when the milk temperature is 40°C.

(c) Obtain a 99% prediction interval for a single pH observation to be made when milk temperature = 40°C.

(d) Would you recommend using the data to calculate a 95% confidence interval for the mean pH when the temperature is 90°C ? Why or why not?

Solution:

(a) x = temperature (°C) under specified experimental conditions and y = milk pH, n = 16

$S_{xy} = \sum xy - \frac{(\sum x)(\sum y)}{n} = -53.52, \quad S_{xx} = \sum x^2 - \frac{(\sum x)^2}{n} = 7325.75$

$b = \frac{s_{xy}}{s_{xx}} = -0.0073, \quad a = \bar{y} - b\bar{x} = 6.8431$

The required regression equation is: $\hat{y} = 6.8431 - 0.0073x$

To test the hypothesis $H_0 : \beta = 0$ versus $H_a : \beta \neq 0$ we use the test statistic, $t = \frac{b\text{-hypothesized value}}{S_b}$

We will first calculate $S_b = \frac{b-0}{s_b}$.

$\text{SSR} = \sum y^2 - a \sum y - b \sum xy = 0.0177, \quad S_e = \sqrt{\frac{SSR}{n-2}} = 0.0356, \text{ and } S_b = \frac{S_e}{\sqrt{S_{xx}}} = 0.000416$

Therefore, $t = \frac{b}{s_b} = -17.548$, and the critical value at $\alpha = 0.01$ is $t_{14,0.005} = 2.9768$

Since, $|t| > t_{14,0.005}$, we reject $H_0 : \beta = 0$

Thus we might conclude that there is a useful linear relationship between x and y, and knowledge of x is useful for predicting y.

(b) The required calculation for 95% prediction interval for y^* are as follows $S_{a+b(x^*)} = S_e\sqrt{\frac{1}{n} + \frac{(x^*-\bar{x})^2}{S_{xx}}} = 0.0089$

The t critical value for the df 14 and 95% prediction level is $t_{14,0.025} = 2.1448$

The prediction interval formula is given by: $\mathrm{a} + \mathrm{b}\,(x^*) \pm t_{\text{critical value}} \sqrt{S_e^2 + S_{a+b(x^*)}^2}$

Substituting the calculated values in the above formula we obtain the required prediction interval as:

$6.8431 - (0.0073 * 40) \pm 2.1448\sqrt{0.0356^2 + 0.0089^2} = (6.4724, 6.6298)$

(c) For calculating the 99% prediction interval for y^* we use the previously stated prediction interval formula. The critical value for t distribution with df 14 and 99% prediction level is $t_{14,0.005} = 2.9768$

Substituting the calculated values in the above formula we obtain the required prediction interval as:

$6.8431 - (0.0073 * 40) \pm 2.9768\sqrt{0.0356^2 + 0.0089^2} = (6.4419, 6.6603)$

(d) The required calculation for 95% prediction interval for y^* are as follows

$$S_{a+b(x^*)} = S_e\sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}} = 0.021716$$

The t critical value for the df 14 and 95% prediction level is $t_{14,0.025} = 2.1448$

The prediction interval formula is given by:

$\mathrm{a} + \mathrm{b}\,(x^*) \pm t_{\text{critical value}} \sqrt{S_e^2 + S_{a+b(x^*)}^2}$

Substituting the calculated values in the above formula we obtain the required prediction interval as:

$6.8431 - (0.0073 * 90) \pm 2.1448\sqrt{0.0356^2 + 0.021716^2} = (6.0967, 6.2755)$.

Since, the prediction interval is not wide and the actual prediction also lies within the prediction interval. So, we will recommend using the data to calculate a 95% confidence interval for the mean pH when the temperature is 90°C.