{LINEAR REGRESSION}

What is Regression? 1

A technique for determining the **statistical relationship** between two or more variables where a change in a dependent variable (DV) is associated with, and depends on, a change in one or more independent variables (IV).

Simple **L**inear **R**egression is used when we have, one independent variable and one dependent variable.

01

02

Multiple **L**inear **R**egression is used when we have more than one independent variable and one dependent variable.

What is Regression? 2

A technique for determining the **statistical relationship** between two or more variables where a change in a dependent variable(DV) is associated with, and depends on, a change in one or more independent variables(IV).

Since it is a statistical relationship therefore understanding of

- Hypothesis testing
- Statistical tests (t test/ ANOVA)
- Significance (alpha)
- p value
- Standard Error & Confidence Interval
- R^2 and Adjusted R^2

→ Required

Types of Linear Regression

01

SIMPLE LINEAR REGRESSION

has **one** independent and one dependent variable

sales is a function of price of the product

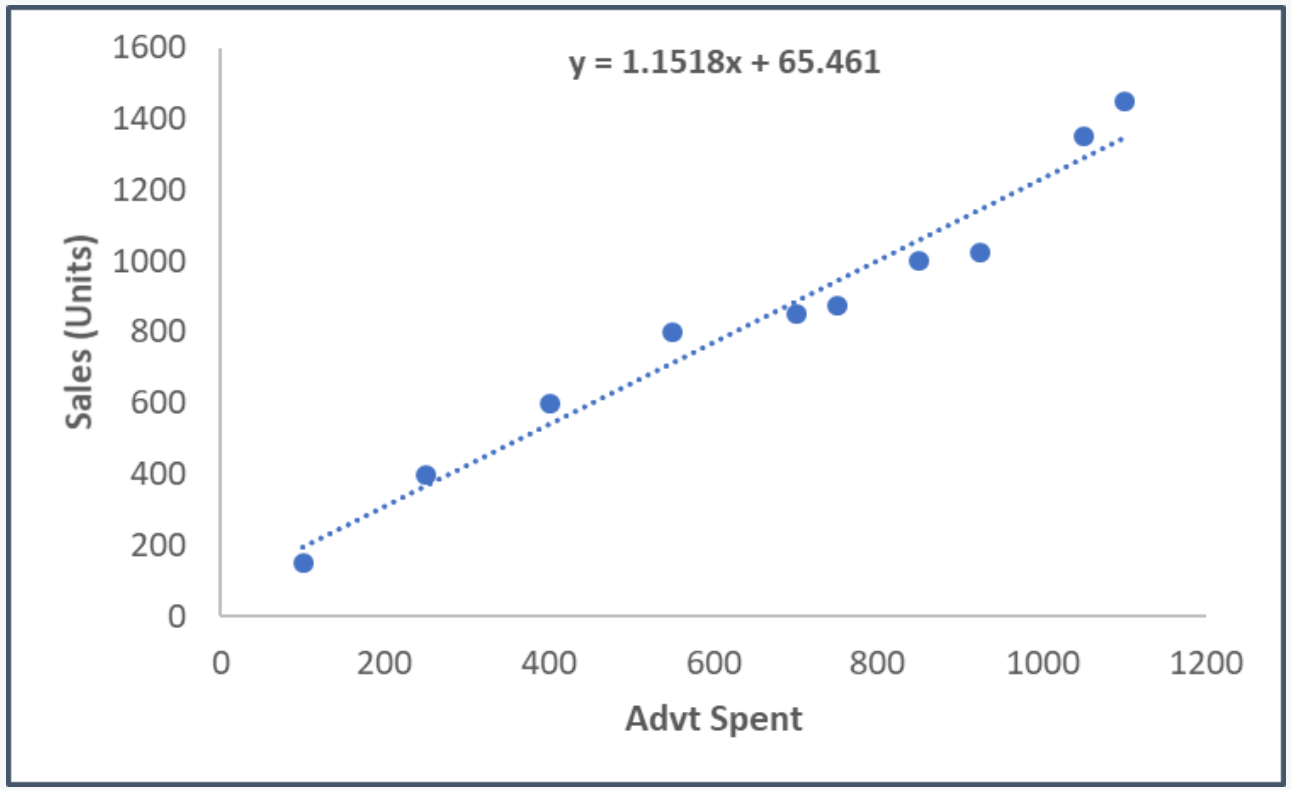
02

MULTIPLE LINEAR REGRESSION

has **multiple** independent variables and one dependent variable

sales is a function of price and advt budget of the product

Linear Regression



Uses a best-fit straight line (regression line) constructed using the **Ordinary Least Squares method (OLS)**

What is “Linear” about “Linear Regression”?

Linearity is the property of a mathematical relationship or function.

Graphically represented as a straight line.

Mean of the dependent variable is a linear combination of the parameters (regression coefficients) and the predictor variables.

Linear Regression Equation (Population) 1

β_0 – y-intercept or the constant

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 .... \beta_n X_n + \varepsilon$$

Y – dependent variable

Linear Regression Equation (Population) 2

X_1, X_2, X_n – independent variables

$\beta_1, \beta_2, \beta_n$ – Slope of regression line or regression co-efficients

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 .... \beta_n X_n + \varepsilon$$

$\beta_1 X_1$ = effect of an *independent variable* on the expected value of *dependent variable*

Linear Regression Equation (Population) 3

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \ldots \beta_n X_n + \varepsilon$$

ε – Error term i.e. unexplained variation in the dependent variable

Linear Regression Equation (Population) 4

β_0 – y-intercept or the constant

X_1, X_2, X_n – independent variables

$\beta_1, \beta_2, \beta_n$ – Slope of regression line or regression co-efficients

$\beta_1 X$ = effect of an *independent variable* on the expected value of *dependent variable*

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \ldots \beta_n X_n + \varepsilon$$

Y – dependent variable

ε – Error term i.e. unexplained variation in the dependent variable

How Well a Regression Line Fits The Data

Accuracy of prediction ….. R squared.

Goodness of Fit

Goodness of Fit describes how well a model fits a set of observations.

01

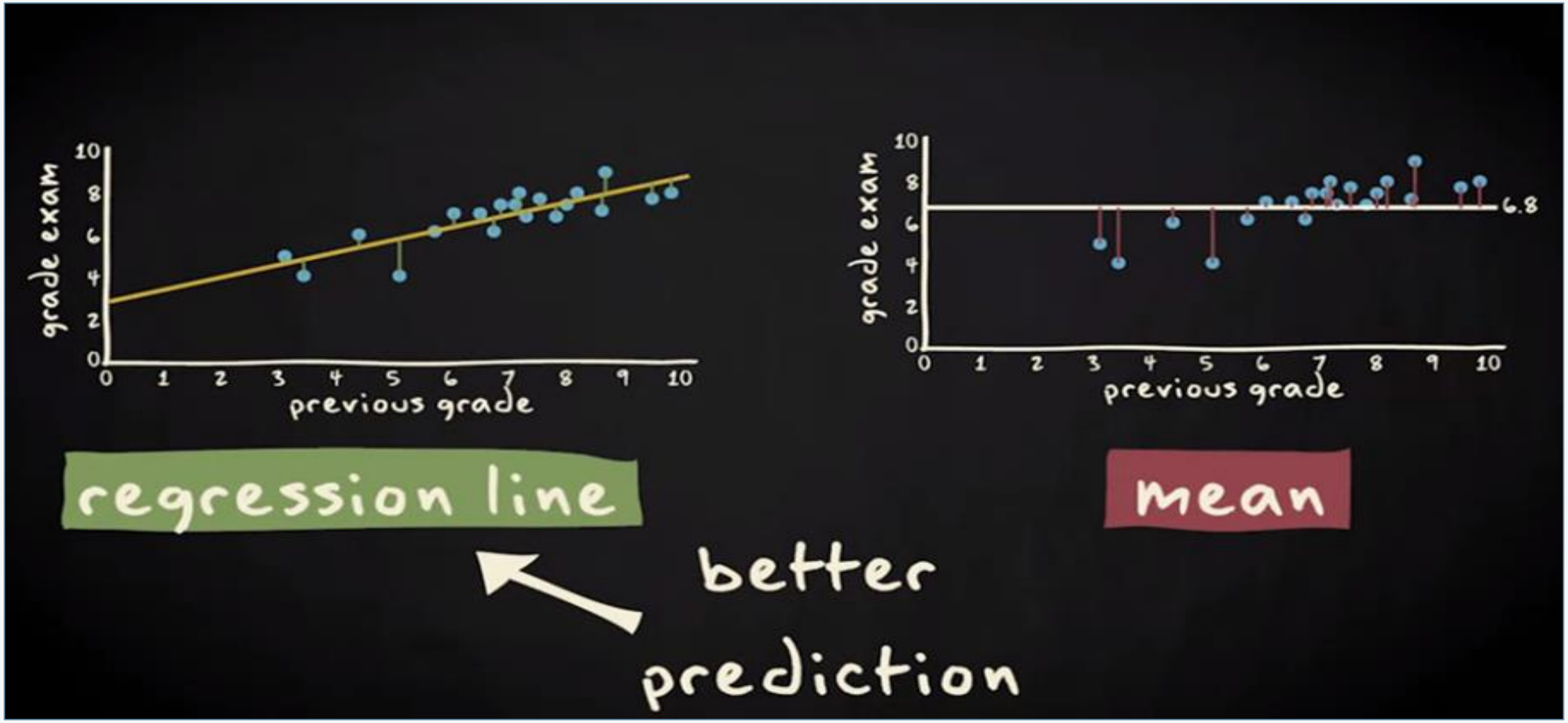
02

If model fits observations well, more accurate will be the predictions.

Measures of goodness of fit summarize discrepancy between observed values and predicted values from the model.

03

Over Mean


grade exam
previous grade
regression line
mean
6.8
better
prediction

The R^2 value

Co-efficient of Determination.

Identify how good the model is in predicting the dependent variable.

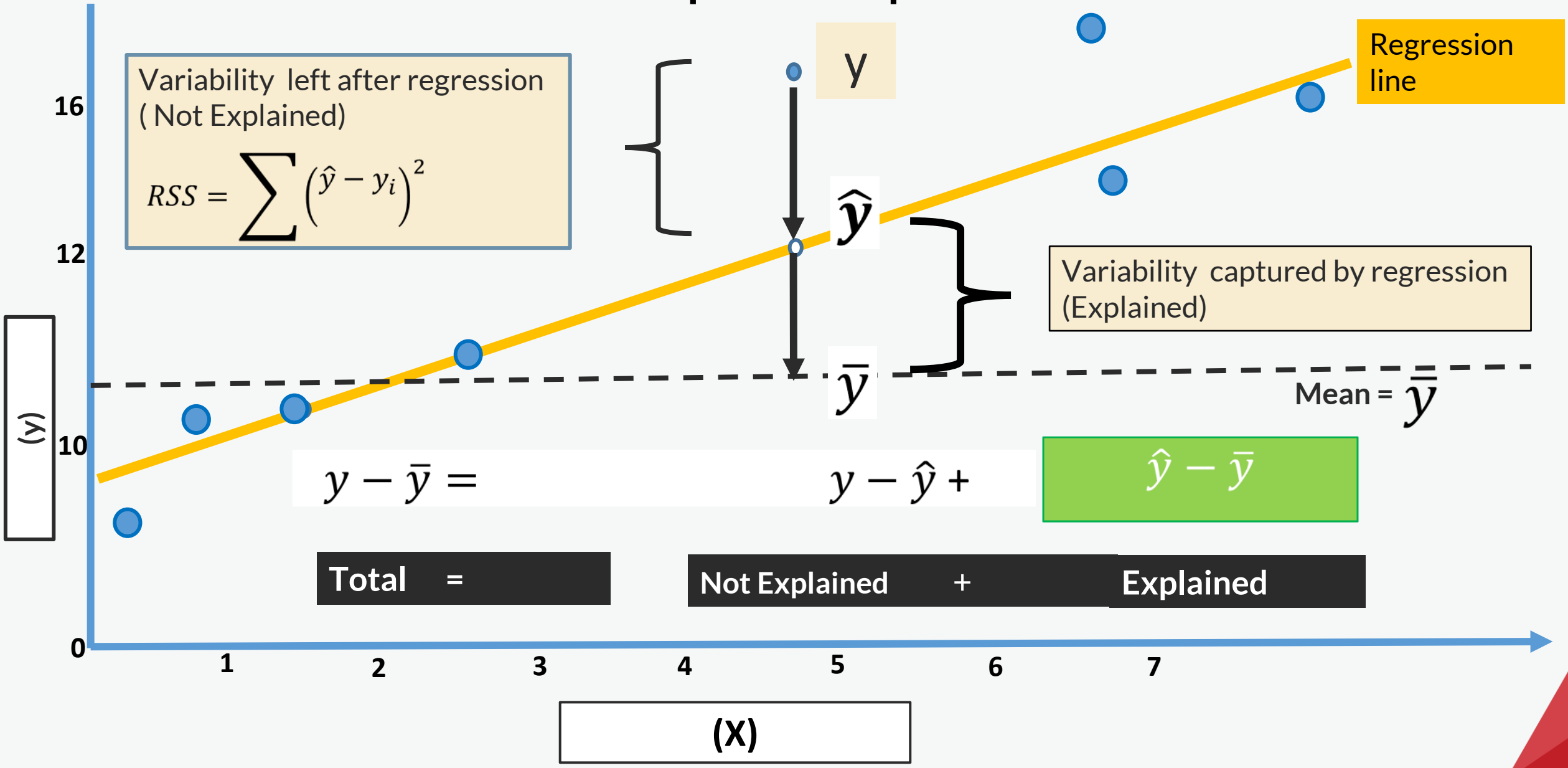
How well model fits the data.

Measures strength of relationship between the model and dependent variable.

Percentage of the variance in the dependent variable that the independent variables explain collectively.

Larger the R^2, better regression model fits data.

Concept of R –squared



Therefore

R-Squared = Proportion of Variance Explained.

R –Squared = (TSS –RSS)/TSS.

R –Squared = 1- (Unexplained/Total) = Explained /Total.

R^2= Explained / Total

Calculating R^2 value

$$R^2 = \frac{SSR}{SST} = \frac{SSR}{SSR + SSE}$$

SSE should be low for a high R^2

R^2 – what % of SST can be explained by the model

R^2 = 384831/394000

R^2 = **97.67%**

97.7% of variation in Profit (y) is explained by Sales (x)

R-Squared Ranges From 0 to 100%

R^2= Explained / Total

When Explained = Total i.e. Regression is able to capture the total variability

- R-Squared = 100%

When Explained = 0 i.e. Regression is not able to capture the total variability

- R-Squared = 0%

Therefore

While r is the direction and strength of the relationship.

R^2 conveys

1. How much better a regression line predicts the dependent variable than the mean of that variable.

2. How much of the variance in the dependent variable is explained by the independent variables.

Interpreting R^2 in 2 ways

$R^2 = 0.69$

1. Prediction error is 69 % smaller than when mean is used.

2. The amount of variance in the dependent variable (y) that is explained by the independent variables (X).

The OLS Method

- Ordinary Least Squares (OLS) is a linear least squares method.

- Estimates the unknown parameters (b_0 and b_1) in a model.

How does OLS method work?

1. Applies the principal of least squares.

2. Minimizes **sum of the squares of the differences** between observed variable and those predicted by the linear function.

3. The smaller the differences, the better the model.

In the Add-Ins available box, select the Solver Add-in check box, and then click OK

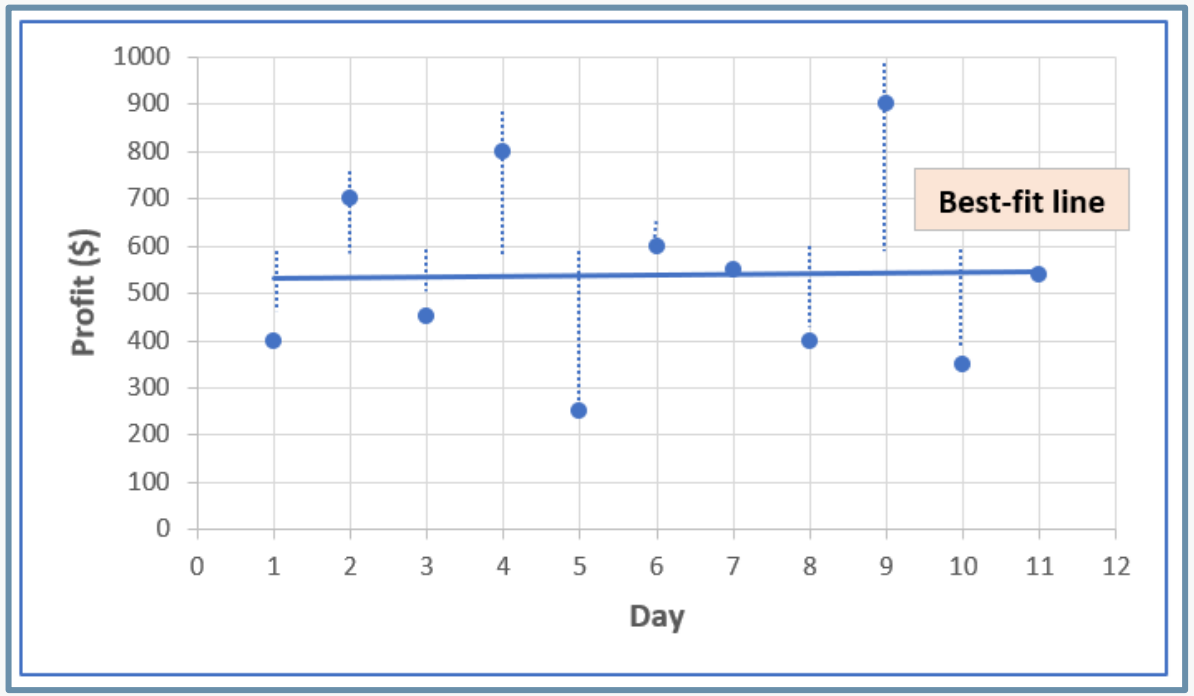
The OLS method

X & Y variables		Predicted Y	Obs - pred	(Obs – pred)^2
X	Y	Y'	Y-Y'	(Y-Y')^2
1.00	1.00	1.210	-0.210	0.044
2.00	2.00	1.635	0.365	0.133
3.00	1.30	2.060	-0.760	0.578
4.00	3.75	2.485	1.265	1.600
5.00	2.25	2.910	-0.660	0.436
		SUM	**0**	**2.791**

Sum of Squares of differences $\sum(Y - Y')^2 = \mathbf{2.791}$

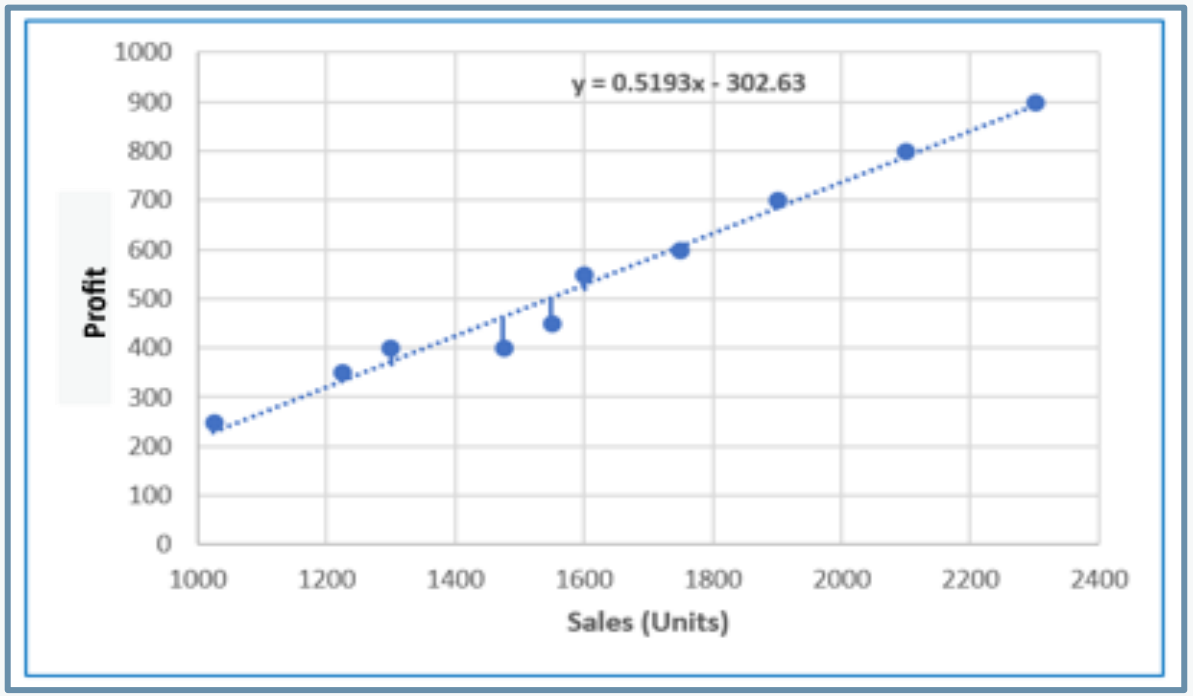
The Regression Line

INTERCEPT ONLY MODEL



Sum of Residual Squares (SSE) = 394000

MODEL WITH ONE X VARIABLE



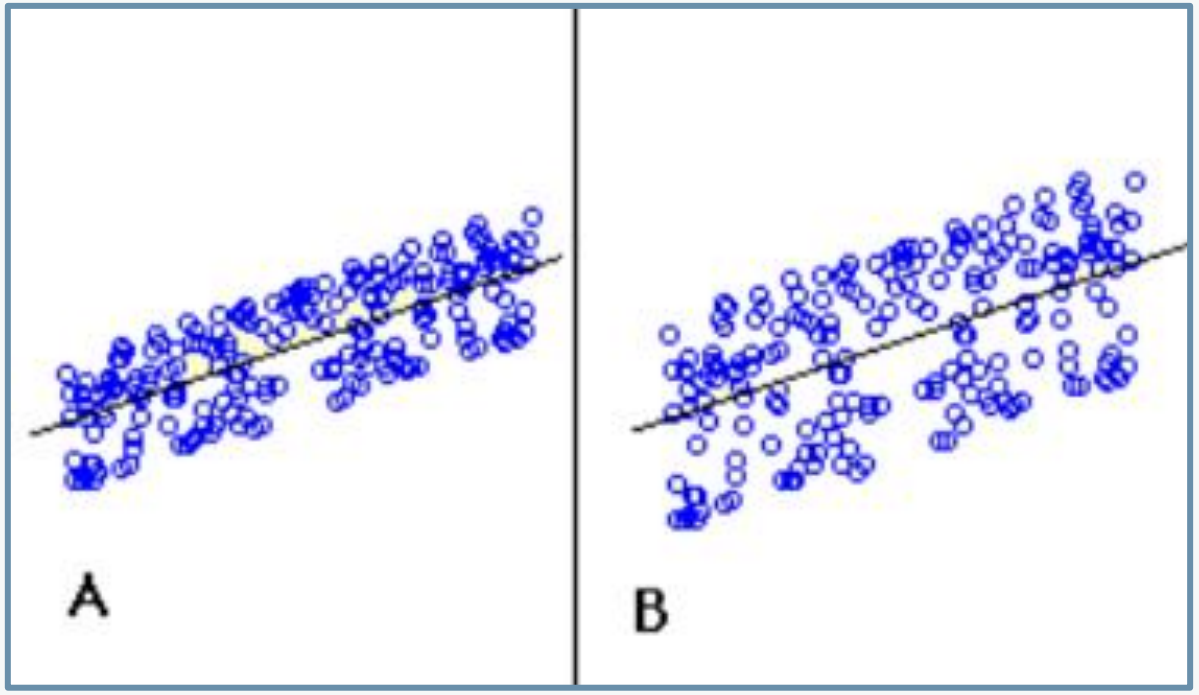
Sum of Residual Squares (SSE) = 9169

SSE has decreased significantly after the independent variable (Sales) was introduced, thereby producing a better regression model

Measures of predictive accuracy

Which model is more accurate?

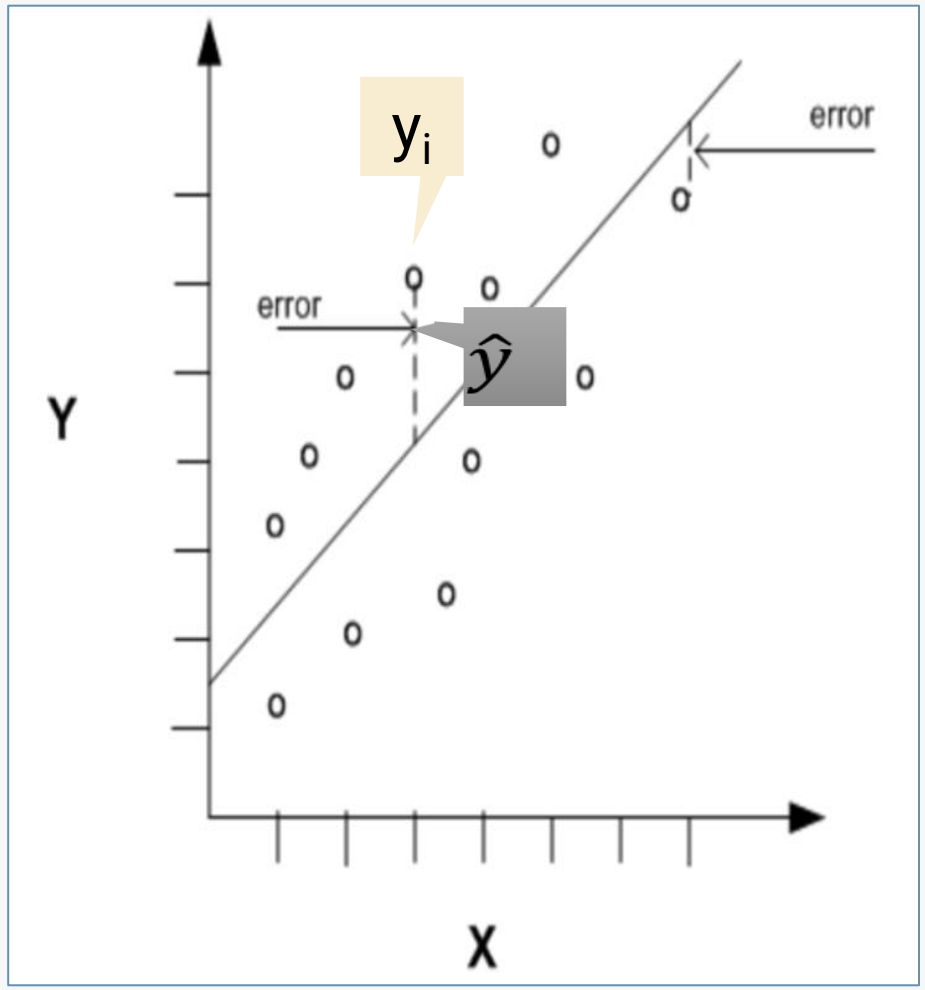
The actual values are closer to the regression line in Model A.



Measures of Predictive Accuracy

- Mean absolute error (MAE)
- Root Mean Square Error (RMSE)
- Mean Absolute Percentage Error (MAPE)

Mean Absolute Error (MAE)



Averages sum of absolute differences (error) between observed and predicted values

$$MAE = \frac{1}{n}\sum |y - \hat{y}|$$

For average

Absolute value of the error

Less prone to the effect of outliers

Calculating MAE

Day	Sales x_i	Profit y_i	$\hat{y}$	**Abs_error**
1	1300	400	372.51	27.5
2	1900	700	684.12	15.9
3	1550	450	502.35	52.3
4	2100	800	787.98	12.0
5	1025	250	229.69	20.3
6	1750	600	606.22	6.2
7	1600	550	528.31	21.7
8	1475	400	463.40	63.4
9	2300	900	891.85	8.1
10	1225	350	333.56	16.4
Average				**24.4**

$$absolute\ error = abs(y - \hat{y})$$

Root Mean Square Error

MSE: Square of the error before averaging the residual errors

RMSE = $\sqrt{MSE}$

Lower RMSE Score ⇒ a better fit

More weightage to larger errors- sensitive to outliers

$$MSE = \frac{1}{n}\sum (y - \hat{y})^2$$

Square of error

Calculating RMSE

Day	Sales x_i	Profit y_i		Error^2
1	1300	400	372.51	756
2	1900	700	684.12	252
3	1550	450	502.35	2740
4	2100	800	787.98	144
5	1025	250	229.69	412
6	1750	600	606.22	39
7	1600	550	528.31	470
8	1475	400	463.40	4019
9	2300	900	891.85	66
10	1225	350	333.56	270
Average				**917**

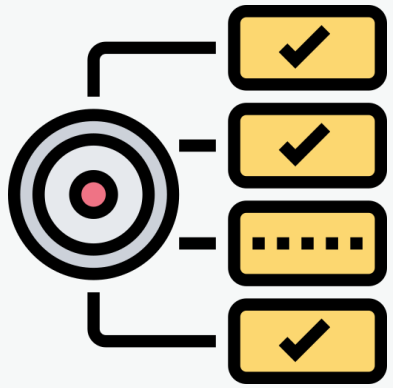
MSE = 917

RMSE = √917 = 30.3

Mean Absolute Percentage Error (MAPE)

MAPE measures how far the model's predictions are off from their corresponding outputs on average

$$MAPE = \frac{100\%}{n} \sum \left| \frac{y - \hat{y}}{y} \right|$$



Mean Absolute Percentage Error (MAPE)

Clear interpretation since percentages are easier to conceptualize

MAPE Value	Prediction Accuracy
MAPE ≤ 10%	High
10% < MAPE ≤ 20%	Good
20% < MAPE ≤ 50%	Reasonable
MAPE > 50%	Low

Calculating MAPE

Day	Sales x_i	Profit y_i	$\hat{y}$	Abs_error	Abs_error / Y
1	1300	400	372.51	27.5	0.07
2	1900	700	684.12	15.9	0.02
3	1550	450	502.35	52.3	0.12
4	2100	800	787.98	12.0	0.02
5	1025	250	229.69	20.3	0.08
6	1750	600	606.22	6.2	0.01
7	1600	550	528.31	21.7	0.04
8	1475	400	463.40	63.4	0.16
9	2300	900	891.85	8.1	0.01
10	1225	350	333.56	16.4	0.05
Average					**0.06**

MAPE = 0.06 X 100 = 6%

Calculating SST, SSR, SSE

Day	Sales x_i	Profit y_i	Ypred			
1	1300	400	372.5	19600	28052	756
2	1900	700	684.1	25600	20770	252
3	1550	450	502.3	8100	1418	2740
4	2100	800	788.0	67600	61496	144
5	1025	250	229.7	84100	96289	412
6	1750	600	606.2	3600	4385	39
7	1600	550	528.3	100	137	470
8	1475	400	463.4	19600	5868	4019
9	2300	900	891.9	129600	123800	66
10	1225	350	333.6	36100	42616	270
SUM				**394000**	**384831**	**9169**
mean	1622.5	540		SST	SSR	SSE

Sum of Squares Total
SST = $\sum(y_i - \bar{y})^2 = 394000$

Sum of Squares Regression
SSR = $\sum(\hat{y}_i - \bar{y})^2 = 384831$

Sum of squares Error
SSE = $\sum(y_i - \hat{y}_i)^2 = 9169$

SSR + SSE = SST
384831 + 9169 = 394000

In sum

Simple linear regression is essentially **a comparison of two** models.

1. One where independent variable does not even exist.

2. Other uses the best fit regression line.

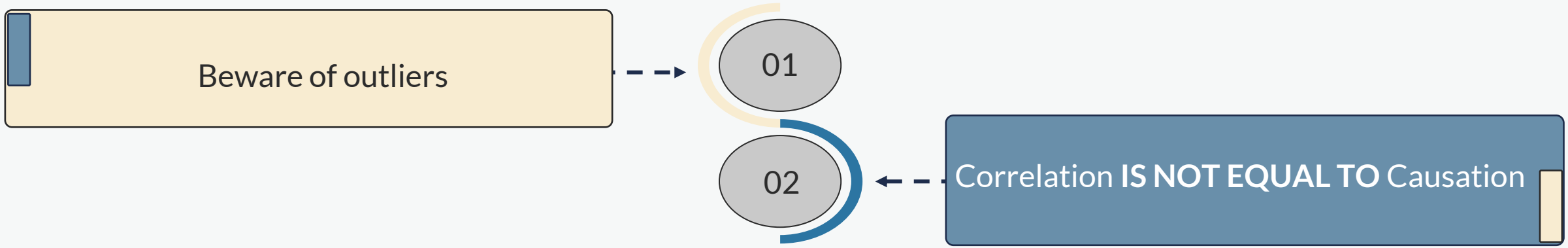
If there is only one variable, then the best prediction for the other values is the mean of the dependent variable.

The difference between the best fit line and the observed values is called the residual (error).

Residuals are squared and added together to generate sum of squares (SSE).

Simple linear regression is designed to find the best fitting line through the data that minimizes SSE.

Interpreting regression results



{THANK YOU}