

Regression Analysis

Correlation and Regression

- **Scatter Plot** suggests relationship between two variables but does not quantify it.
- **Correlation Analysis:** A technique to quantify the strength of relationship between output and input variable in terms of correlation coefficient r. Example : Calories consumed and weight gained , Advt. expenditure and sales revenue .
- **Regression Analysis:** Statistical analysis technique to identify functional relationship between an input variable x and an output variable y and to predict the output variable using regression equation which can give the value of output for given input value(s). Example: For a known Advt. expenditure we can estimate the sales revenue using the regression equation **Y= a+bx**

Analysis for Data type :
Input X – continuous
Output Y - continuous

Scatter Diagrams

D M A I C

Scatter diagrams or plots provides a graphical representation of the relationship of two continuous variables

How does a Scatter Diagram help?

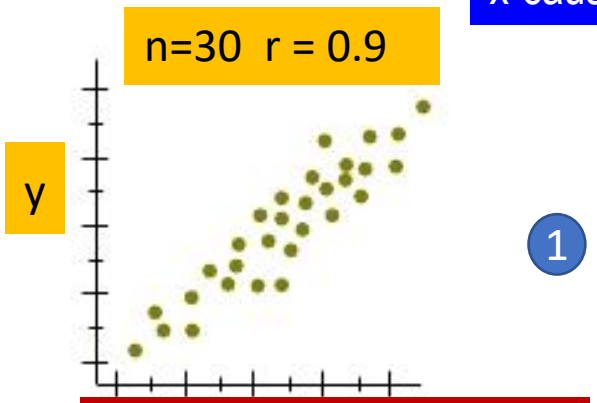
- ✓ Lets you “see” patterns in data
- ✓ Supports or refutes theories about the data and thereby helps create or refine hypotheses
- ✓ Predicts effects under other circumstances (be careful extending predictions beyond the range of data used)
- ✓ **Be Careful-** Correlation does not guarantee causation. Correlation by itself **does not imply a cause and** effect relationship!

How to draw inferences of a scatter diagram?

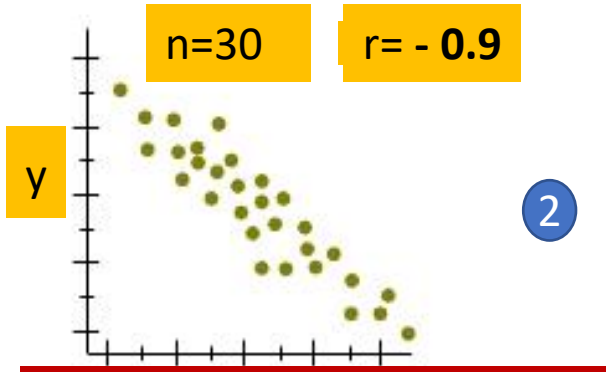
- ✓ Judge strength of relationship by width or tightness of scatter
- ✓ Determine direction of the relationship, e.g. If X increases, and Y decreases, it is negative correlation, similarly if X increases, and Y increases, it is positive correlation

Scatter Diagram Interpretation

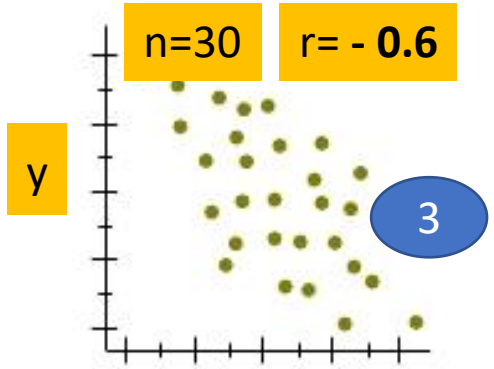
x-cause (Continuous), y – effect (Continuous)



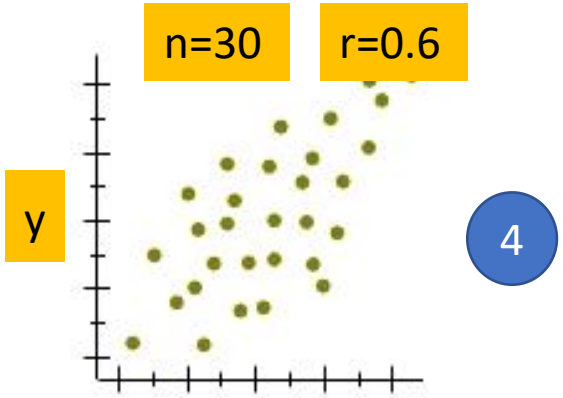
x-cause
Strong Positive Correlation



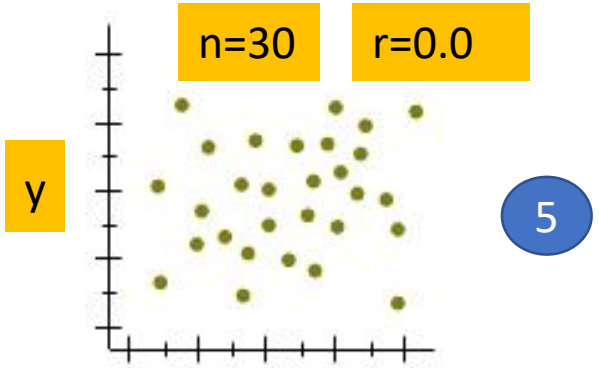
x-cause
Strong Negative Correlation



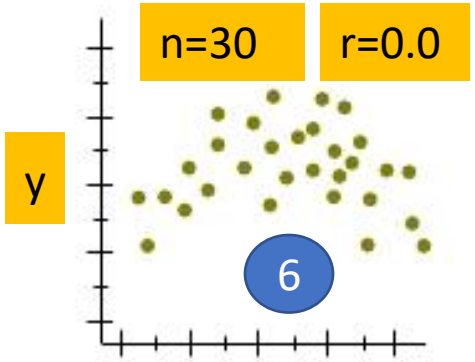
Weak Negative Correlation



Weak Positive Correlation



No Correlation



No Linear Correlation

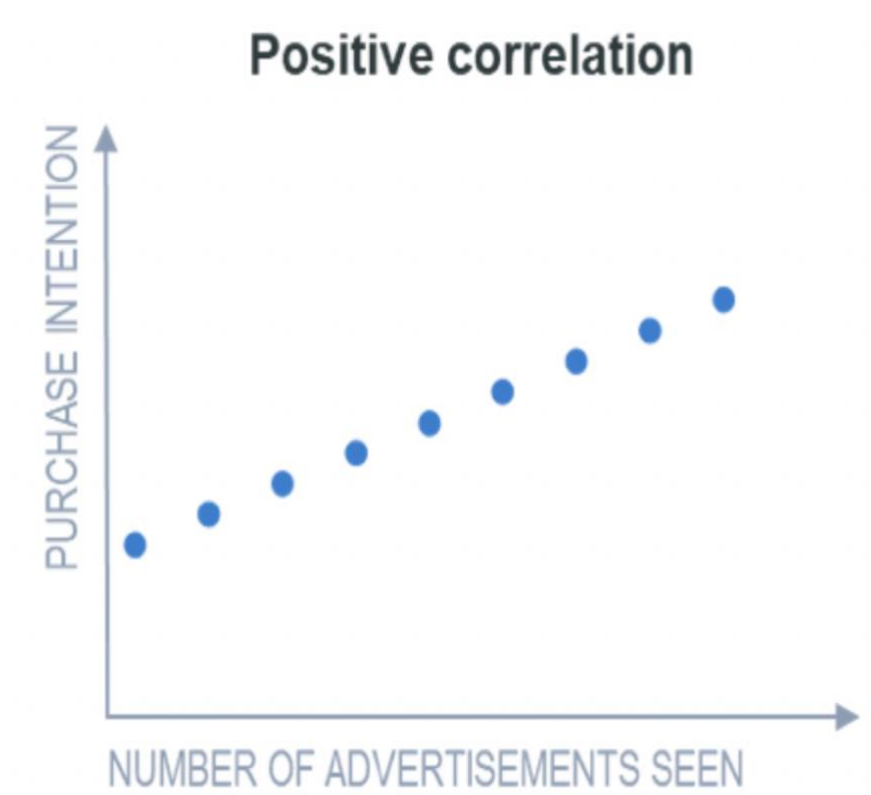
Judge strength of relationship by width or tightness of scatter

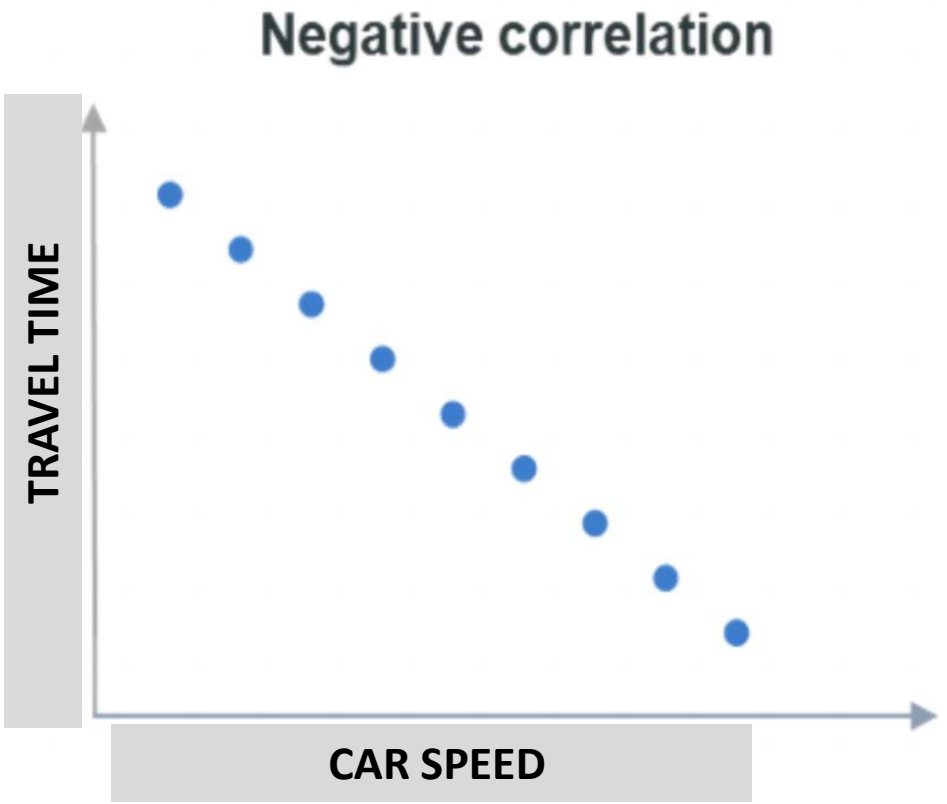
If x increases and y increases, then it is + ve Correlation

If x decreases and y increases, then it is - ve Correlation

CORRELATION

Correlation is usually defined as a measure of the *linear* relationship between two quantitative variables (e.g., height and weight).





When the values of one variable increase as the values of the other increase, this is known as *positive correlation*

When the values of one variable decrease as the values of another increase to form an inverse relationship

The most common measure of correlation is *correlation coefficient*, denoted by 'r'

The value of r will lie between -1 to +1

Positive correlation _ Examples

1. The more time you spend running on a treadmill, the more calories you will burn.
2. The less time I spend marketing my business, the fewer new customers I will have.
3. The more hours you spend in direct sunlight, the more severe your sunburn.
4. The more money she saves, the more financially secure she feels.
5. As the temperature goes up, ice cream sales also go up.
6. The more gasoline you put in your car, the farther it can go.
7. The longer someone invests, the more compound interest he will earn.

Negative correlation _ Examples

1. A student who has many absences has a decrease in grades.
2. As weather gets colder, air conditioning costs decrease.
3. If a train increases speed, the length of time to get to the final point decreases.
4. If a car decreases speed, travel time to a destination increases.
5. If a car tire has more air, the car may use less gas per mile.
6. The warmer it is outside, the fewer layers of clothing one has to wear to be warm.
7. As one exercises more, his body weight becomes less.
8. The older a man gets, the less hair that he has.
9. The more one works, the less free time one has.
10. If the temperatures outside decrease dramatically, heating bills will increase.

Interpretation of Correlation Coefficient (r)

- The value of correlation coefficient 'r' ranges from -1 to +1
- If r = +1, then the correlation between the two variables is said to be perfect and positive
- If r = -1, then the correlation between the two variables is said to be perfect and negative
- If r = 0, then there exists no correlation between the variables

r value	Degree of relationship
1	Perfect
> 0.80	Strong (Good)
> 0.65	Moderate
< 0.65	Weak (Poor)
0	No relationship
	+ or – indicates the direction of relationship

The population parameter for correlation coefficient is expressed as

ρ (rho)

Correlation & causation

Well, according to the Bureau of Statistics correlation is, "A statistical measure (expressed as a number) that describes the size and direction of a relationship between two or more variables."

While causation "Indicates that one event is the result of the occurrence of the other event; i.e. there is a causal relationship between the two events. This is also referred to as cause and effect."

The classic causation vs correlation example that is frequently used is that smoking is correlated with alcoholism, but doesn't cause alcoholism. While smoking causes an increase in the risk of developing lung cancer.

Correlation Quiz 2

6.The determination of population confidence interval based on sample data is called---------- statistics

a) Descriptive
b) Inferential
c) Normal
d) Conclusive

b) Inferential

A confidence interval is an indicator of your measurement's precision. It is also an indicator of how stable your estimate is, which is the measure of how close your measurement will be to the original estimate if you repeat your experiment.

Correlation - Quiz 1

Which of the following statement(s) is/are true:

- 1. Correlation coefficient falls between 1 and – 1 — True
- 2. The nearer the correlation coefficient is to 0, more is the relationship between two variables linear. — False
- 3. Negative correlation means the value of independent variabl decreases as the dependent variable increases. — True
- 4. Correlation does not determine causation — True
- 5 .Regression is used to predict relationship between a discrete and a continuous variable — False
- 6. 2 sample t is used to determine if there is any difference in va f samples taken from two different populations — False

Correlation Quiz 2

1. The purpose of applying Correlation analysis is to

a) Find out the correlation coefficient that exists between the given variables

b) Ascertain the degree of relationship between the given variables

c) Ascertain the nature of relationship that exists between the given variables

d) All of the above

d) All of the above

Correlation Quiz 2

2. In ‘Karl Pearson’s Correlation analysis’, the coefficient between two variables under consideration is found to be -0.1. Which of the following can be a valid inference?

a) There is a negative relationship between the two variables

b) There is a weak negative relationship between the two variables

c) A non-linear relationship may exist between the two variables

d) All of the above

d) All of the above

Correlation Quiz 2

3. In ‘Karl Pearson’s Correlation analysis’, the value of coefficient is – 1.05. Which of the following can be a valid inference?

a) There is a strong negative relationship between the two variables
b) There is a weak negative relationship between the two variables
c) There is an error in calculation
d) All of the above

c) There is an error in calculation

Correlation Quiz 2

4. While applying the Coefficient of correlation analysis on Y and X:

a) Both X and Y should be quantitative
b) Y should be quantitative, X can be qualitative or quantitative
c) Both X and Y should be qualitative
d) Y can be qualitative or quantitative, X should be quantitative

a) Both X and Y should be quantitative

Correlation Quiz 2

5. If ‘Y’ increases in all the cases as ‘x’ increases, the correlation coefficient will be:

a) Positive
b) Negative
c) 1
d) -1

a) Positive

Correlation Quiz 2

6.If X and Y are perfectly correlated , the most appropriate inference is

a) X causes Y
b) X is atleast one of the causes of Y
c) X has perfect linear association with Y
d) All of the above

C

Learning Objectives

In this chapter, you learn:

- How to use regression analysis to predict the value of a dependent variable based on an independent variable
- The meaning of the regression coefficients b_0 and b_1
- How to evaluate the assumptions of regression analysis and know what to do if the assumptions are violated
- To make inferences about the slope and correlation coefficient

Regression Analysis

- Regression Analysis is mathematical measure of average relationship between two or more variables.
- Regression analysis is a statistical tool used in prediction of value of unknown variable from known variable.

Regression analysis generates a line that quantifies the relationship between X and Y.

A Note on Terminology

- These terms are often used interchangeably:

- **Regression equation**
- **Regression line**
- **Prediction equation**
- **Prediction line**
- **Fitted line or fits*Model**

Introduction to Regression Analysis

- Regression analysis is used to:
  - Predict the value of a dependent variable based on the value of at least one independent variable
  - Explain the impact of changes in an independent variable on the dependent variable

Dependent variable: the variable we wish to predict or explain

Independent variable: the variable used to predict or explain the dependent variable

Simple Regression

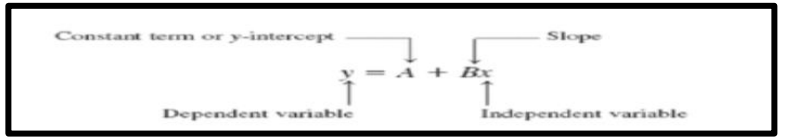
Definition

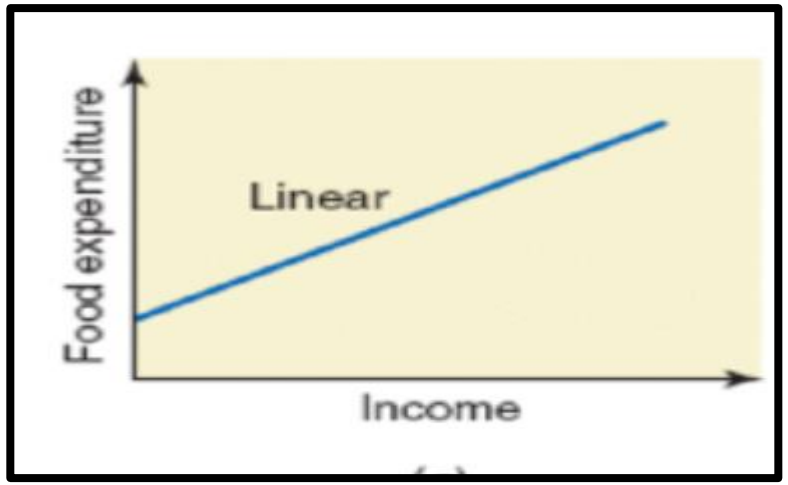
A regression model is a mathematical equation that describes the relationship between two or more variables. A ***simple regression*** model includes only two variables: one independent and one dependent. The dependent variable is the one being explained, and the independent variable is the one used to explain the variation in the dependent variable.

Linear Regression

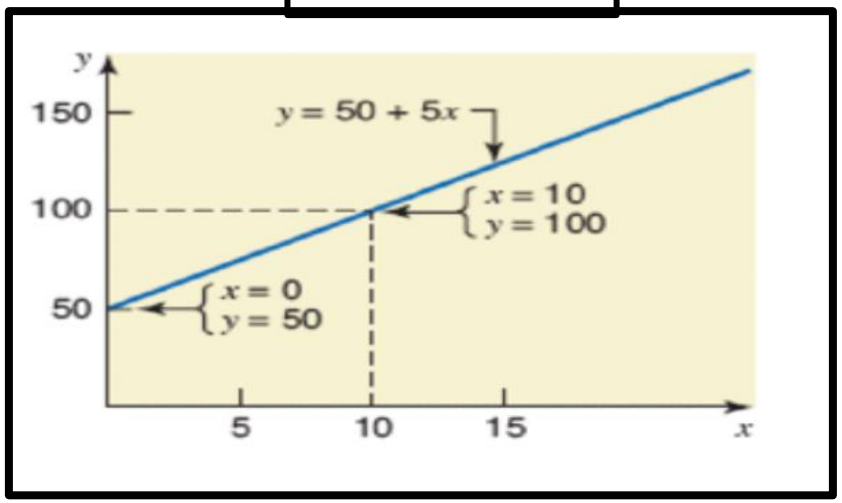
Definition

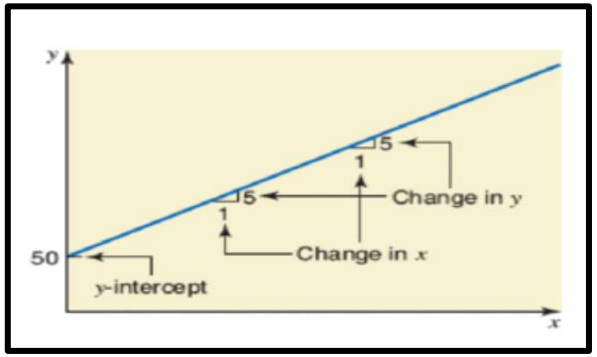
A (simple) regression model that gives a straight-line relationship between two variables is called a ***linear regression*** model.





Regression equation
Y= 50+5x





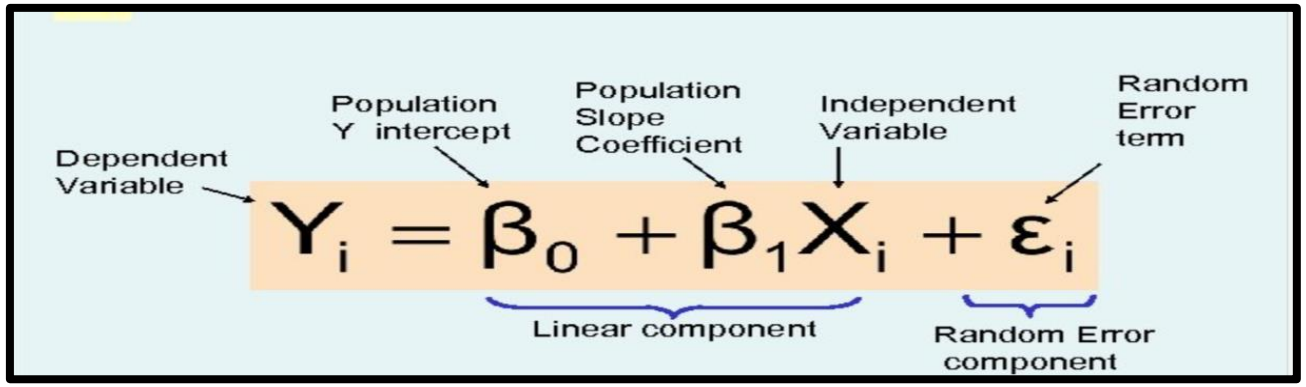
Regression Analysis

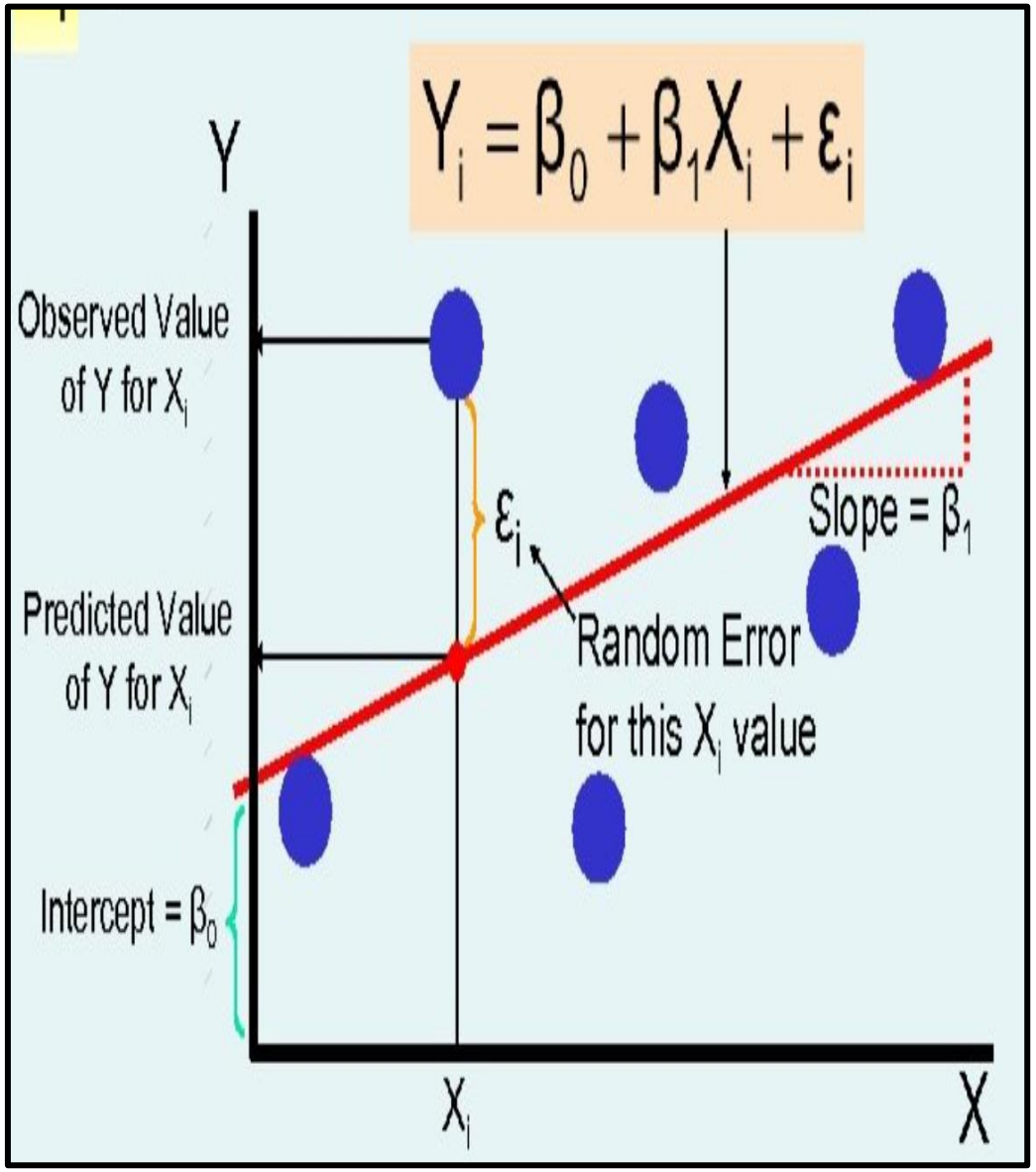
- In **Regression Analysis**, we use the independent variable, X, to estimate the value of the dependent variable, Y.
  - In Linear Regression Analysis, the relationship between the variables is linear.
  - Both variables must be at least interval scale.
  - The least squares criterion is used to determine the equation.
    That is, the term $\Sigma(Y-Y')^2$ is minimized.

There are 2 regression lines . To estimate Y ,use Y on X regression line and to estimate X use X on Y regression Line

Simple Linear Regression Model

- Only **one** independent variable, X
- Relationship between X and Y is described by a linear function
- Changes in Y are assumed to be related to changes in X





Simple Linear Regression Equation (Prediction Line)

The simple linear regression equation provides an estimate of the population regression line

Estimated (or predicted) Y value for observation i

Estimate of the regression intercept

Estimate of the regression slope

Value of X for observation i

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

- b_0 is the estimated average value of Y when the value of X is zero
- b_1 is the estimated change in the average value of Y as a result of a one -unit change in X

The Least Squares Method

b_0 and b_1 are obtained by finding the values of that minimize the sum of the squared differences between Y and $\hat{Y}$:

$$\min \sum (Y_i - \hat{Y}_i)^2 = \min \sum (Y_i - (b_0 + b_1 X_i))^2$$

QUIZ

1. What does a simple linear regression analysis examine?

a

a. The relationship between only two variables

b. The relationship between one dependent and one independent variable

c. The relationship between many variables

d. The relationship between two dependent and one independent variable

2. Which of the following are correct?

a ,d

a. The intercept/constant (β_0) is the mean-Y when X=0

b. The intercept is the the amount of change in mean-Y when X=0

c. The coefficient is the mean-Y at a certain value of X

d. The coefficient (β) is the amount of change in mean-Y for every unit increase in X

3. What does the least squares method do exactly?

c

a. Minimizes the distance between the data points

b. Finds the least problematic regression line

c. Finds those (best) values of the intercept and slope that provide us with the smallest value of the residual sum of squares

d. Finds those (best) values of the intercept and slope that provide us with the smallest value of the sum of residuals

4. Which of the following measures is optimal for comparing the goodness of the fit of competing regression models involving the same dependent variable?

d

a. The intercept

b. The coefficient

c. R-square

d. Standard deviation of the residuals

Simple Linear Regression Example

- A real estate agent wishes to examine the relationship between the selling price of a home and its size (measured in square feet)

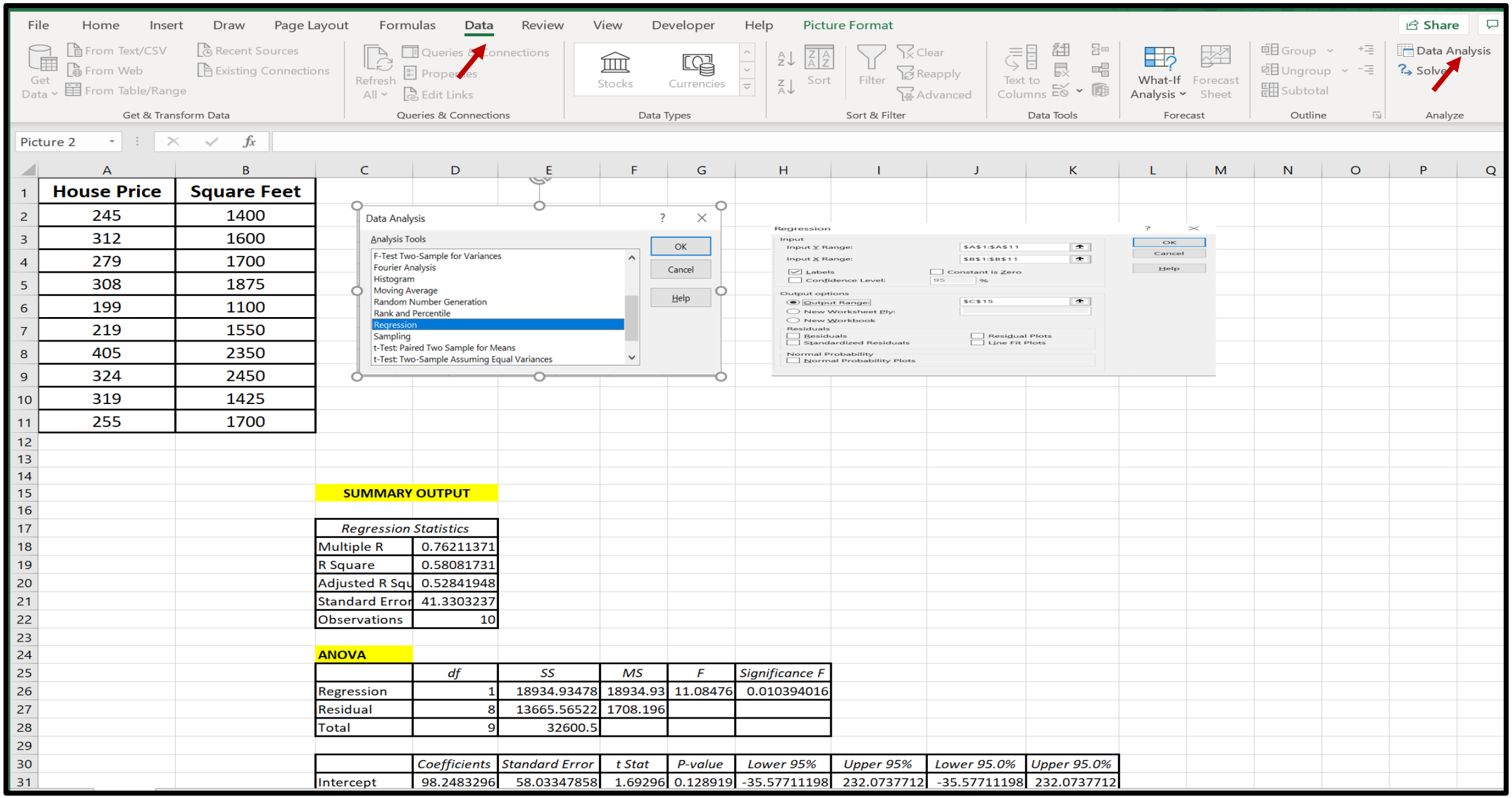
- A random sample of 10 houses is selected
  - Dependent variable (Y) = house price in $1000s
  - Independent variable (X) = square feet



Simple Linear Regression Example: Data

House Price in $1000s (Y)	Square Feet (X)
245	1400
312	1600
279	1700
308	1875
199	1100
219	1550
405	2350
324	2450
319	1425
255	1700





	A	B
1	**House Price**	**Square Feet**
2	245	1400
3	312	1600
4	279	1700
5	308	1875
6	199	1100
7	219	1550
8	405	2350
9	324	2450
10	319	1425
11	255	1700

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.76211371
R Square	0.58081731
Adjusted R Squ	0.52841948
Standard Error	41.3303237
Observations	10

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	18934.93478	18934.93	11.08476	0.010394016
Residual	8	13665.56522	1708.196		
Total	9	32600.5			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	98.2483296	58.03347858	1.69296	0.128919	-35.57711198	232.0737712	-35.57711198	232.0737712

Example: Excel Output

Regression Statistics	
Multiple R	0.76211
R Square	0.58082
Adjusted R Square	0.52842
Standard Error	41.33032
Observations	10

The regression equation is:

$$\widehat{\text{house price}} = 98.24833 + 0.10977\ (\text{square feet})$$

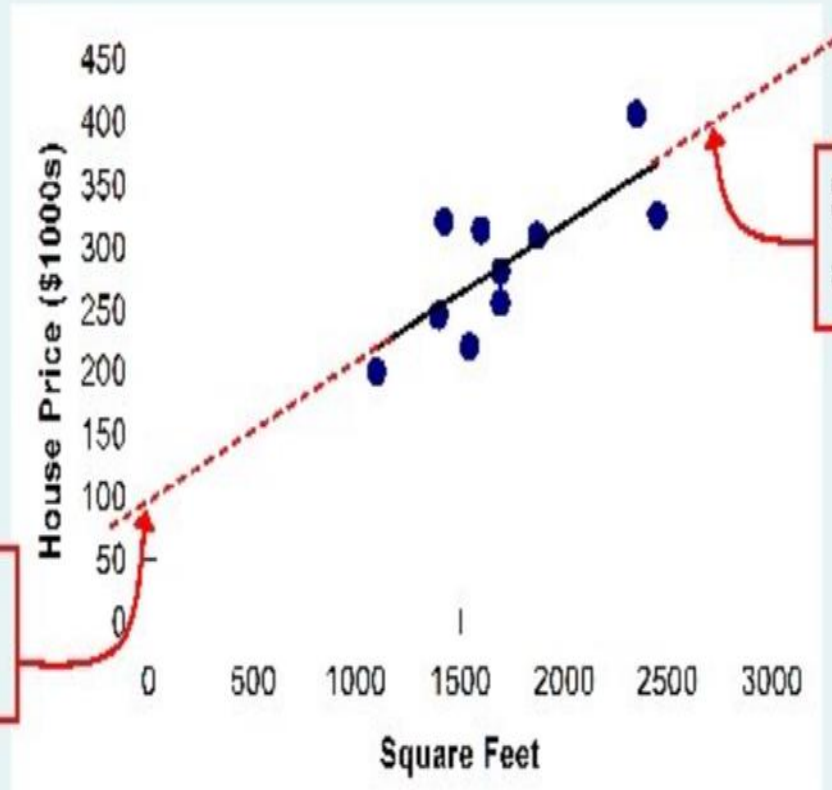
ANOVA	df	SS	MS	F	Significance F
Regression	1	18934.9348	18934.9348	11.0848	0.01039
Residual	8	13665.5652	1708.1957		
Total	9	32600.5000			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	98.24833	58.03348	1.69296	0.12892	-35.57720	232.07386
Square Feet	0.10977	0.03297	3.32938	0.01039	0.03374	0.18580



House price model: Scatter Plot and Prediction Line





$$\widehat{\text{house price}} = 98.24833 + 0.10977\ (\text{square feet})$$

$$\widehat{\text{house price}} = 98.24833 + 0.10977\ (\text{square feet})$$

- b_0 is the estimated average value of Y when the value of X is zero (if X = 0 is in the range of observed X values)
- Because a house cannot have a square footage of 0, b_0 has no practical application

- b_1 estimates the change in the average value of Y as a result of a one-unit increase in X
  - Here, $b_1 = 0.10977$ tells us that the mean value of a house increases by .10977($1000) = $109.77, on average, for each additional one square foot of size

Predict the price for a house with 2000 square feet:

$$\widehat{\text{house price}} = 98.25 + 0.1098\ (\text{sq.ft.})$$

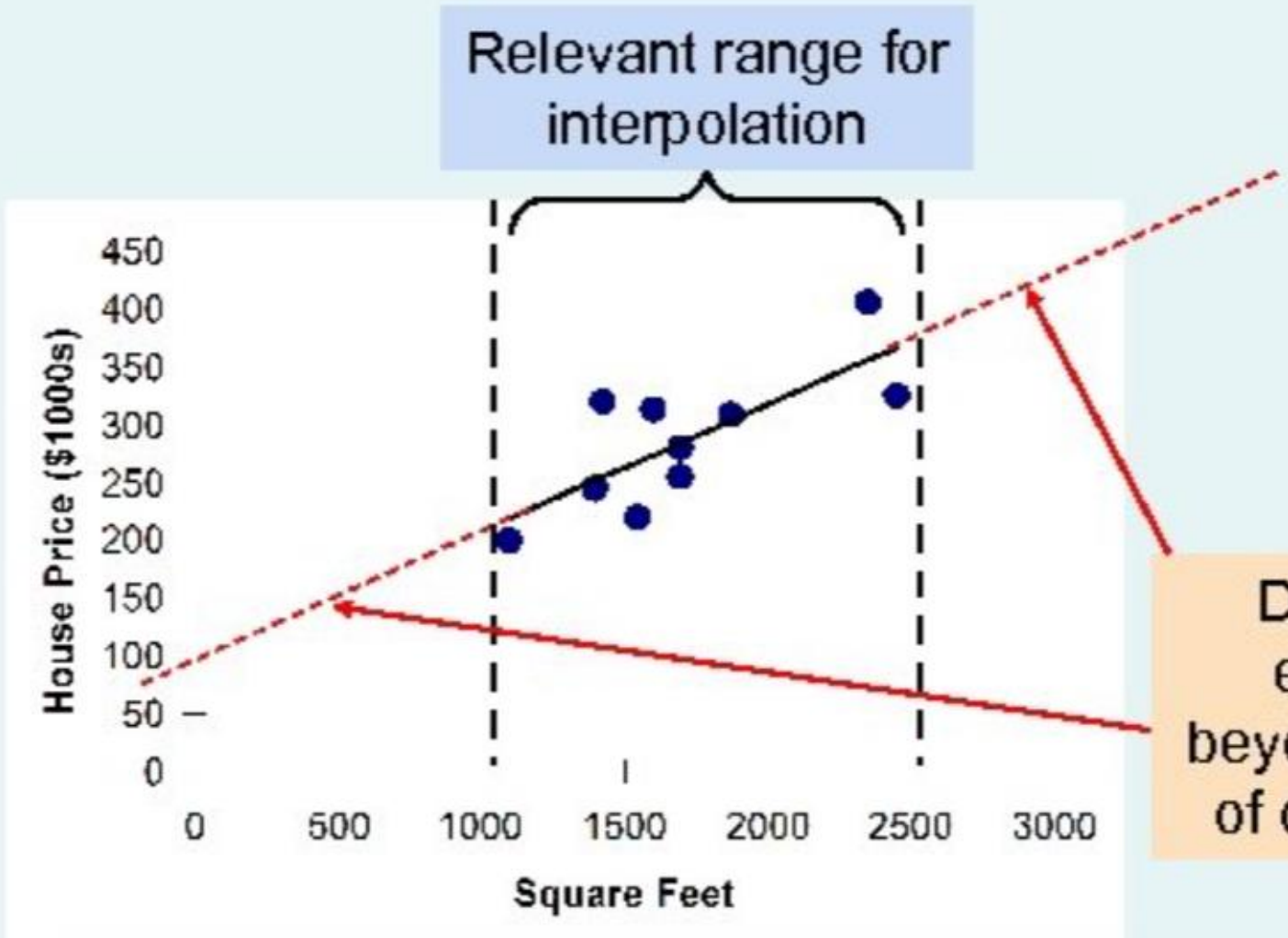
$$= 98.25 + 0.1098(2000)$$

$$= 317.85$$

The predicted price for a house with 2000 square feet is 317.85($1,000s) = $317,850

- When using a regression model for prediction, only predict within the relevant range of data

House Price in $1000s (Y)	Square Feet (X)
245	1400
312	1600
279	1700
308	1875
199	1100
219	1550
405	2350
324	2450
319	1425
255	1700



Coefficient of Determination, r^2

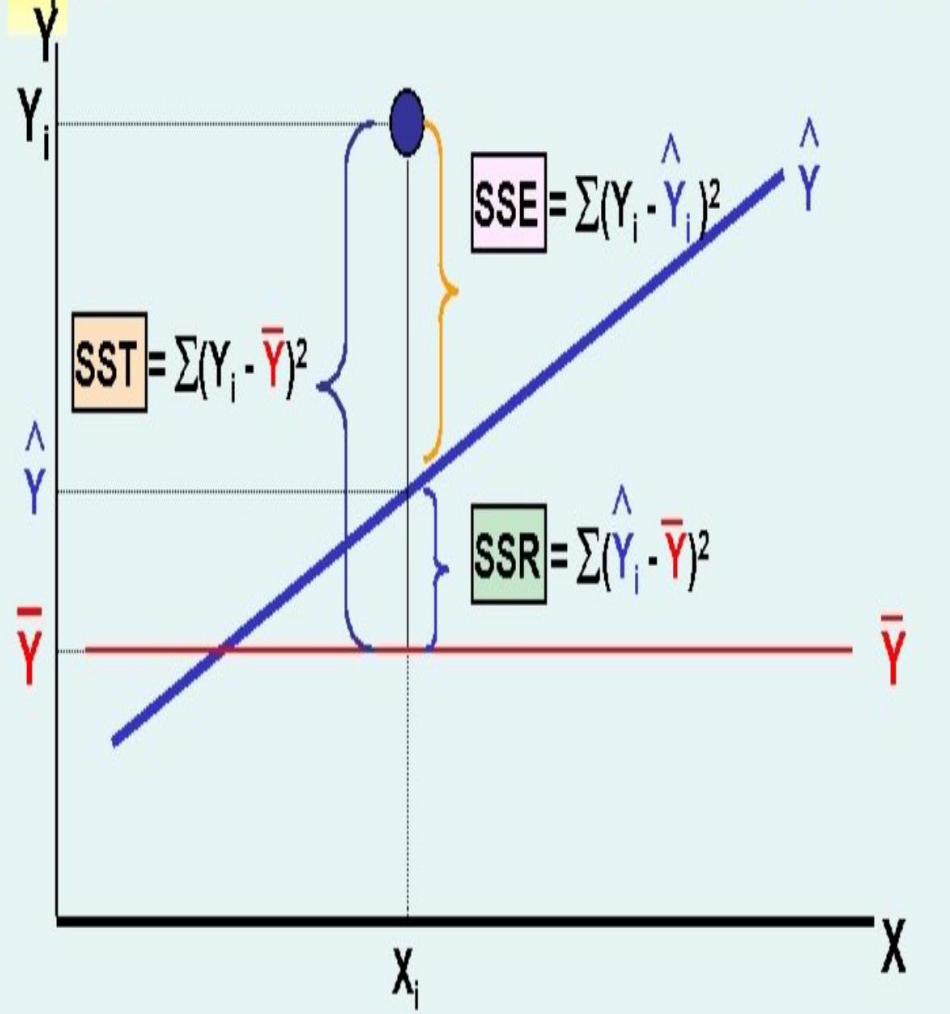
- The coefficient of determination is the portion of the total variation in the dependent variable that is explained by variation in the independent variable
- The coefficient of determination is also called r -squared and is denoted as r^2

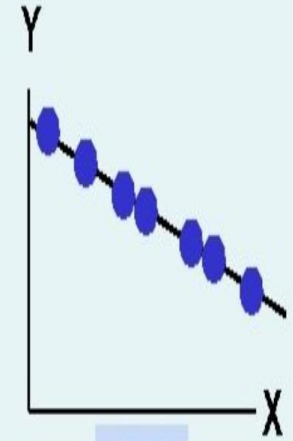
$$r^2 = \frac{SSR}{SST} = \frac{\text{regression } \textit{sum} \text{ of squares}}{\textit{total} \text{ sum of squares}}$$

note: $0 \le r^2 \le 1$

Measures of Variation

(continued)

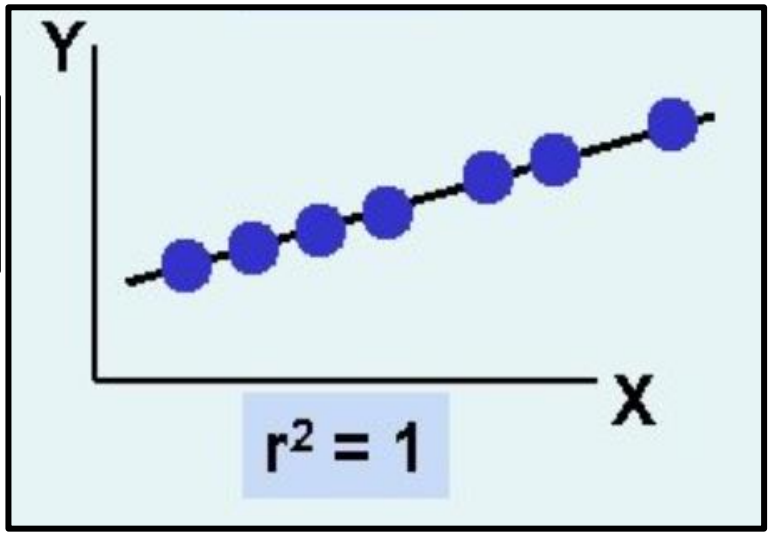


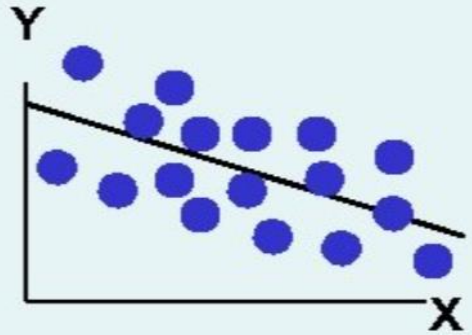


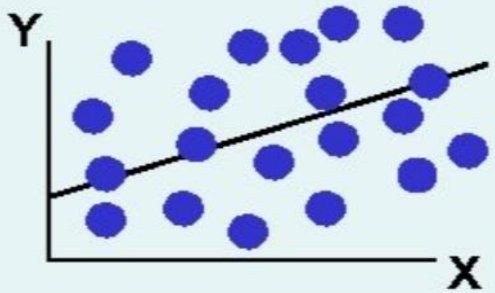
$r^2 = 1$

Perfect linear relationship between X and Y:

100% of the variation in Y is explained by variation in X



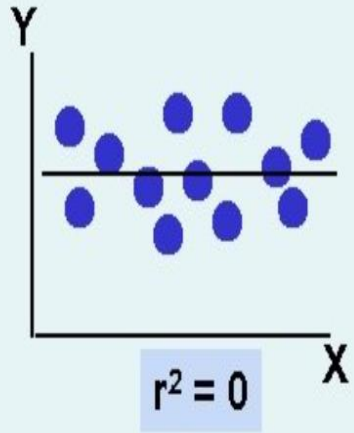




$0 < r^2 < 1$

Weaker linear relationships between X and Y:

Some but not all of the variation in Y is explained by variation in X



$r^2 = 0$

No linear relationship between X and Y:

The value of Y does not depend on X. (None of the variation in Y is explained by variation in X)

Coefficient of Determination, r^2 in Excel

Regression Statistics	
Multiple R	0.76211
R Square	0.58082
Adjusted R Square	0.52842
Standard Error	41.33032
Observations	10

$$r^2 = \frac{SSR}{SST} = \frac{18934.9348}{32600.5000} = 0.58082$$

58.08% of the variation in house prices is explained by variation in square feet

ANOVA

	df	SS	MS	F	Significance F
Regression	1	18934.9348	18934.9348	11.0848	0.01039
Residual	8	13665.5652	1708.1957		
Total	9	32600.5000			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	98.24833	58.03348	1.69296	0.12892	-35.57720	232.07386
Square Feet	0.10977	0.03297	3.32938	0.01039	0.03374	0.18580

Standard Error of Estimate

- The standard deviation of the variation of observations around the regression line is estimated by

$$S_{YX} = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n-2}}$$

Where

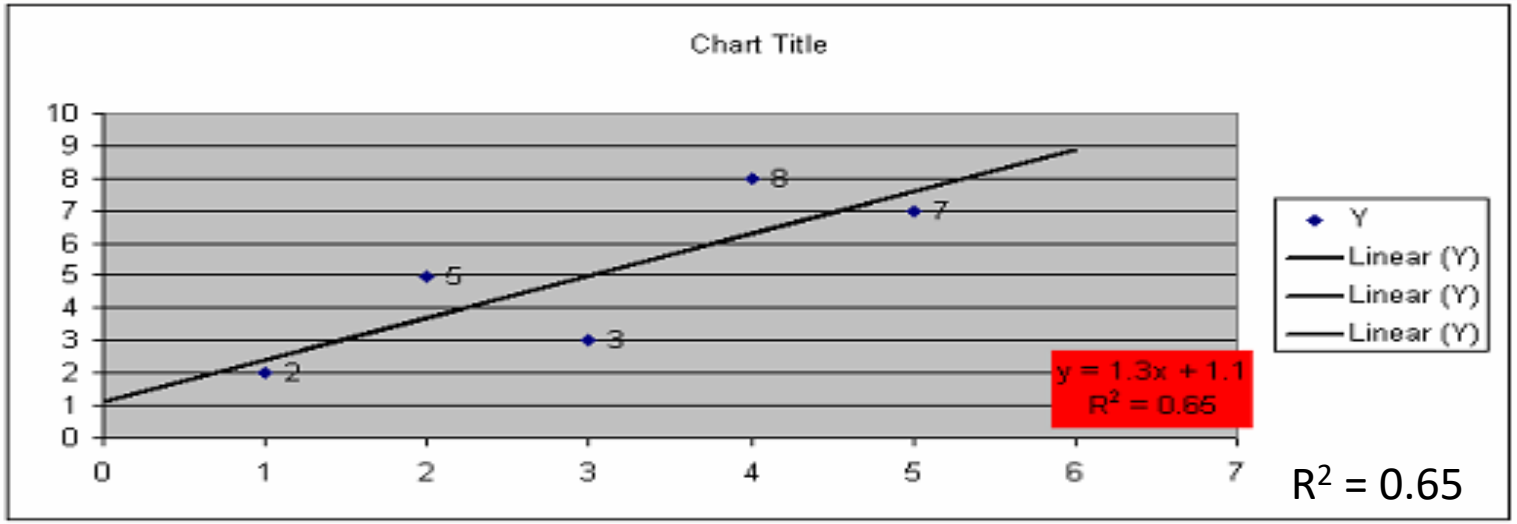
SSE = error sum of squares

n = sample size

Microsoft Excel - Prob2.xls

File Edit View Insert Format Tools Data Window Help Adobe PDF

	X	Y
1	1	2
2	2	5
3	3	3
4	4	8
5	5	7



$R^2 = 0.65$

X = Adv Expenditure in ten thousands
Y = sales revenue in Lacs

X	Y	Y_c=1.3X + 1.1	Y-Y_c	(Y-Y_c)^2
1	2			
2	5			
3	3			
4	8			
5	7			

Sheet4 Sheet1 Sheet2 Sheet3

Draw AutoShapes

Ready NUM

Example: Regression Analysis

- Is there a relationship between calories consumed and weight gained?
- Please find the following
  - Rsq=
  - r=
  - Regression Equation
- If 3300 calories are consumed, how much weight will be gained?

Weight gained	Calories Consumed
0	1500
200	2300
900	3400
200	2200
300	2500
0	1600
0	1400
10	1900
600	2800
1100	3900
100	1670
150	1900
350	2700
700	3000

The regression equation is

Weight gained = - 755 + 0.463 Calories Consumed

Weight gained for 3300 calories consumed

y = -755 + 0.463 * 3300 = 772.9

Predictor	Coef	SE Coef	T	P
Constant	-754.64	82.56	-9.14	0.000
Calories Consumed	0.46307	0.03369	13.74	0.000

S = 91.3702 R-Sq = 94.0% R-Sq(adj) = 93.5%

The Multiple Regression Model

Idea: Examine the linear relationship between 1 dependent (Y) & 2 or more independent variables (X_i)

Multiple Regression Model with k Independent Variables:

Y-intercept | Population slopes | Random Error

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} + \varepsilon$$

MULTIPLE REGRESSION - EXAMPLE

The marketing data contains the impact of three advertising medias; Facebook, Youtube and newspapers on sales. Check if the data can be used for predicting sales (in million dollars) from the budget (in thousand dollars) invested in the three advertising medias.

Regression Equation

Sales (in million $) = 3.527 + 0.04576 Youtube (in 1000$) + 0.18853 Facebook (in 1000$) - 0.00104 Newspaper (in 1000$)

