Regression Analysis

Unit 2

Unit Objectives

- At the end of the unit, the students are expected to:
  1. State what is regression analysis.
  2. Distinguish the response and explanatory variable
  3. Use MSExcel Data Analysis Toolpack
  4. Interpret results generated by the Data Analysis Toolpack
  5. Use the MSExcel scatterplot to show relationship between variables.
  6. Predict the response variable based on the value of the explanatory variable.

Unit Contents

- Define is regression analysis?
- Response and explanatory variables
- Correlation and the Scattergram
- The MSExcel Analysis Toolpack
- Simple Linear Regression
- Multiple Regression Analysi

Regression Analysis in Business

- Businesses are interested in predicting future production, consumption, investment, prices, profits, sales etc.
- Sociology
  - Sociological studies
  - Economic planning
    - Projections of population, birth rates, death rates, etc.

Regression Analysis

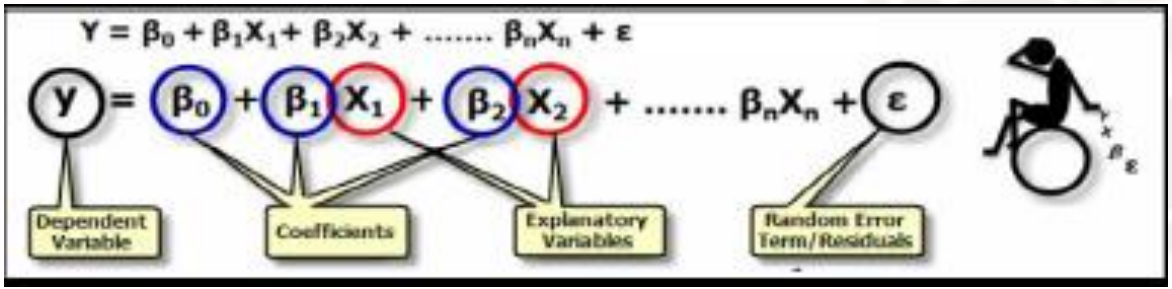
- Definition
  - Technique in studying the dependence of one variable (dependent variable) on one or more variables (explanatory variable,
  - to estimate or predict the average value of the dependent variables in terms of the known or fixed values of the independent variable
- Purposes
  - Estimate the relationship that exists between the dependent variable and the explanatory variable
  - Determine the effect of each of the explanatory variable on the dependent variable, controlling the effects of all other explanatory variables.
  - Predict the value of the dependent variable for a given explanatory variable.

Regression Analysis

- Regression analysis is the most often applied technique of statistical analysis and modeling.
- Regression Models
  - Simple Linear Regression Model
    - Involves only one dependent variable
  - Multiple Regression Model
    - Multiple dependent variables
    - One dependent variable in multiple forms

Variables

- Independent/Explanatory/Predictor Variable
  - Basis of the estimation
- Dependent or response variables
  - The variable to be predicted or estimated



Which is the dependent variable?

- Level of education and income
- Price of house and its square footage
- Sales volume and ads expenses

Which is the dependent variable?

1. Smoking and your health
   A. Smoking status
   B. Survival status

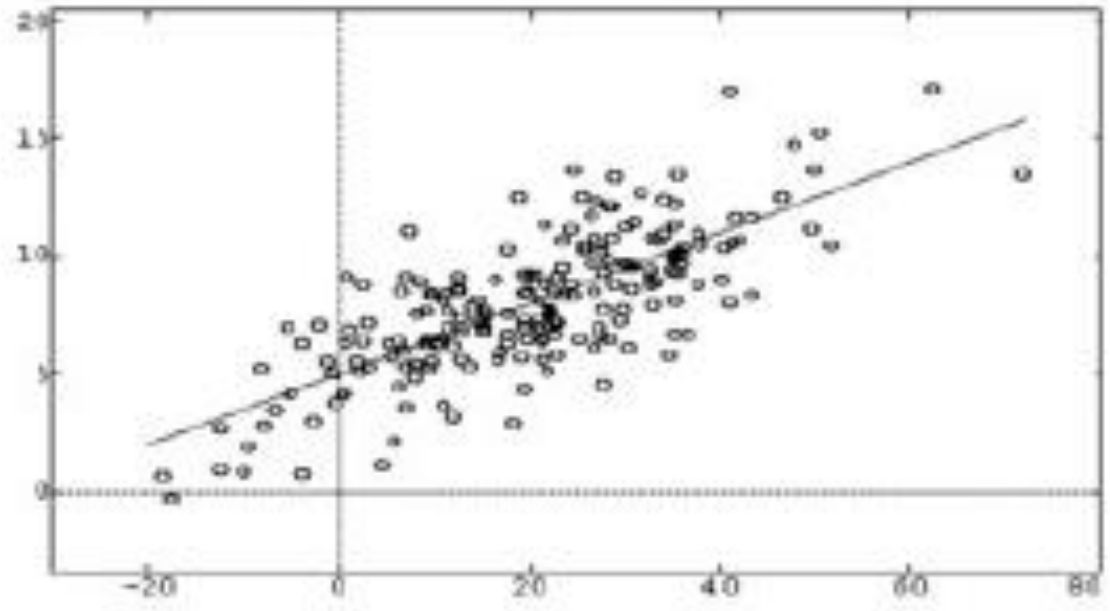
2. Air pollution
   A. Carbon dioxide level of a country
   B. Country's amount of gasoline use for automobiles
3. Student achievement
   A. GWA
   B. Number of hours spent in studying

Check your understanding!!!

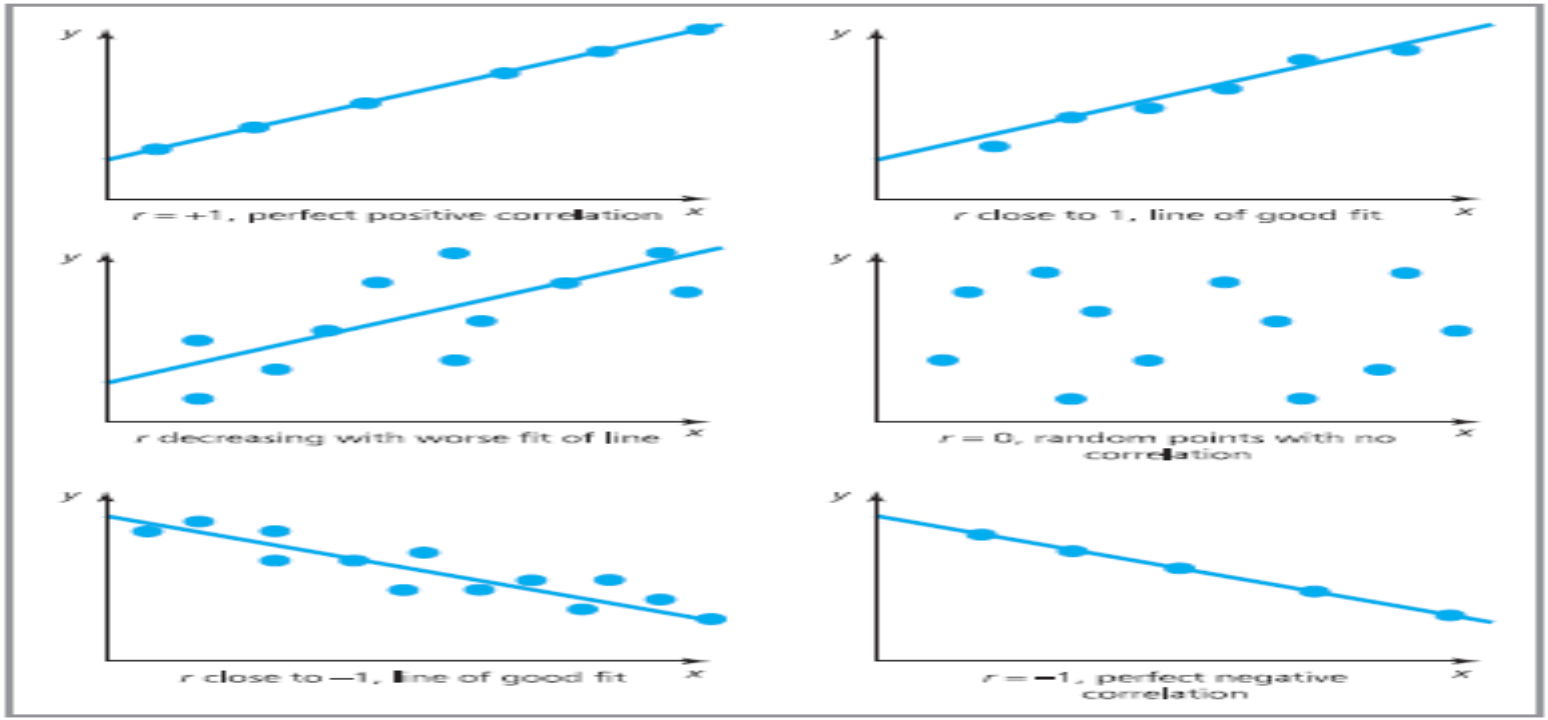
Open your Canvas!!!

Scatter Diagrams

- Scatterplot



Coefficient of Correlation



Measuring the strength of a relationship

- Coefficient of Determination = R^2
  - r^2 is the proportion of variability in Y that is explained by the regression equation.

$$\text{coefficient of determination} = \left[\frac{n\Sigma xy - \Sigma x \Sigma y}{\sqrt{[n\Sigma x^2 - (\Sigma x)^2] \times [n\Sigma y^2 - (\Sigma y)^2]}}\right]^2$$

- Correlation Coefficient – r
  - Measures the degree or strength of the linear relationship

$$r = \pm\sqrt{r^2}$$

Linear Regression

- LR finds values for the constants $\boldsymbol{a}$ and $\boldsymbol{b}$ that define the line of best fit through a set of points.
- Approach to LR
  - Draws a scattergram
  - Identifies a linear relationship
  - Draws a line of best fit through the data
  - Uses the line to predict a value of the dependent variable from a known value of the independent variable.

Simple Regression Analysis

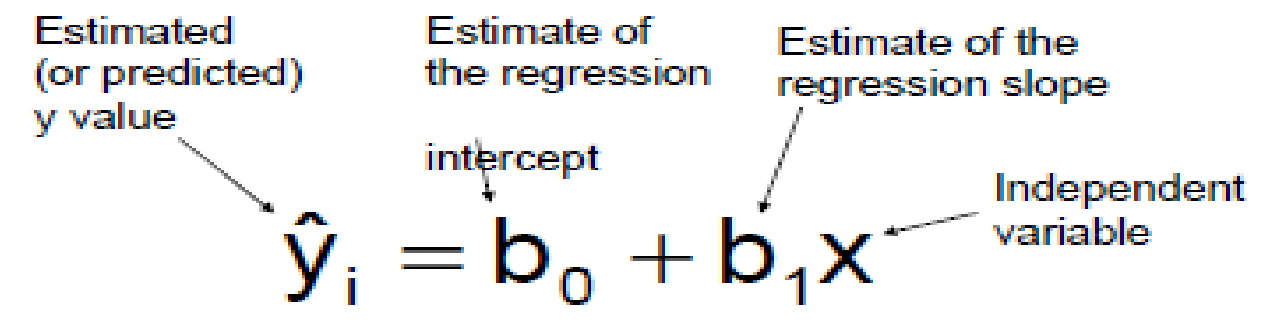
- Implicit assumption
  - There is a relationship that exists between the variables
  - There is a random error that cannot be predicted

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

  - Where:
    - Y = dependent variable (response variable)
    - X = independent variable
    - b_0 = intercept (value of Y when X = 0)
    - β_1 = slope of the regression line
    - ε= random error

Estimated Regression Model

- The sample regression line provides an estimate of the population regression line



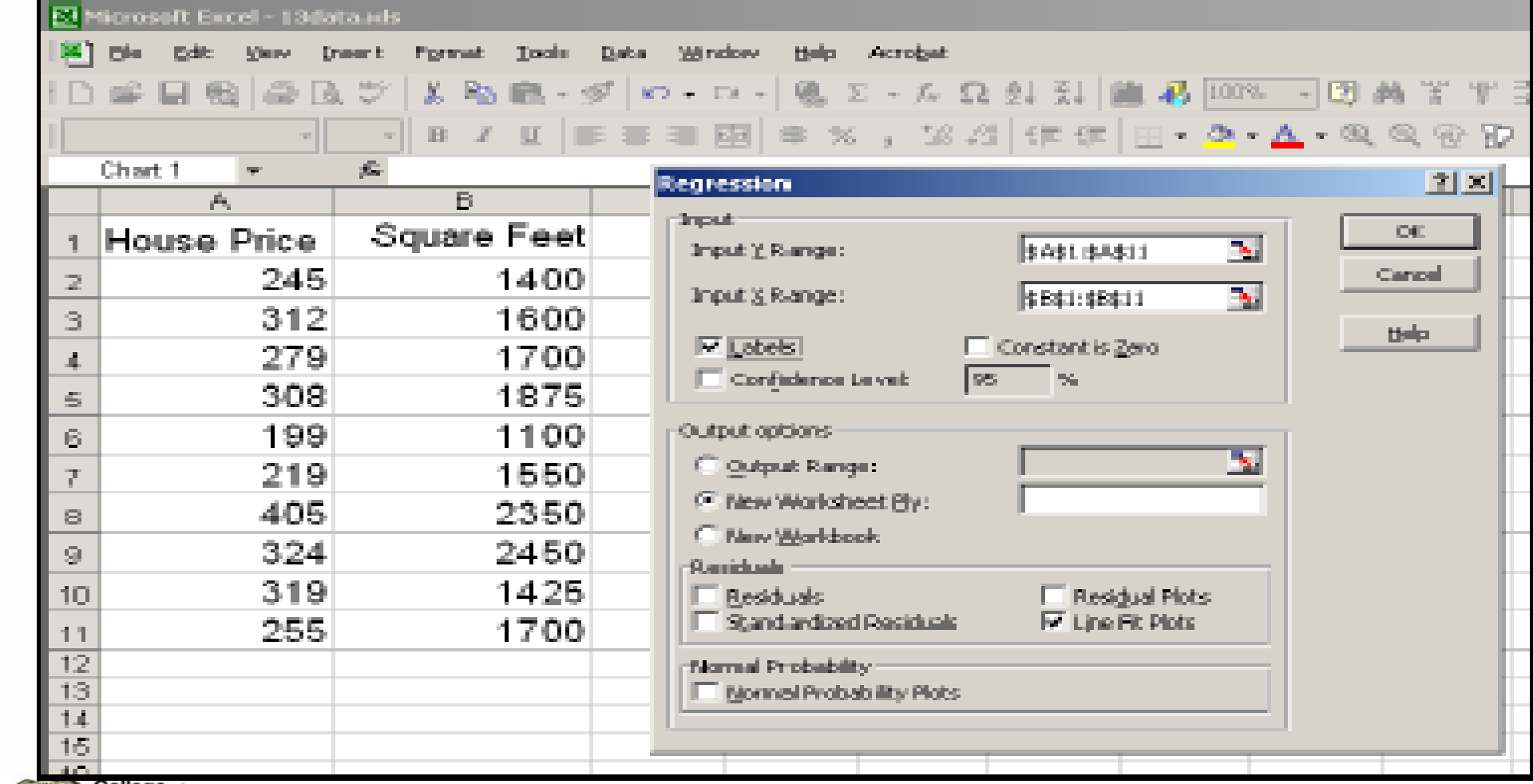
The individual random error terms, have a mean of zero.

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.

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


Microsoft Excel - 13data.xls
File Edit View Insert Format Tools Data Window Help Acrobat
Chart 1
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
12
13
14
15
Regression
Input
Input Y Range: $A$1:$A$11
Input X Range: $B$1:$B$11
Labels
Constant is Zero
Confidence Level: 95 %
OK
Cancel
Help
Output options
Output Range:
New Worksheet Ply:
New Workbook
Residuals
Residuals
Residual Plots
Standardized Residuals
Line Fit Plots
Normal Probability
Normal Probability Plots

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:

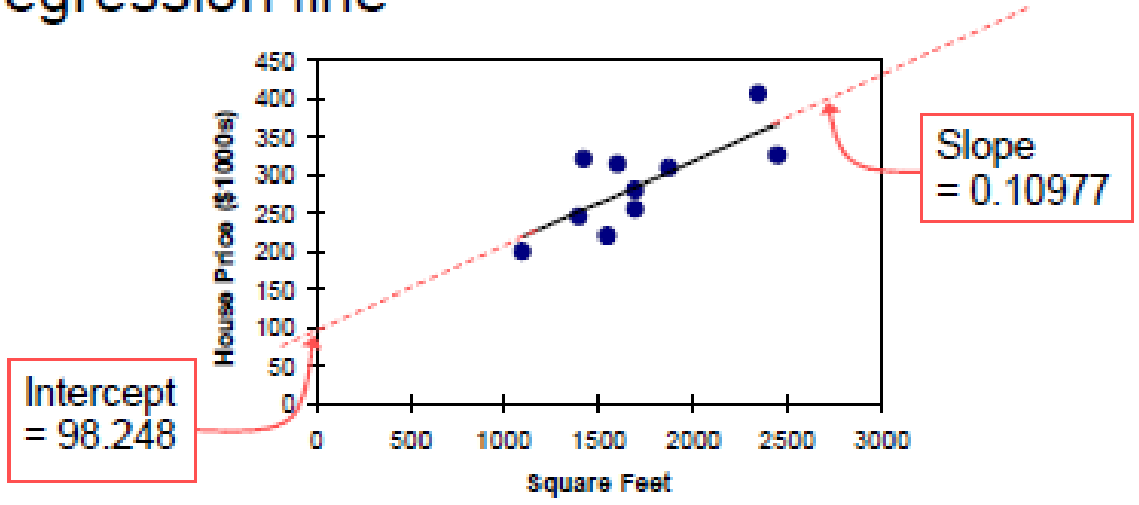
$$\widehat{\text{house price}} = 98.24833 + 0.10977\ (\text{square feet})$$

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

Graphical Presentation

- House price model: scatter plot and regression line



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

Interpretation of the intercept, b_0

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

- 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)
  - Here, no houses had 0 square feet, so b_0 = 98.24833 just indicates that, for houses within the range of sizes observed, $98,248.33 is the portion of the house price not explained by square feet

Interpretation of the slope, b_1

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

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

Predicting the dependent variable

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

Estimated Regression Equation:

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

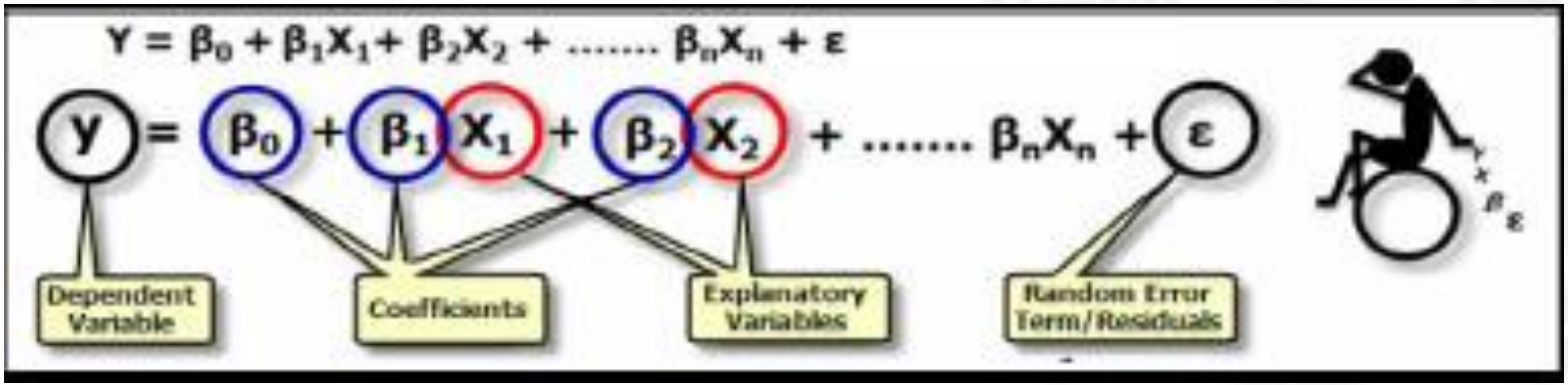
How?

Predict the price for a house with 2000 square feet

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

Multiple Regression Analysis

- A multiple regression model has more than one independent variable
-



Multiple Regression - Example

- A real estate agent wishes to examine the relationship between the selling price of a home, its size (measured in square feet) and the number of bedrooms.
- The data is given in the table.

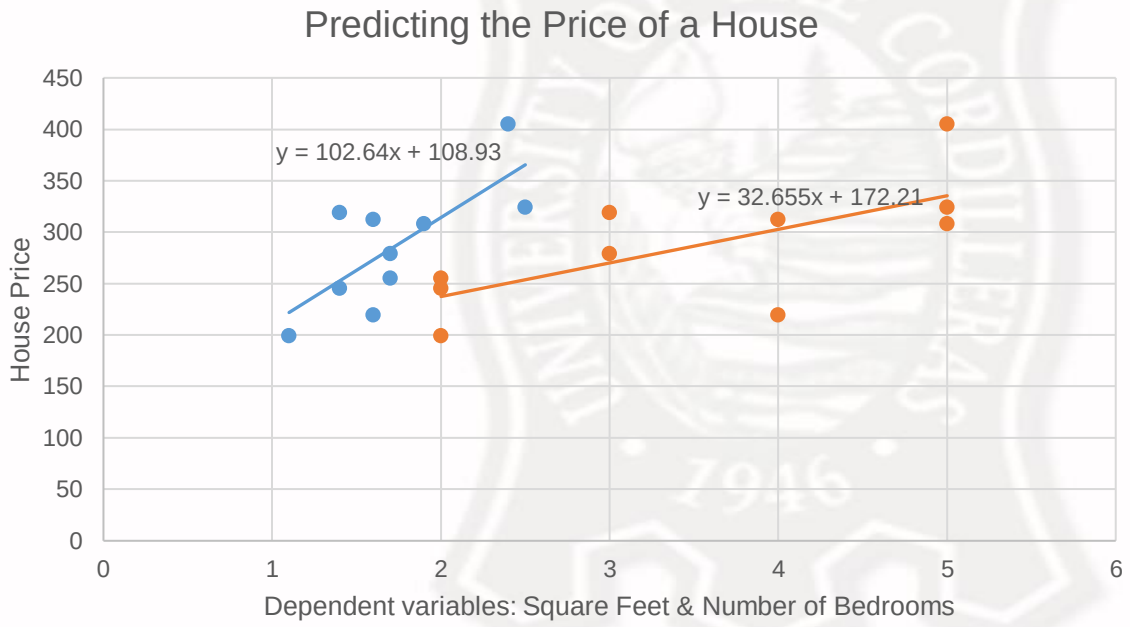
House Price in $1000s (y)	Square Feet (x1) (100's)	Number of Bedrooms (x2)
245	1.4	2
312	1.6	4
279	1.7	3
308	1.9	5
199	1.1	2
219	1.6	4
405	2.4	5
324	2.5	5
319	1.4	3
255	1.7	2

Multiple Regression - Example

The "blue" scatterplot and trendline represent the scatterplot for square footage and the house price.

The "orange" scatterplot and trendline represent the scatterplot for number of bedrooms and the house price.

How? Watch…
https://www.youtube.com/watch?v=iHotbLzZIKI



Multiple Regression - Example

- The summary table is generated using MS Excel Data Analysis – Correlation.
- The table shows that a *positive moderate correlation (0.75)* exists between square footage and house price.
- The table also shows that a *positive moderate correlation (0.69)* exists between number of bedrooms and house price.

	House Price in $1000s (y)	*Square Feet (x1) (100's)*	*Number of Bedrooms (x2)*
House Price in $1000s (y)	1		
Square Feet (x1) (100's)	0.74578059	1	
Number of Bedrooms (x2)	0.688690119	0.790719883	1

Multiple Regression - Example

- The summary output is generated by MSExcel Data Analysis – Regression.
- The Multiple R which is 0.77 (round up value of 0.76311) represents a moderate positive correlation between the house price and the two independent variable taken all together.

$$\hat{Y} = 114.83 + 73.90x_1 + 12.53x_2$$

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.76311
R Square	0.58233
Adjusted R Square	0.463
Standard Error	44.104
Observations	10

ANOVA

	df	*SS*	*MS*	*F*	*gnificance F*
Regression	2	18984.4	9492.19	4.87991	0.04709
Residual	7	13616.1	1945.16		
Total	9	32600.5			

	Coefficients	*tandard Er*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*
Intercept	114.821	60.4693	1.89883	0.09938	-28.166	257.808	-28.166	257.808
Square Feet (x1) (100's)	73.8986	54.9173	1.34563	0.22037	-55.96	203.757	-55.96	203.757
Number of Bedrooms (x2)	12.5242	18.9198	0.66196	0.52918	-32.214	57.2623	-32.214	57.2623

Multiple Regression - Example

- Estimating the price of the house.
  - How much is the price of a house with an area of 2000 square feet and with 5 bedrooms?

$$\hat{Y} = 114.83 + 73.90x\_1 + 12.53x\_2$$

$$\hat{Y} = 114.83 + 73.90(2.0) + 12.53(5)$$

$$\hat{Y} = 325.28$$

  - So, a house which is 2000 ft2 with 5 bedrooms is **$325,280**


QUESTIONS

Your turn. Regression Analysis

12.12 A regional airline transfers passengers from small airports to a larger regional hub airport. The airline's data analyst was assigned to estimate the revenue (in thousands of dollars) generated by each of the 22 small airports based on two variables: the distance from each airport (in miles) to the hub and the population (in hundreds) of the cities in which each of the 22 airports is located. The data is given in the following table.

Airport	Revenue	Distance	Population	Airport	Revenue	Distance	Population
1	233	233	56	12	267	205	96
2	272	209	74	13	338	214	96
3	253	206	67	14	243	183	73
4	296	232	78	15	252	230	55
5	268	125	73	16	269	238	91
6	296	245	54	17	242	144	64
7	276	213	100	18	233	220	60
8	235	134	98	19	234	170	60
9	253	140	95	20	450	170	240
10	233	165	81	21	340	290	70
11	240	234	52	22	200	340	75

With MSExcel, perform the following. Each number should be answered with supporting sheets.

Sheet 1. Determine the correlation of the independent variable with the dependent variables? Which variable has the highest correlation with the dependent variable?

Sheet 2. Super-impose the scatterplots in one graphing plane: revenue versus distance, revenue versus population What are the linear regression equations?

Sheet 3. What is the multiple regression equation that predicts the revenue?

* Estimate the revenue of the airport which is located in a city with 89,000 residents and whose distance from the hub is 200 miles.