Quantitative Methods

Lecture-13

BITS Pilani

Pilani Campus

BITS Pilani
Pilani Campus

Simple Linear Regression

(Ch 12 Business Statistics, Levine et al.)

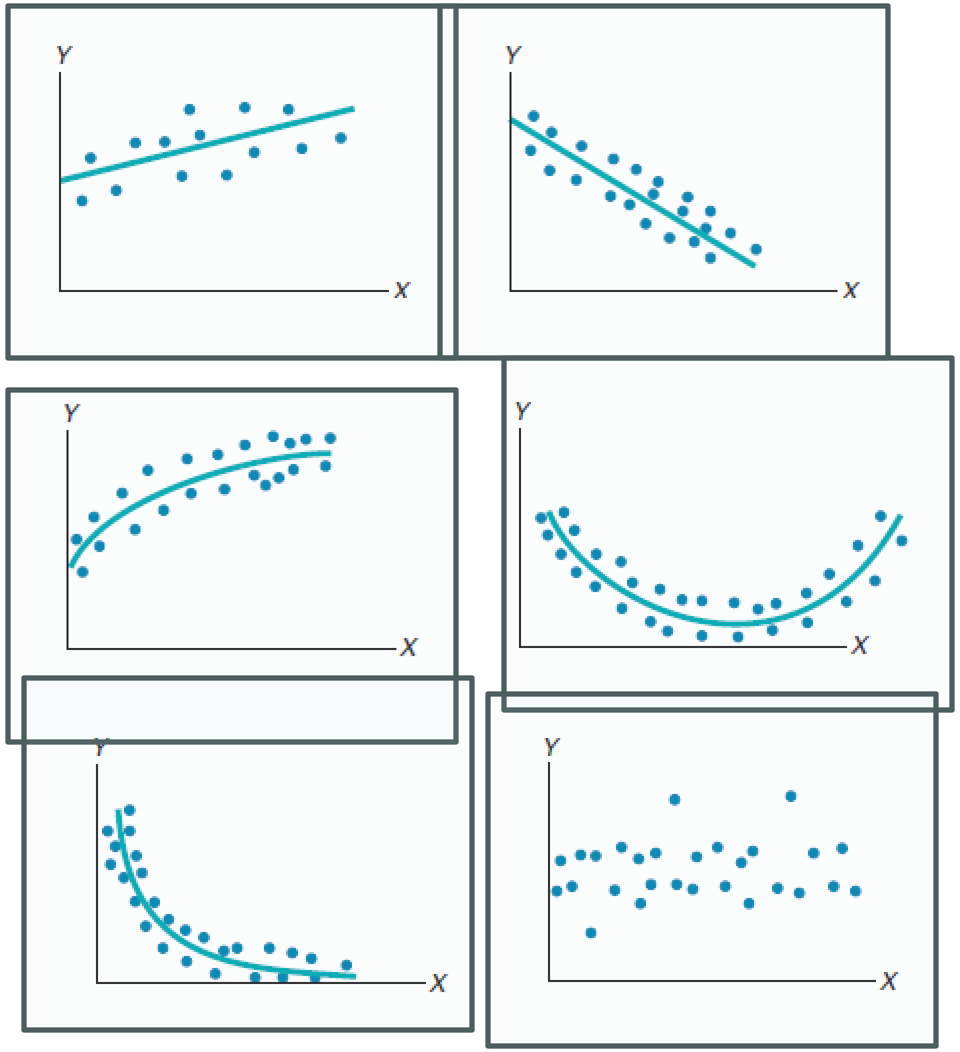
Types of relationships

Linear relationships

- Positive Linear Relationship: When X increases, Y increases
- Negative Linear Relationship: When X increases, Y decreases

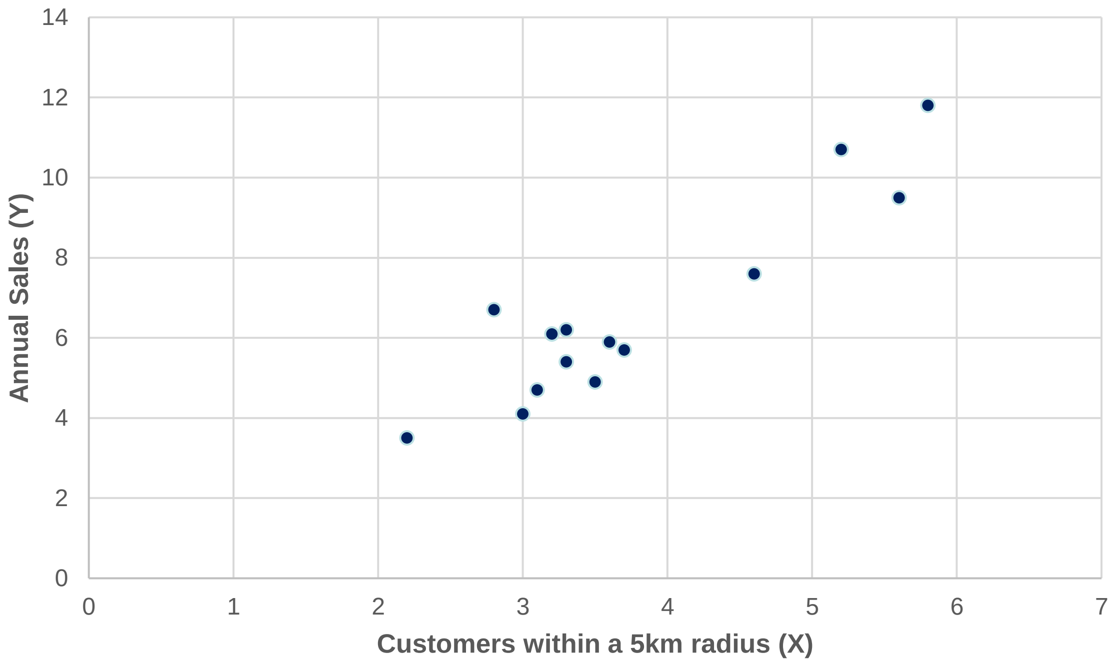
Non-linear relationships and No relationships

- Positive Curvilinear Relationships
- Negative Curvilinear Relationships
- U shaped Curvilinear Relationships
- No Relationships



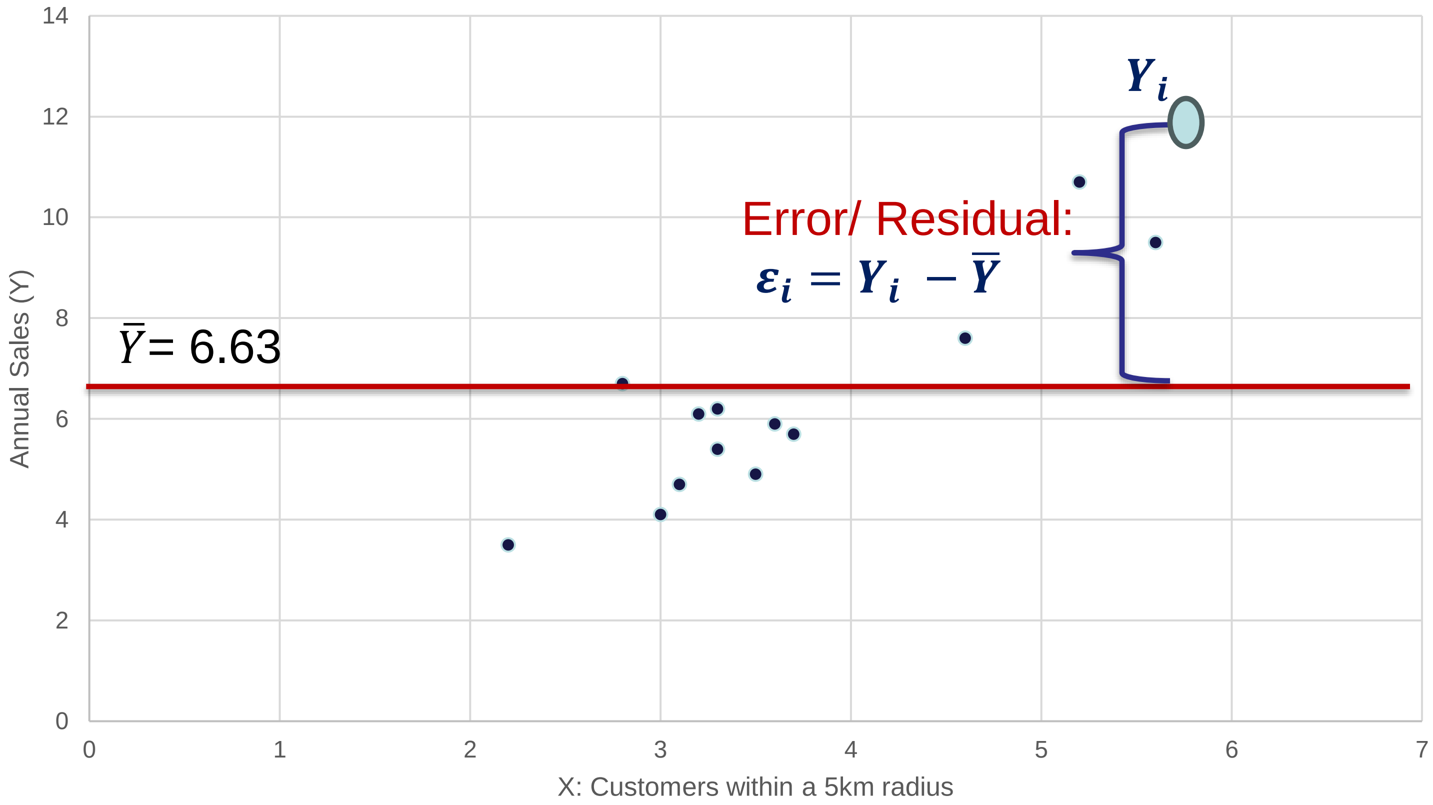
Sample data and the scatter plot

Store	Customers (Lakhs)	Annual Sales (Lakhs)
1	3.7	5.7
2	3.6	5.9
3	2.8	6.7
4	5.6	9.5
5	3.3	5.4
6	2.2	3.5
7	3.3	6.2
8	3.1	4.7
9	3.2	6.1
10	3.5	4.9
11	5.2	10.7
12	4.6	7.6
13	5.8	11.8
14	3	4.1



Is there a relationship between X and Y?

Predictions: Point Estimate

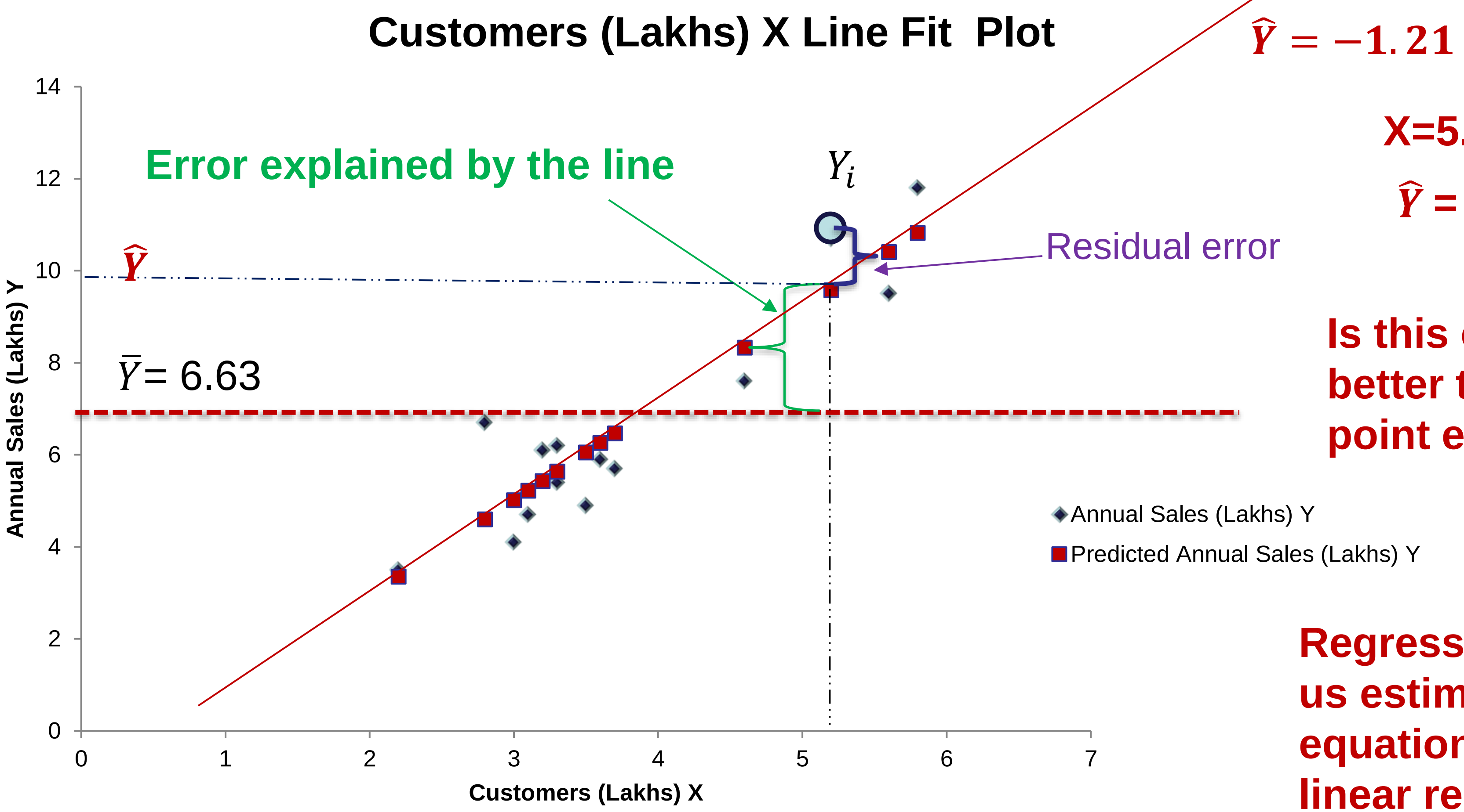


What is the point estimate of Y, without taking the relationship between X and Y into account?

Mean is the point estimate without taking the relationship into account.

Is every value of Y equal to $\bar{Y}$?

Taking the relationship into account: Fitting a line.



$\hat{Y} = -1.21 + 2.07X$

X=5.3

$\hat{Y}$ **= 9.8**

Is this estimate better than the point estimate?

Regression helps us estimate the equation of the linear relationship.

Simple Linear Regression

- Two variables (X and Y).
- They are assumed to have a linear relationship (increasing or decreasing).
- Y is our variable of business interest.
- We want to predict the value of Y, given certain value of X.
- X, is called an independent variable. Its values is determined outside the system (Exogenous).
- Y, is called dependent variable. Sometimes also referred as outcome or response variable.
- Value of Y is determined within the system (Endogenous).
- When independent variable X changes, Y also changes in a predictable way.

Linear relationship: equation of a line

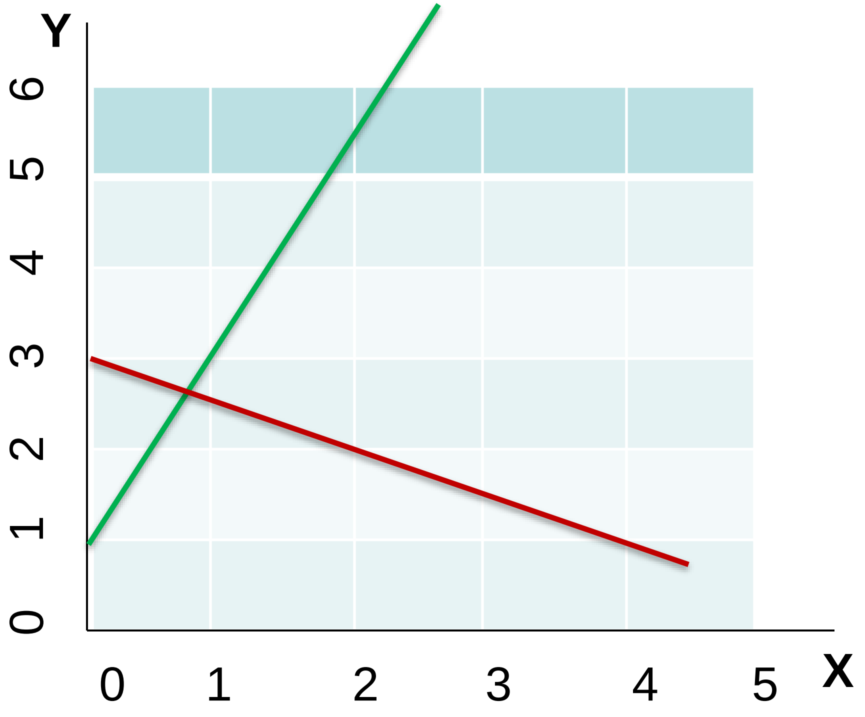
- Independent variable (X) is shown on X axis.
- Dependent variable is (Y) is shown on Y axis.
- Equation of a line: $Y = \beta_0 + \beta_1 X$

<u>Intercept (β_0)</u>

- The point where the line meets Y axis.
- Value of Y, when X is 0

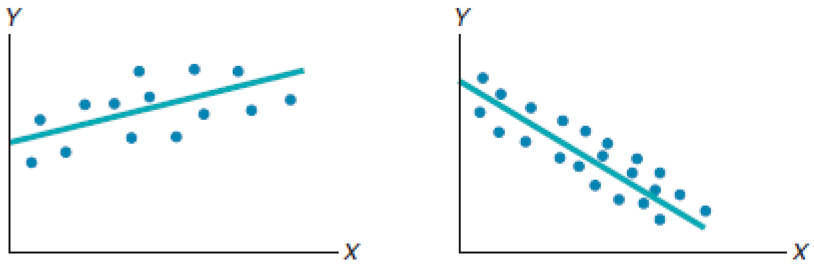
<u>Slope (β_1)</u>

- When value of X goes up by 1 unit, the value of Y goes up by β_1 units.
- **Green line:** Y = 1 + 2X (intercept? Slope?)
- $Intercept\text{: } \beta_0 = 1, Slope\text{: } \beta_1$ = 2
- **Red line:** Y = 3 - 0.5X (Slope β_1 is negative: -0.5)



Simple Linear Regression Model

- Population regression equation.
- $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$
- β_0, β_1 *are intercept and slope population* ***parameters***
- ε_i: Random Error
- *Expected value of Y is the line.*
- $E[Y] = \beta_0 + \beta_1 X$
- Line is the best average fit.



Fitting a regression line from sample data

- We fit a line based on the sample data. If certain assumptions are met, this line can be used to make population predictions.

SIMPLE LINEAR REGRESSION EQUATION: THE PREDICTION LINE

The predicted value of Y equals the Y intercept plus the slope multiplied by the value of X.

$$\hat{Y}_i = b_0 + b_1 X_i$$

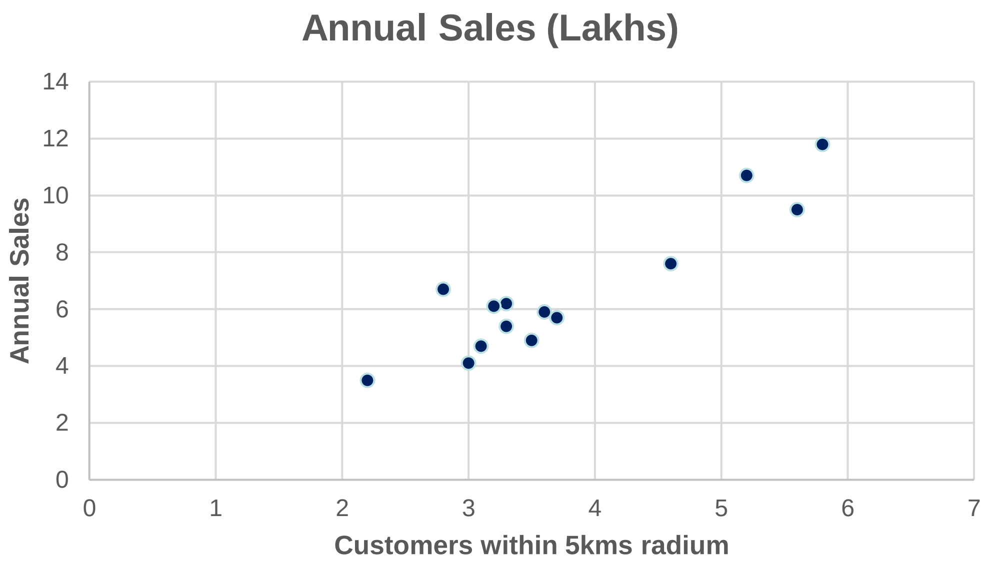
where

$\hat{Y}_i$ = predicted value of Y for observation i

X_i = value of X for observation i

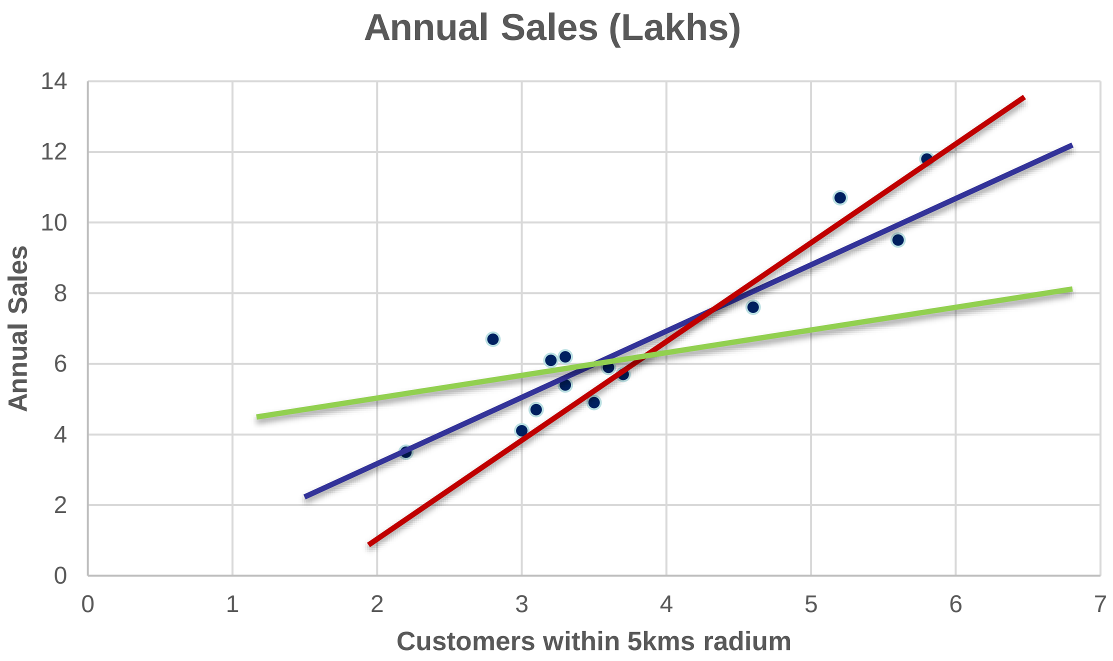
b_0 = sample Y intercept

b_1 = sample slope



- Notice the “hat” on top of the dependent variable. “Hat” represents estimated values of Y.

How to fit a line? Which of the lines fits the best?



- <u>Least Square Method</u>
- The line that minimizes the sum of squared <u>residual errors</u>, is the best fit.
- min $\sum_{All\ values}(Y_i - \widehat{Y}i)^2$
- min $\sum_{All\ values}(Y_i - b_0 - b_1 X_i)^2$
- Solving it gives the formula for the intercept and the slope estimates (b_0, b_1).

Linear Regression: The coefficients

Solving the square errors for minimization, we get

- $b_1 = \frac{SSXY}{SSX}$
- SSXY= $\sum_i^n (X_i - \bar{X})(Y_i - \bar{Y}) = \sum_i^n (X_i Y_i) - \frac{(\sum_{i=1}^n X_i)(\sum_{i=1}^n Y_i)}{n}$
- SSX = $\sum_i^n (X_i - \bar{X})^2 = \sum_{i=0}^n (X_i)^2 - \frac{(\sum_{i=0}^n X_i)^2}{n}$
- $b_0 = \bar{Y} - b_1 \bar{X}$
- $\bar{Y} = \frac{\sum_i^n (Y_i)}{n}$
- $\bar{X} = \frac{\sum_i^n (X_i)}{n}$

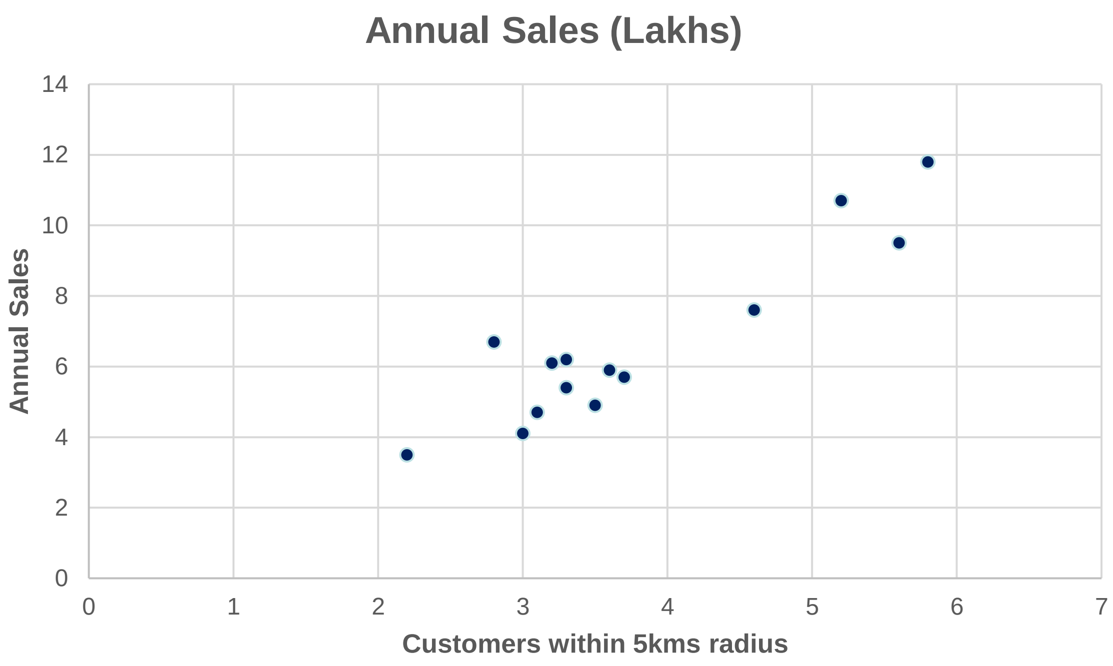
X_i	Y_i	$(X_i - \bar{X})$	$(Y_i - \bar{Y})$	$(X_i - \bar{X})^2$

Linear Regression Example

- You are the CEO of the Coconut water branded outlet business. You would like to get a strategy to identify where to open new stores.
- From your experience you find that your sales directly depend on number of potential customers within 5 sq km radius of the stores.
- You can find number of potential customers within 5 sq km radius by using a market research firm.
- Yow would like to build a linear regression model to be able to predict potential sales.
- Linear Model: An equation of the line that can help you predict the dependent variable.

Sample data

Store	Customers (Lakhs)	Annual Sales (Lakhs)
1	3.7	5.7
2	3.6	5.9
3	2.8	6.7
4	5.6	9.5
5	3.3	5.4
6	2.2	3.5
7	3.3	6.2
8	3.1	4.7
9	3.2	6.1
10	3.5	4.9
11	5.2	10.7
12	4.6	7.6
13	5.8	11.8
14	3	4.1



Linear Regression: Working through a business problem

Store	Customers (Lakhs) X	Annual Sales (Lakhs) Y	X-Xbar	(X-Xbar)^2	Y-Ybar	(X-Xbar)(Y-Ybar)
1	3.7	5.7	-0.08	0.0062	-0.93	0.073
2	3.6	5.9	-0.18	0.0319	-0.73	0.130
3	2.8	6.7	-0.98	0.9576	0.07	-0.070
4	5.6	9.5	1.82	3.3176	2.87	5.230
5	3.3	5.4	-0.48	0.2290	-1.23	0.588
6	2.2	3.5	-1.58	2.4919	-3.13	4.939
7	3.3	6.2	-0.48	0.2290	-0.43	0.205
8	3.1	4.7	-0.68	0.4605	-1.93	1.309
9	3.2	6.1	-0.58	0.3347	-0.53	0.306
10	3.5	4.9	-0.28	0.0776	-1.73	0.482
11	5.2	10.7	1.42	2.0205	4.07	5.787
12	4.6	7.6	0.82	0.6747	0.97	0.798
13	5.8	11.8	2.02	4.0862	5.17	10.454
14	3	4.1	-0.78	0.6062	-2.53	1.969
Mean	3.78	6.63	SSX	15.5236	SSXY	32.199
			b1	**2.0742**		
			b0	**-1.2089**		

Estimated regression line: $\hat{Y} = $ **−1.2089 + 2.0742*X**

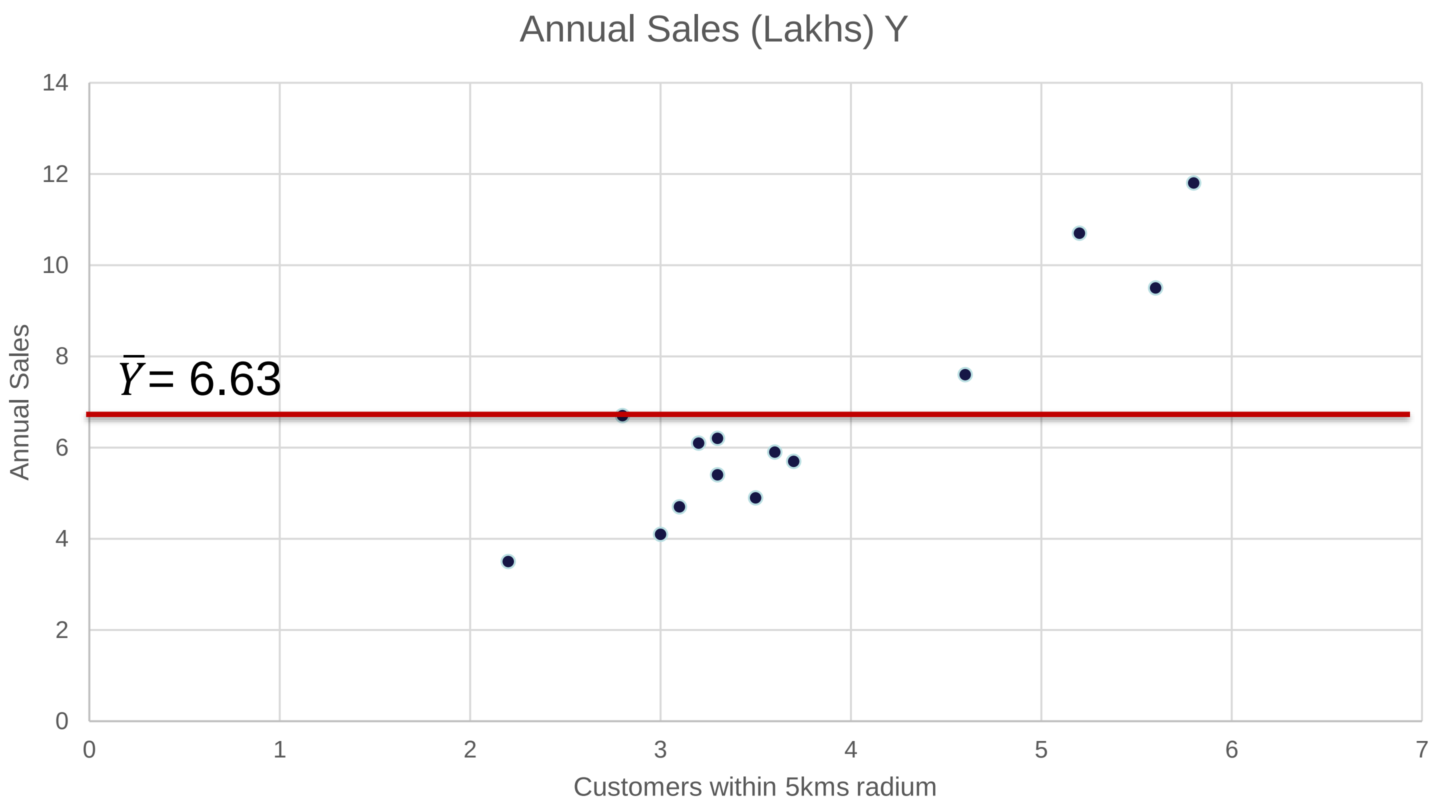
Linear Regression: Predictions and Cautions

- Regression Equation: $\hat{Y} = -1.21 + 2.1 * X$
- X: Number of customers within 5kms radius and $\hat{Y}$: Predicated sales.
- Your MR firm calculated potential customers within 5kms rage to be 4 Lakhs.
- What is your predicted annual sales?
- $\hat{Y} = -1.21 + 2.1*4 = 7.2$ Lakh INR.

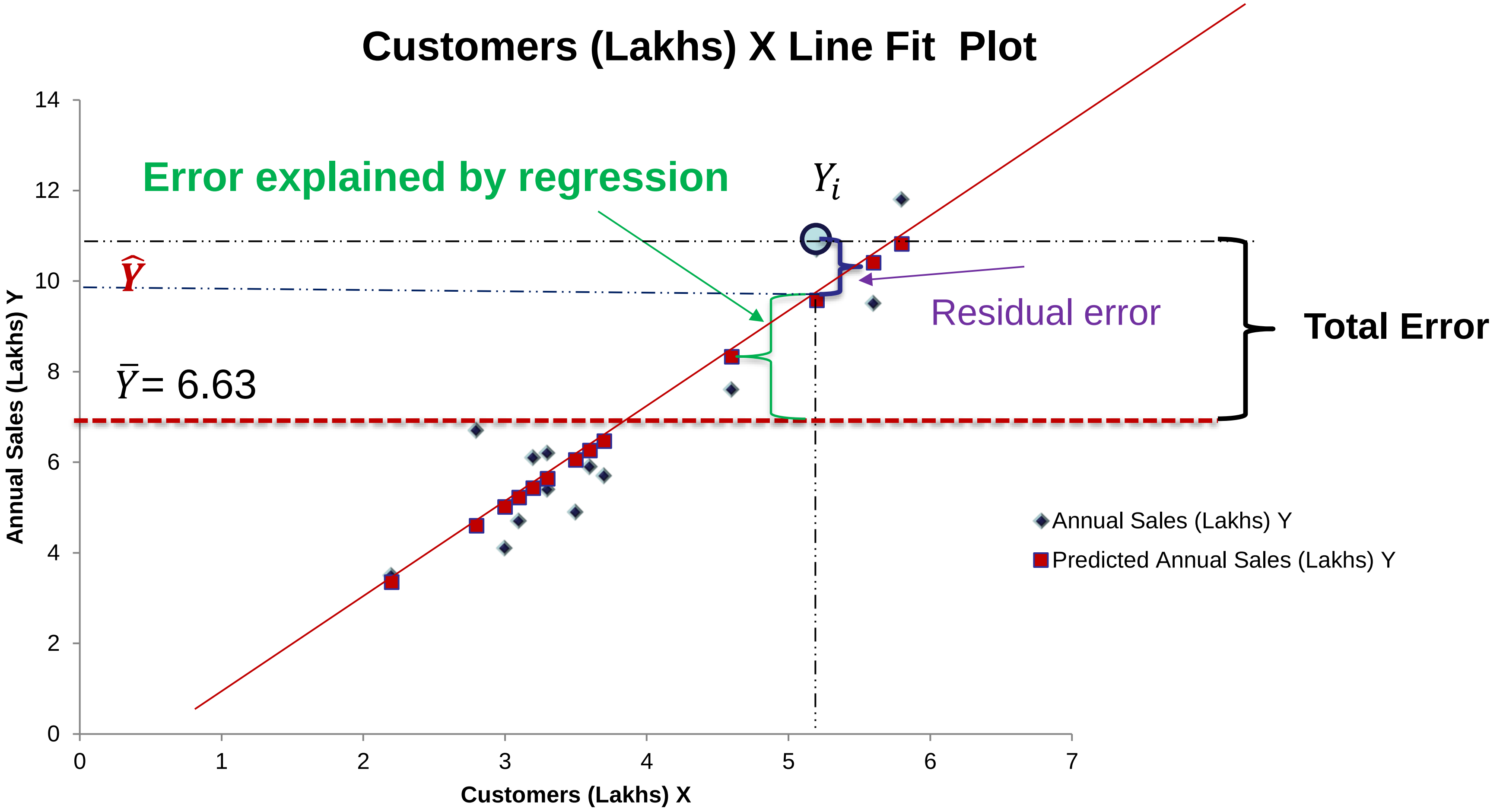
<u>Caution:</u>

- Interpolation v/s Extrapolation.
- You have used a range of X values, from your sample, to estimate the regression equation (2.2 – 5.8 Lakh)
- Predictions may be invalid out of these range of values. We should not use regression for <u>extrapolation.</u>
- Predictions within the range are called <u>interpolations</u>.

Measures of Variation: Prediction without additional info.

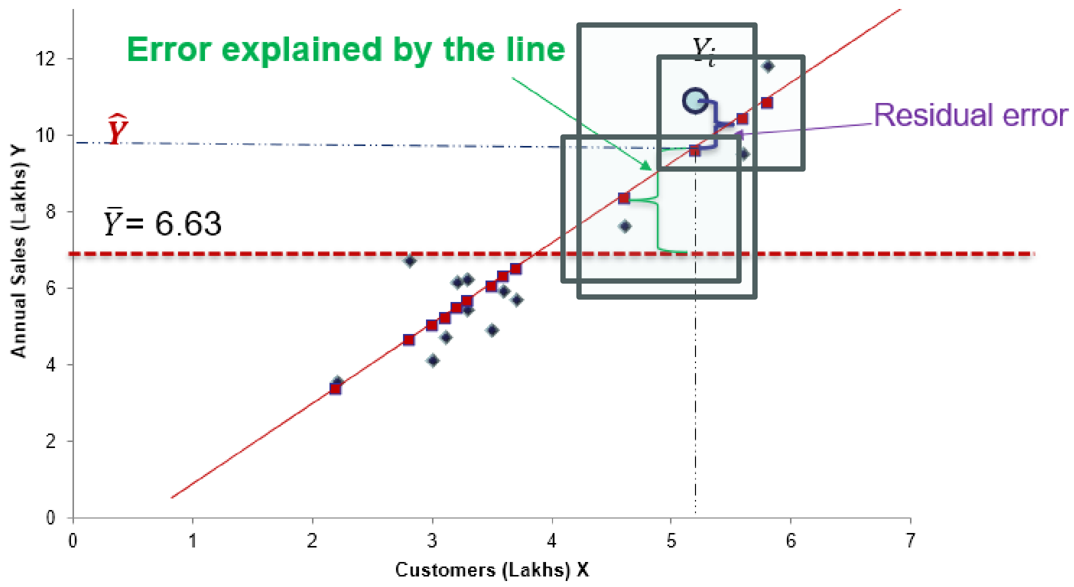


Taking the relationship into account: Fitting a line.



Measures of Variation: SST = SSR + SSE

- Total Sum Of Squares Variation (SST): Measure of variation of Yi around the mean
- Total Variation = Explained by the regression + Residual Variation
- $SST = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$
- Regression Sum Of Squares Variation (SSR): Variation in Y explained by the regression on Variable X.
- $SSR = \sum_{i=1}^{n}(\hat{Y} - \bar{Y})^2$
- Error Sum of Squares (SSE): Variation in Y not explained by X (Due to other factors).
- $SSE = \sum_{i=1}^{n}(Y_i - \hat{Y})^2$
- **SST = SSR + SSE**



Linear Regression: The Coefficient of Determination

- Total Variation = Explained by the regression + Residual Variation
- Total Sum Of Squares Variation (SST): Measure of variation of Yi around the mean
- Regression Sum Of Squares Variation (SSR): Variation in Y **explained by the regression** on Variable X.
- Error Sum of Squares (SSE): Variation in Y **not** explained by X (Due to other factors).
- SST = SSR + SSE

What proportion of the total variation is explained by the regression?

- r^2 **= SSR / SST** **;** This is called "The Coefficient of Determination".
- The proportion of variation in the values of Y, explained by the linear relationship between independent variable X with the dependent variable Y.
- **Correlation Coefficient: r** (How do you know if it is +ve or –ve?)
- Depends on the sign of slope *b1*

Linear Regression: Standard Error of the Estimate

- Regression line does not predict values exactly.
- There is residual error.
- Standard error of the estimate of Y
- $S_{XY} = \sqrt{\frac{SSR}{n-2}}$
- SSR: Residual sum of square.
- $SSR = \sum_{i=1}^{n}(\hat{Y} - \bar{Y})^2$
- S_{XY}: Standard error of the estimate.
- The standard deviation measures variation around the mean. S_{XY} measures the variation around the regression line.

Linear Regression: XLSX

- Data >> Enable Data Analysis ToolPak

Inference about the slope: t-test

- Regression line: $\hat{Y} = b_0$+b_1*X
- b_1: Slope of the regression line.
- What does b_1 = 0 mean?.
- There is no linear relationship between X and Y.
- Regression output gives us the t-test results
- H_0: b_1 = 0
- **Is $\boldsymbol{b_o}$ the intercept and $\boldsymbol{b_1}$ the slope significant (null of zero value is rejected) in the following output?**

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	-1.20884	0.994874	-1.21507	0.247707	-3.37648	0.958806	-3.37648	0.958806
Customers (Lakhs) X	2.074173	0.253629	8.177972	3E-06	1.521562	2.626784	1.521562	2.626784

Inference about the slope: F-test

- The test uses the ratio of SSR (Regression sum of squares) and SSE (Error sum of squares) to check the significance of the slope parameter.
- F-statistics follows an F distribution with 1 and n-2, numerator and denominator degrees of freedom resp.

$$F_{STAT} = \frac{MSR}{MSE}$$

$$MSR = \frac{SSR}{1} = SSR$$

$$MSE = \frac{SSE}{n-2}$$

Source	df	Sum of Squares	Mean Square (variance)	F
Regression	1	SSR	$MSR = \frac{SSR}{1} = SSR$	$F_{STAT} = \frac{MSR}{MSE}$
Error	$n-2$	SSE	$MSE = \frac{SSE}{n-2}$	
Total	$n-1$	SST		

Reject H_0 if $F_{STAT} > F_\alpha$;
otherwise, do not reject H_0.

Q&A