

REPORT

AIM:

In this section, we aim at predicting the number of sales made (Sales) from the number of sales calls made (Calls).

Thus, the dependent variable (y) is Sales and the independent variable is Calls (x).

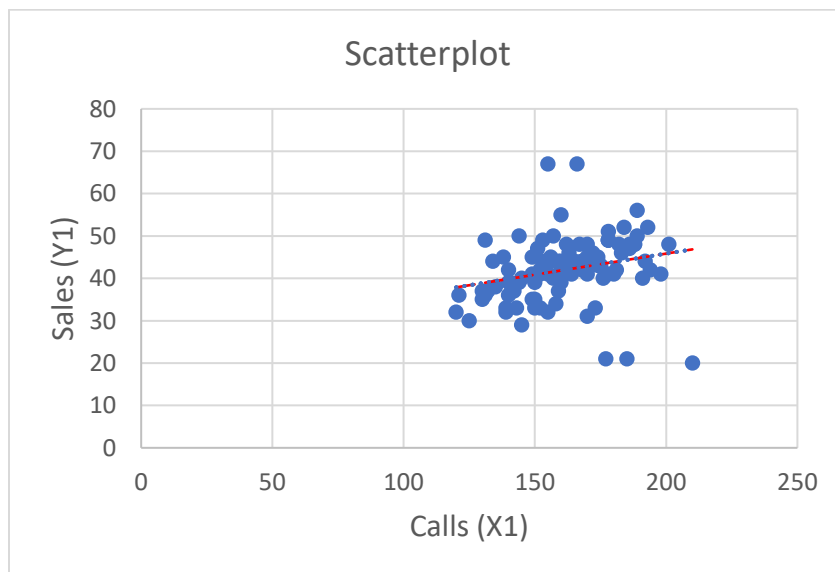
We have a set of 100 observed data values of the number of sales calls made and the number of sales made this week.

We use this data observation to predict the number of sales made (Sales) from the number of sales calls made (Calls).

DATA VISUALIZATION:

We plot the scatterplot for the number of sales and the number of sales calls made and observe a slightly increasing trend in the number of sales made with the increase in the number of sales calls made.

We also observe a few outliers in the data.



LINE OF BEST FIT:

We fit a linear regression model and obtain the line of "best fit".

The line of best fit, which describes the relationship between the number of sales and the number of sales calls made is:

Number of sales = $25.927 + 0.0996 * \text{number of sales calls}$

This means that on an average, with an increase in the number of sales calls by 1, the number of sales increases by 0.0996

CORRELATION ANALYSIS:

The correlation between the number of sales and the number of sales calls made is 0.249

This low positive value of coefficient of correlation (0.249) indicates a weak positive linear relationship between the number of sales made this week (Sales) and the number of sales calls made this week (Calls).

COEFFICIENT OF DETERMINATION:

The coefficient of determination between the number of sales and the number of sales calls made is 0.062

This represents that the fitted regression line explains only 6.2% of the variability of the response data around its mean.

This is indicative of poor performance of the model in predicting the number of sales on the basis of the number of sales calls.

TESTING THE SIGNIFICANCE OF THE REGRESSION MODEL:

We perform a hypothesis testing analysis to test the model.

We obtain the value of the F statistic as:

$F(1,98) = 6.4975$ with p-value = 0.01235

At level of significance 0.1, the linear regression model obtained to predict the number of sales from the number of sales calls made is statistically significant.

Also, at level of significance 0.05, the linear regression model obtained to predict the number of sales from the number of sales calls made is statistically significant.

Thus, the linear regression model is statistically significant in predicting the number of sales from the number of sales calls made.

SLOPE PARAMETER:

The estimated slope parameter is 0.096 which means that on an average, with an increase in the number of sales calls by 1, the number of sales increases by 0.0996

The standard error of the slope parameter is 0.039

On testing the null hypothesis $H_0: b = 0$ against the alternative hypothesis $H_1: b \neq 0$,

The value of the t statistic is 2.549 and the p-value of the hypothesis test is 0.01235

Thus, at level of significance 0.05 and 0.1, we reject the null hypothesis $H_0: b = 0$ against the alternative hypothesis $H_1: b \neq 0$.

This indicates that the slope parameter is statistically significant.

The confidence interval for the slope parameter, using a 95% confidence level is between 0.022 and 0.177

CONFIDENCE AND PRECTION INTERVALS:

We use the obtained linear regression model to predict the number of sales when the number of sales calls made is 160.

We obtain that the expected number of sales is 41.8597 when the number of sales calls made is 160.

The 99% confidence interval for the number of sales, when the number of sales calls is 160 is equal to (40.08592, 43.63361)

This means that with 99% confidence, the mean value of the number of sales lies between 40.08592 and 43.63361 when the number of sales calls is 160.

The 99% prediction interval for the number of sales, when the number of sales calls is 160 is equal to (24.11223, 59.60729)

This means that the actual value of the number of sales will lie between 24.11223 and 59.60729, when the number of sales calls is 160, with 99% probability.

We can extrapolate the regression line to predict the value of the dependent variable for the values of the independent variable that are outside the range of the sample values. But that requires strong assumptions and is thus generally avoided as it is often inappropriate and may yield incredible results.

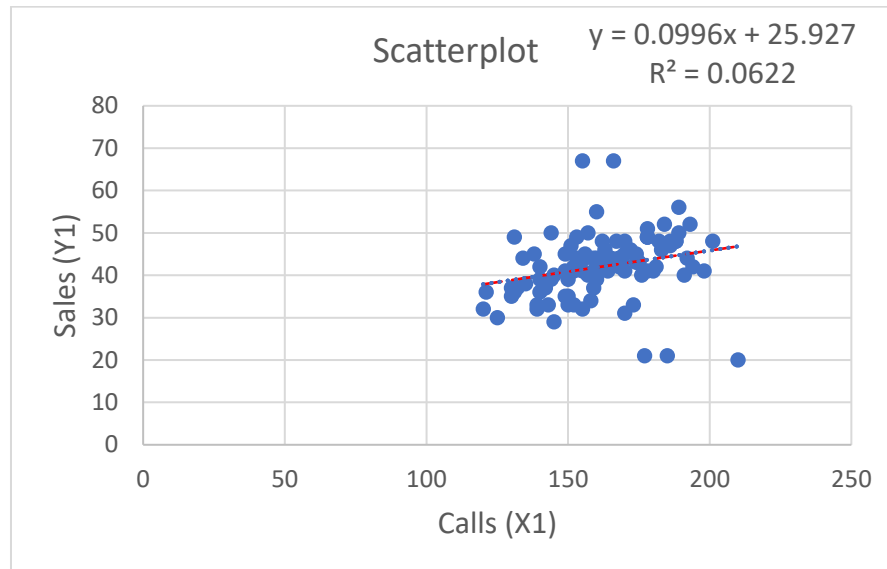
APPENDIX

The dependent variable (Y) is SALES (Y).

The independent variable (X) is CALLS (X1).

1.

The scatterplot generated for the dependent variable SALES and the independent variable CALLS, including the graph of the best fit line is:



2.

	Coefficients
Intercept	25.9271188
Calls (X1)	0.09957902

Hence, the equation of the “best fit” line, which describes the relationship between SALES (y) and CALLS (x) is:

$$y = 25.927 + 0.0996x$$

3.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.24935706
R Square	0.06217894
Adjusted R Square	0.05260934
Standard Error	7.46665747
Observations	100

The coefficient of correlation is the Multiple R statistic in the above output.

Hence, the coefficient of correlation is 0.249

This low positive value of coefficient of correlation (0.249) indicates a weak positive linear relationship between the number of sales made this week (SALES) and the number of sales calls made this week (CALLS).

4.

The coefficient of determination is the R square statistic in the above output.

Hence, the coefficient of determination is 0.062

The coefficient of determination is a statistical measure of how close the data are to the fitted regression line.

Here, it indicates that the fitted regression line explains 6.2% of the variability of the response data around its mean.

5.

ANOVA

	df	SS	MS	F	Significance F
Regression	1	362.2445743	362.244574	6.49754704	0.012353773
Residual	98	5463.595426	55.7509737		
Total	99	5825.84			

	Coefficients	Standard Error	t Stat	P-value
Intercept	25.9271188	6.365130567	4.07330511	9.4123E-05
Calls (X1)	0.09957902	0.039065477	2.54902865	0.01235377

H0: $b = 0$

H1: $b \neq 0$

The value of the test statistic for testing this hypothesis is given by t Stat in the above output.

Thus, the value of the test statistic is equal to 2.549

Under H0, this test statistic follows a t-distribution with $df = 1$

The p-value of this test statistic is given by the P-value in the above output.

Thus, the p-value of the test statistic is equal to 0.012

The level of significance = 0.1

Since the level of significance is greater than the p-value,

We reject H_0 against H_1 .

Thus, we have enough evidence to conclude that the slope is not equal to 0.

6.

Based on the above steps,

The independent variable has a weak positive linear relationship with the dependent variable (since the coefficient of correlation = 0.249)

The fitted regression line can explain only 6.2% of the variability of the response data around its mean (since the coefficient of determination = 0.062)

The hypothesis testing rejects the null hypothesis of the value of the slope being zero at level of significance = 0.1

Hence, we conclude that the independent variable can statistically significantly predict the dependent variable by a linear regression model fit at level of significance = 0.1

7.

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	25.9271188	6.365130567	4.07330511	9.4123E-05	13.29572471	38.558513
Calls (X1)	0.09957902	0.039065477	2.54902865	0.01235377	0.022054854	0.17710319

The confidence interval for b, using a 95% confidence level is given by the Lower 95% and Upper 95% of the Calls (X1) in the above output.

Hence, the confidence interval for b, using a 95% confidence level is between 0.022 and 0.177

8.

The selected value of the independent variable CALLS for which we will be predicting the interval is $x^* = 160$

The predicted value of the dependent variable $SALES = \hat{y} = 25.927 + 0.0996 * 160 = 41.859762$

The t-value at 99% confidence interval with $df = n - 2 = 98$ is $t_{n-2}^* = 2.365002$

The mean of the values of the dependent variable $= \bar{x} = 161.81$

The variance of the values of the dependent variable $= s_x^2 = 365.3139$

The standard error $= s_y = 7.4666$

The sample size $= n = 100$

The confidence interval is given by the formula

$$\hat{y} \pm t_{n-2}^* s_y \sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{(n-1)s_x^2}}$$

On substituting the values, the 99% confidence interval is equal to 41.859762 - 1.773846 and 41.859762 + 1.773846

Thus, the 99% confidence interval for the dependent variable SALES, when the value of the independent variable CALLS is 160 is equal to (40.08592, 43.63361)

This means that with 99% confidence, the mean value of the dependent variable SALES lies between 40.08592 and 43.63361 when the value of the variable CALLS is 160.

9.

The prediction interval is given by the formula

$$\hat{y} \pm t_{n-2}^* s_y \sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{(n-1)s_x^2}}$$

On substituting the values, the 99% prediction interval is equal to 41.859762 - 17.74753 and 41.859762 + 17.74753

Thus, the 99% prediction interval for the dependent variable SALES, when the value of the independent variable CALLS is 160 is equal to (24.11223, 59.60729)

A prediction interval is an estimate of an interval in which future observations will fall, with a certain probability, given what has already been observed.

This means that the actual value of the SALES will lie between 24.11223 and 59.60729, when the value of the variable CALLS is 160, with 99% probability.

10.

We can extrapolate the regression line to predict the value of the dependent variable for the values of the independent variable that are outside the range of the sample values. But that requires strong assumptions and is thus generally avoided as it is often inappropriate and may yield incredible results.