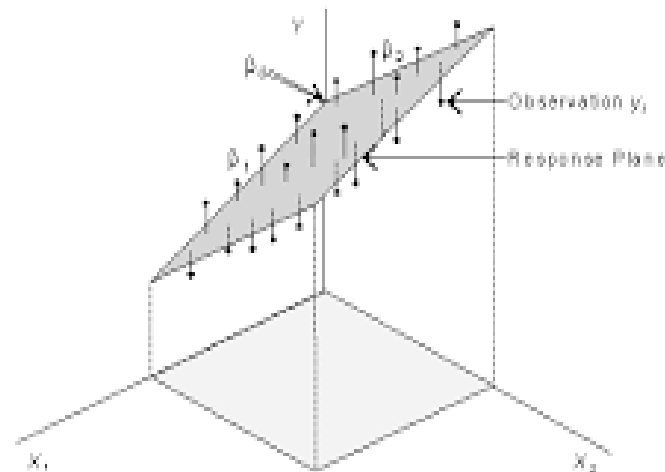


# MULTIPLE REGRESSION ANALYSIS AND MODEL BUILDING

---



# INTRODUCTION TO MULTIPLE REGRESSION ANALYSIS

---

# MULTIPLE REGRESSION ANALYSIS

- Why need to know?
  - Many practical situations involve analyzing the relationships among three or more variables.
  - Example: an automobile manufacturer would be interested in the relationship between her company's automobile sales and the variables that influence those sales such as **competitors' sales**, and **advertising**, as well as economic variables such as **disposable personal income**, the **inflation rate**, and the **unemployment rate**.
- When multiple independent variables **are to be included in an analysis simultaneously**, multiple linear regression is very useful.

## Multiple Regression Model Population

$$y = \beta_0 + \beta_1x_1 + \beta_2x_2 + \cdots + \beta_kx_k + \varepsilon$$

where:

$\beta_0$  = Population's regression constant

$\beta_j$  = Population's regression coefficient for each variable  $x_j = 1, 2, \dots, k$

$k$  = Number of independent variables

$\varepsilon$  = Model error

### Estimated Multiple Regression Model

$$\hat{y} = b_0 + b_1x_1 + b_2x_2 + \cdots + b_kx_k$$



## Simple Linear Regression Model

$$\hat{y} = \beta_0 + \beta_1 x$$

1. estimated simple regression is a equation for **a straight line** in a two-dimensional space

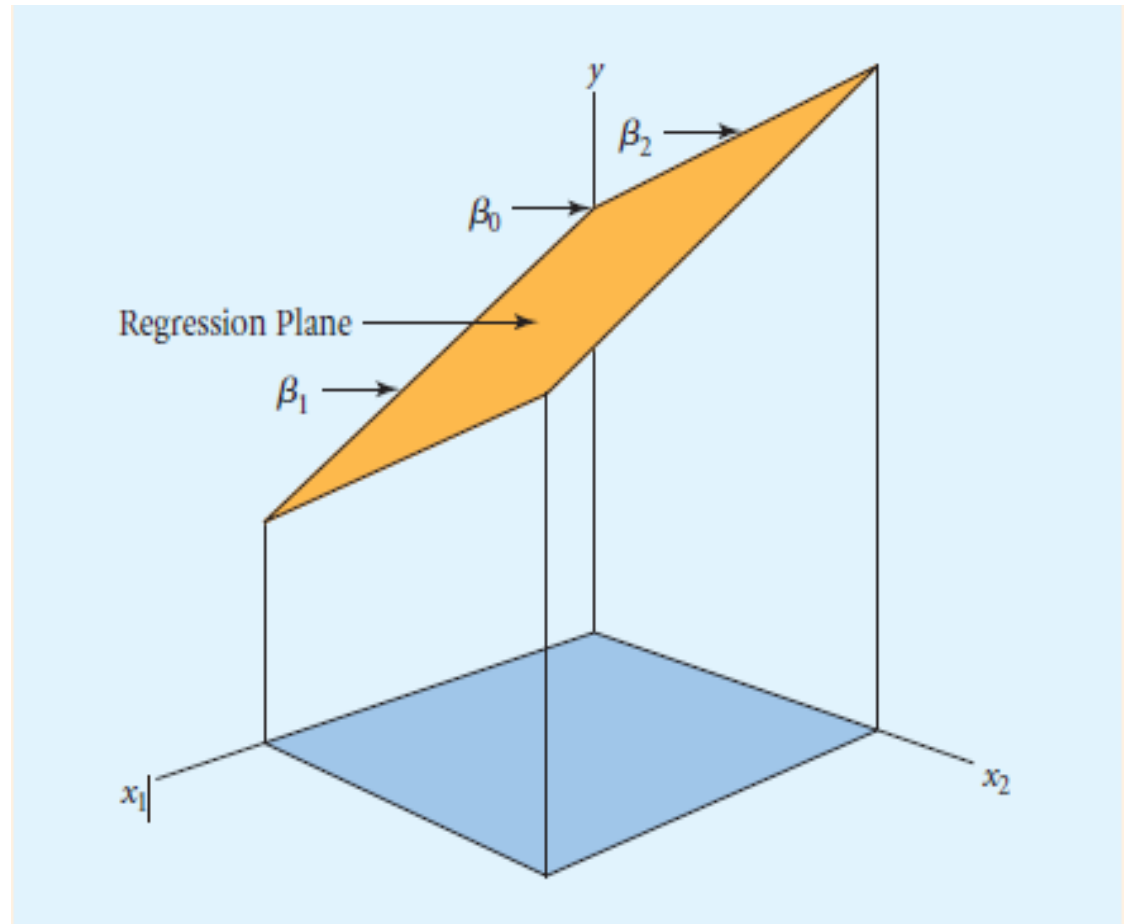
## Multiple Regression Model

$$\hat{y} = \beta_0 + \beta_1 x + \beta_2 x_2 + \beta_3 x_3 + \dots + \beta_k x_k$$

1. estimated multiple regression model forms a **hyperplane** (or response surface) through multidimensional space.
2. **regression hyperplane** represents the relationship between the dependent variable and the  $k$  independent variables.

# MULTIPLE REGRESSION ANALYSIS

Figure: Multiple Regression Hyperplane for Population



\*\* Multiple Regression analysis is usually performed with the aid of a computer and appropriate software.

# BASIC MODEL BUILDING CONCEPT

---

STATISTICAL MODEL USING MULTIPLE  
REGRESSION ANALYSIS

## STATISTICAL MODEL BUILDING CONCEPT



## What is MODEL?

A representation of an actual system using either physical or mathematical portrayal

Statistical Model-building process consisting 3 components:

1. Model Specification
2. Model Building
3. Model Diagnosis



## MODEL SPESIFICATION

- The process included are:
  - Determine the dependent variable
  - Decide which independent variables to be included in the model
  - Obtain the sample data for all variables.



## MODEL BUILDING

- The process of constructing a mathematical equation in which some or all of the independent variables are used to explain the variation in the dependent variable.

## MODEL DIAGNOSIS

- The process of analyzing the quality of the model you have constructed by determining how well a specified model fits the data you just gathered.

## MODEL DIAGNOSIS

- The objective of model building is to help you make better decisions.
  - Sophisticated model does not necessary will produce acceptable result
- The process included are:
  - Examine output values. For example: examine the output value  $R$ -squared and the standard error of the model.
  - Assess the extent to which the model's assumptions appear to be satisfied.



# EXAMPLE: Developing Multiple Regression Model

## Situation:

First City Real Estate executives wish to build a model to predict sales prices for residential property. Such a model will be valuable when working with potential sellers who might list their homes with First City.

### 1. Model Specification:

The response (dependent variable)  $\rightarrow y$  = Prices sales for residential property

The managers selected the following variables as good candidates:

$x_1$  = Home size (in square feet)

$x_2$  = Age of house

$x_3$  = Number of bedrooms

$x_4$  = Number of bathrooms

$x_5$  = Garage size (number of cars)

Data were obtained for a sample of 319 residential properties that had sold within the previous two months in an area served by two of First City's offices. For each house in the sample, the sales price and values for each potential independent variable were collected. The data are in the file **First City**.

# EXAMPLE: Developing Multiple Regression Model

## 2. Model Building:

There is **NO WAY** to determine whether an independent variable will be a good predictor variable by analyzing the individual variable's **descriptive statistics**.

Instead, need to look at the correlation between the independent variables and the dependent variable, which is measured by the **correlation coefficient**. For multiple variables, use **correlation matrix**.

```
> library(readxl)
> FirstCityNew <- read_excel("D:/AIS_STUFF/SUBJEK PENGAJARAN/MANB1123_Business Stat for DS/excel_data/FirstCityNew.xlsx")
> View(FirstCityNew)
> cor(FirstCityNew)
```

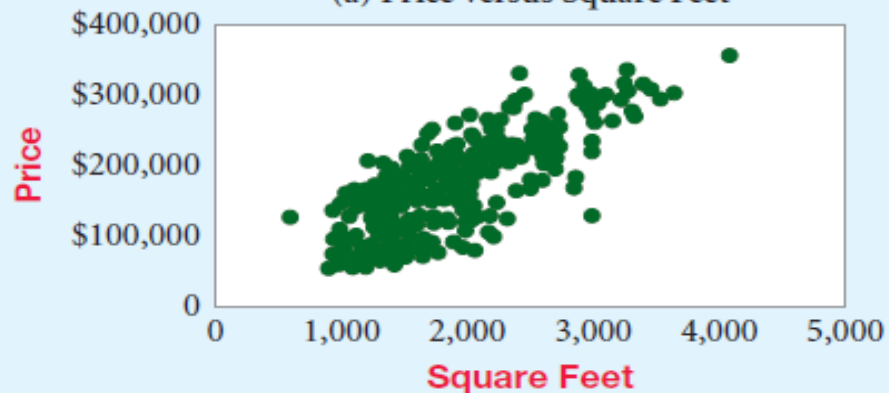
	Price	Sq. Feet	Age	Bedrooms	Bathrooms	Garage
Price	1.0000000	0.74771197	-0.48522184	0.5400880	0.6655043	0.6935385
Sq. Feet	0.7477120	1.00000000	-0.07288341	0.7058603	0.6292896	0.4162613
Age	-0.4852218	-0.07288341	1.00000000	-0.2024017	-0.3871049	-0.4373795
Bedrooms	0.5400880	0.70586025	-0.20240165	1.00000000	0.5996403	0.3120343
Bathrooms	0.6655043	0.62928955	-0.38710488	0.5996403	1.00000000	0.4646015
Garage	0.6935385	0.41626129	-0.43737948	0.3120343	0.4646015	1.00000000

### Correlation Matrix:

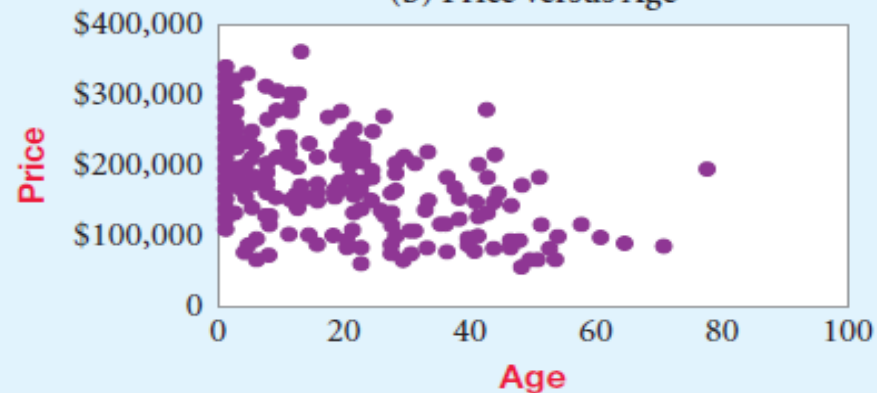
A table showing the pairwise correlation between all variables

```
> |
```

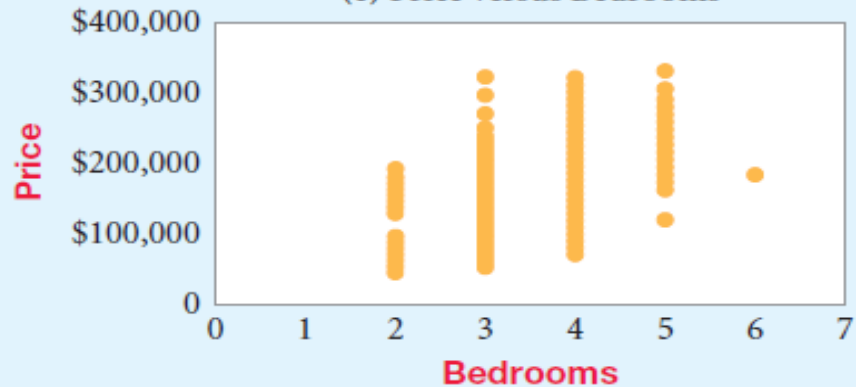
(a) Price versus Square Feet



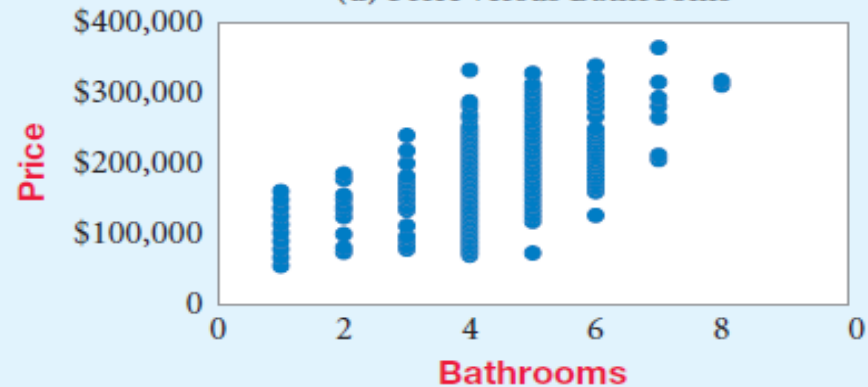
(b) Price versus Age



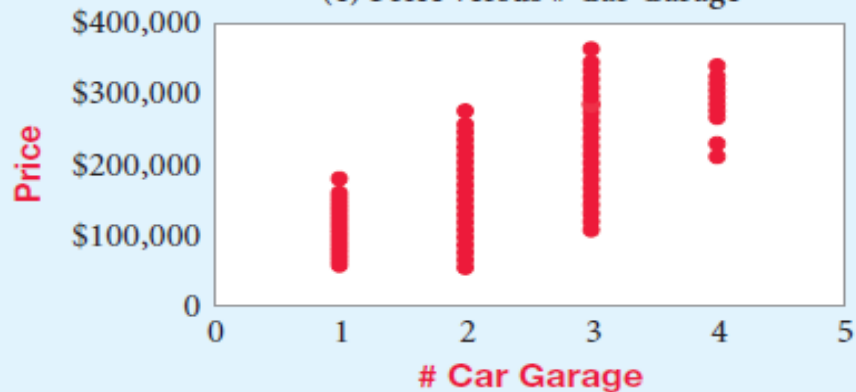
(c) Price versus Bedrooms



(d) Price versus Bathrooms



(e) Price versus # Car Garage



# EXAMPLE: Developing Multiple Regression Model

## Significance Test for Correlation Coefficient:

REMEMBER: the purpose of conducting the test is to identify either the variable (x,y) is correlated to each other or not.

$H_0: \rho = 0$  (no correlation)

$H_A: \rho \neq 0$  (correlation exists)

The test is conducted with a significance level of  $\alpha = 0.05$

The degree of freedom is:  $n-2 = 319 - 2 = 317$

The critical t (see t-table) for two-tailed test is approximately  $\pm 1.96$ .

Decision rule: any t-value which is greater than 1.96 and smaller than -1.96,  $H_0$  will be rejected.

or look at p-value, which if less than significance level of  $\alpha = 0.05$ ,  $H_0$  will be rejected.

```
> library(readxl)
Warning message:
package 'readxl' was built under R version 3.2.5
> FirstCityNew <- read_excel("D:/AIS_STUFF/SUBJEK PENGAJARAN/MANB1123_Business Stat for DS/excel_data/FirstCityNew.xlsx")
> View(FirstCityNew)
> FCModel = lm(Price~Sq.Feet+Age+Bedrooms+Bathrooms+Garage,data = FirstCityNew)
> summary(FCModel)
```

```
Call:
lm(formula = Price ~ Sq.Feet + Age + Bedrooms + Bathrooms + Garage,
    data = FirstCityNew)
```

Residuals:

```
      Min       1Q   Median       3Q      Max
-106752  -15052    2587   17602   77565
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	31127.602	9539.669	3.263	0.00122 **
Sq.Feet	63.066	4.017	15.700	< 2e-16 ***
Age	-1144.437	112.780	-10.148	< 2e-16 ***
Bedrooms	-8410.379	3002.511	-2.801	0.00541 **
Bathrooms	3521.954	1580.997	2.228	0.02661 *
Garage	28203.542	2858.692	9.866	< 2e-16 ***

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 27350 on 313 degrees of freedom
Multiple R-squared:  0.8161,    Adjusted R-squared:  0.8131
F-statistic: 277.8 on 5 and 313 DF,  p-value: < 2.2e-16
```


```
>
```

e.g: The correlation between sales price and house square feet is statistically significant

The results indicates that a significant linear relationship between each independent variable and sales price.

# EXAMPLE: Developing Multiple Regression Model

## Computing the regression equation:

```
Console ~/ 
> FCModel = lm(Price~Sq.Feet+Age+Bedrooms+Bathrooms+Garage,data = FirstCityNew)
> summary(FCModel)
```

Call:  
lm(formula = Price ~ Sq.Feet + Age + Bedrooms + Bathrooms + Garage,  
data = FirstCityNew)

Residuals:

	Min	1Q	Median	3Q	Max
	-106752	-15052	2587	17602	77565

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )	
(Intercept)	31127.602	9539.669	3.263	0.00122	**
Sq.Feet	63.066	4.017	15.700	< 2e-16	***
Age	-1144.437	112.780	-10.148	< 2e-16	***
Bedrooms	-8410.379	3002.511	-2.801	0.00541	**
Bathrooms	3521.954	1580.997	2.228	0.02661	*
Garage	28203.542	2858.692	9.866	< 2e-16	***

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 27350 on 313 degrees of freedom  
Multiple R-squared: 0.8161, Adjusted R-squared: 0.8131  
F-statistic: 277.8 on 5 and 313 DF, p-value: < 2.2e-16

```
> options(show.signif.stars = F)
> anova(FCModel)
```

Analysis of Variance Table

Response: Price	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Sq.Feet	1	7.1172e+11	7.1172e+11	951.449	< 2.2e-16
Age	1	2.3744e+11	2.3744e+11	317.418	< 2.2e-16
Bedrooms	1	7.4847e+09	7.4847e+09	10.006	0.0017136
Bathrooms	1	9.4429e+09	9.4429e+09	12.624	0.0004396
Garage	1	7.2811e+10	7.2811e+10	97.336	< 2.2e-16
Residuals	313	2.3414e+11	7.4804e+08		



# EXAMPLE: Developing Multiple Regression Model

## Computing the regression equation:

The estimate of the multiple regression model are:

$$\hat{y} = 31127.6 + 63.1(sq. feet) - 1144.4(age) - 8410.4(bedroom) + 3522.0(bathroom) + 28203.5(garage)$$


The coefficients for each independent variable represent an estimate of the average change in the dependent variable for a 1-unit change in the independent variable.

For example, for houses of the same age, with the same number of bedrooms, baths, and garages, a 1-square-foot increase in the size of the house is estimated to increase its price by an average of \$63.10

## Computing the multiple coefficient of Determination:

Is used to determine the proportion of variation in the dependent variable that is explained by the dependent variable's relationship to all the independent variables in the model.

# EXAMPLE: Developing Multiple Regression Model

```
Console ~/ 
> FCModel = lm(Price~Sq.Feet+Age+Bedrooms+Bathrooms+Garage,data = FirstCityNew)
> summary(FCModel)
```

```
Call:
lm(formula = Price ~ Sq.Feet + Age + Bedrooms + Bathrooms + Garage,
    data = FirstCityNew)
```

```
Residuals:
    Min       1Q   Median       3Q      Max
-106752  -15052    2587   17602   77565
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 31127.602   9539.669   3.263  0.00122 **
Sq.Feet      63.066     4.017  15.700 < 2e-16 ***
Age        -1144.437   112.780 -10.148 < 2e-16 ***
Bedrooms    -8410.379  3002.511  -2.801  0.00541 **
Bathrooms    3521.954  1580.997   2.228  0.02661 *
Garage      28203.542  2858.692   9.866 < 2e-16 ***
```

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 27350 on 313 degrees of freedom
Multiple R-squared:  0.8161, Adjusted R-squared:  0.8131
F-statistic: 277.8 on 5 and 313 DF,  p-value: < 2.2e-16
```

```
> options(show.signif.stars = F)
> anova(FCModel)
Analysis of Variance Table
```

```
Response: Price
      Df Sum Sq Mean Sq F value    Pr(>F)
Sq.Feet  1 7.1172e+11 7.1172e+11 951.449 < 2.2e-16
Age       1 2.3744e+11 2.3744e+11 317.418 < 2.2e-16
Bedrooms  1 7.4847e+09 7.4847e+09  10.006 0.0017136
Bathrooms  1 9.4429e+09 9.4429e+09  12.624 0.0004396
Garage     1 7.2811e+10 7.2811e+10  97.336 < 2.2e-16
Residuals 313 2.3414e+11 7.4804e+08
>
```

## Summary from $R^2$ value

More than 82% of the variation in sales price can be explained by the linear relationship of the five independent variables in the regression model to the dependent variable.

$R^2$  value

# EXAMPLE: Developing Multiple Regression Model

## 2. Model Diagnosis:

Why need to diagnosis the regression model?:

To determine how well the regression model perform.

Statistics measurement need to perform for model diagnosis are:

- i. Analyzing R-squared ( $R^2$ )- multiple coefficient of determination
- ii. Analyzing adjusted R-Squared
- iii. Analyzing standard error of the estimate

Several questions which normally asked during model diagnosis:

- a) Is the overall model significant?
- b) Are the individual variables significant?
- c) Is the standard deviation of the model error too large to provide meaningful results?

# EXAMPLE: Developing Multiple Regression Model

## 2. Model Diagnosis:

a) Is the model significant?

Because the regression model is constructed based on a sample of data from the population, therefore it is subject to sampling error. Thus, it is need to test the statistical significance of the overall regression model.


The hypothesis statement:

$$H_0: \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$$

$$H_A: \text{At least one } \beta_i \neq 0$$

The  $F$ -test is a method for testing whether the regression model explains a significant proportion of the variation in the dependent variable (and whether the overall model is significant).

# EXAMPLE: Developing Multiple Regression Model

```
Console ~/   
> FCModel = lm(Price~Sq.Feet+Age+Bedrooms+Bathrooms+Garage,data = FirstCityNew)  
> summary(FCModel)
```

```
Call:  
lm(formula = Price ~ Sq.Feet + Age + Bedrooms + Bathrooms + Garage,  
    data = FirstCityNew)
```

```
Residuals:  
      Min       1Q   Median       3Q      Max  
-106752  -15052    2587   17602   77565
```

```
Coefficients:  
              Estimate Std. Error t value Pr(>|t|)  
(Intercept) 31127.602   9539.669   3.263  0.00122 **  
Sq.Feet      63.066     4.017  15.700 < 2e-16 ***  
Age         -1144.437   112.780 -10.148 < 2e-16 ***  
Bedrooms    -8410.379   3002.511  -2.801  0.00541 **  
Bathrooms    3521.954   1580.997   2.228  0.02661 *  
Garage      28203.542   2858.692   9.866 < 2e-16 ***
```

```
---  
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 27350 on 313 degrees of freedom  
Multiple R-squared:  0.8161,    Adjusted R-squared:  0.8131  
F-statistic: 277.8 on 5 and 313 DF,  p-value: < 2.2e-16
```

```
> options(show.signif.stars = F)  
> anova(FCModel)  
Analysis of Variance Table
```

```
Response: Price  
      Df    Sum Sq   Mean Sq F value    Pr(>F)  
Sq.Feet  1 7.1172e+11 7.1172e+11  951.449 < 2.2e-16  
Age      1 2.3744e+11 2.3744e+11  317.418 < 2.2e-16  
Bedrooms 1 7.4847e+09 7.4847e+09  10.006 0.0017136  
Bathrooms 1 9.4429e+09 9.4429e+09  12.624 0.0004396  
Garage    1 7.2811e+10 7.2811e+10   97.336 < 2.2e-16  
Residuals 313 2.3414e+11 7.4804e+08  
>
```

## Summary from F-test

p-value = 0.000 which is  $< \alpha = 0.01$ , therefore  $H_0$  is rejected.

Conclusion: The regression model *does* explain a significant proportion of the variation in sales price. Thus, the overall model is statistically significant. This means we can conclude that at least one of the regression slope coefficients is not equal to zero.

# EXAMPLE: Developing Multiple Regression Model

## 2. Model Diagnosis:

**Adjusted R-squared ( $R_A^2$ )** - A measure of the percentage of explained variation in the dependent variable that takes into account the relationship between the sample size and the number of independent variables in the regression model.

Residual standard error: 27350 on 313 degrees of freedom

Multiple R-squared: 0.8161, Adjusted R-squared: 0.8131

F-statistic: 277.8 on 5 and 313 DF, p-value: < 2.2e-16

the adjusted  $R$ -squared is 81.3%, only slightly less than  $R^2$  81.6%.

# EXAMPLE: Developing Multiple Regression Model

## 2. Model Diagnosis:

b) Are the individual variables significant?

We have concluded that the overall model is significant. This means *at least* one independent variable explains a significant proportion of the variation in sales price.

To determine which variables are significant, we test the following hypotheses:


$$H_0: \beta_j = 0$$

$$H_A: \beta_j \neq 0 \text{ for all } j$$

*The significant test is conducted using significance level  $\alpha = 0.05$*

From the computer output, we may compare the p-value for each regression slope coefficient with significance level,  $\alpha$ . If the p-value is **less than alpha**, we **reject the null hypothesis** and conclude that **the independent variable is statistically significant in the model**.

# EXAMPLE: Developing Multiple Regression Model

```
Console ~/   
> FCModel = lm(Price~Sq.Feet+Age+Bedrooms+Bathrooms+Garage,data = FirstCityNew)  
> summary(FCModel)
```

```
Call:  
lm(formula = Price ~ Sq.Feet + Age + Bedrooms + Bathrooms + Garage,  
    data = FirstCityNew)
```

```
Residuals:  
      Min       1Q   Median       3Q      Max  
-106752  -15052    2587   17602   77565
```

```
Coefficients:  
              Estimate Std. Error t value Pr(>|t|)  
(Intercept) 31127.602   9539.669   3.263  0.00122 ***  
Sq.Feet      63.066     4.017   15.700 < 2e-16 ***  
Age        -1144.437   112.780  -10.148 < 2e-16 ***  
Bedrooms    -8410.379  3002.511   -2.801  0.00541 **  
Bathrooms    3521.954  1580.997    2.228  0.02661 *  
Garage      28203.542  2858.692    9.866 < 2e-16 ***
```

```
---  
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 27350 on 313 degrees of freedom  
Multiple R-squared:  0.8161,    Adjusted R-squared:  0.8131  
F-statistic: 277.8 on 5 and 313 DF,  p-value: < 2.2e-16
```

```
> options(show.signif.stars = F)  
> anova(FCModel)  
Analysis of Variance Table
```

```
Response: Price  
      Df    Sum Sq   Mean Sq F value    Pr(>F)  
Sq.Feet  1 7.1172e+11 7.1172e+11  951.449 < 2.2e-16  
Age      1 2.3744e+11 2.3744e+11  317.418 < 2.2e-16  
Bedrooms 1 7.4847e+09 7.4847e+09  10.006 0.0017136  
Bathrooms 1 9.4429e+09 9.4429e+09  12.624 0.0004396  
Garage    1 7.2811e+10 7.2811e+10   97.336 < 2.2e-16  
Residuals 313 2.3414e+11 7.4804e+08  
>
```

## Summary from t- test/p-value

We conclude that all five independent variables in the model are significant.



# EXAMPLE: Developing Multiple Regression Model

## 2. Model Diagnosis:

- c) Is the standard deviation of the model error too large to provide meaningful results?

The standard deviation of the regression model (also called the *standard error of the estimate or residual error*), measures the dispersion of observed dependent variable,  $y$ , around values predicted by the regression model.

Sometimes, even though a model has a high  $R^2$ , the standard error of the estimate will be too large to provide adequate precision for confidence and prediction intervals.

The rough prediction range for standard error of the estimate is  $\pm 2s_e$

# EXAMPLE: Developing Multiple Regression Model

## 2. Model Diagnosis:

From the First City Real Estate Company example:

Residual standard error: 27350 on 313 degrees of freedom  
Multiple R-squared: 0.8161, Adjusted R-squared: 0.8131  
F-statistic: 277.8 on 5 and 313 DF, p-value: < 2.2e-16

The standard error = \$27,350

Following the rule of thumb  $\pm 2s_e$ ,  $= 2(27,350) = \pm \$54,700$ .

Thus, the rough prediction range for the price of an individual home is  $\pm \$54,700$ .

# USING QUALITATIVE INDEPENDENT VARIABLES

---

# USING QUALITATIVE INDEPENDENT VARIABLE

- There is situation in which you may wish to use a qualitative, lower level variable as an explanatory variable in a regression model
  - Example: use of variable such as; marital status, gender, education level or job performance.
- How these variable can be incorporated into a multiple regression analysis?
  - Dummy (or indicator) variable – a variable that is assigned a value equal to either 0 or 1, depending on whether the observation possesses a given characteristic.

$x_1 = 1$  if female

$x_1 = 0$  if male

if more than two mutually exclusive (for example, never married, married, divorced)

$x_1 = 1$  if never married, 0 if not

$x_2 = 1$  if married, 0 if not

$x_3 = 1$  if divorced, 0 if not

# USING QUALITATIVE INDEPENDENT VARIABLE

Example:

The population from which the sample was selected consists of executives between the ages of 24 and 60 who are working in U.S. manufacturing businesses. Data for annual salary ( $y$ ) and age ( $x_1$ ) are available. The objective is to determine whether a model can be generated to explain the variation in annual salary for business executives. Even though age and annual salary are significantly correlated ( $r = 0.686$ ) at the  $\alpha = 0.05$  level, the coefficient of determination is only 47%. Suppose we can determine which of the 16 people in the sample had a master of business administration (MBA) degree.

$x_2 = 1$  if holds MBA degree

$x_2 = 0$  if did not hold MBA degree

**TABLE 15.2** | Executive Salary Data Including MBA Variable

Salary(\$)	Age	MBA
65,000	26	0
85,000	28	1
74,000	36	0
83,000	35	0
110,000	35	1
160,000	40	1
100,000	41	0
122,000	42	1
85,000	45	0
120,000	46	1
105,000	50	0
135,000	51	1
125,000	55	0
175,000	50	1
156,000	61	1
140,000	63	0

# USING QUALITATIVE INDEPENDENT VARIABLE

From the statistical packages, the estimated regression equation are:

$$\hat{y} = 6,974 + 2,055x_1 + 35,236x_2$$

Because the dummy variable,  $x_2$ , has been coded 0 or 1 depending on MBA status, incorporating it into the regression model is like having two simple linear regression lines with the same slope, but different intercept.

For instance when  $x_2 = 1$  (respondent who holds MBA degree),

$$\hat{y} = 6,974 + 2,055x_1 + 35,236(1)$$

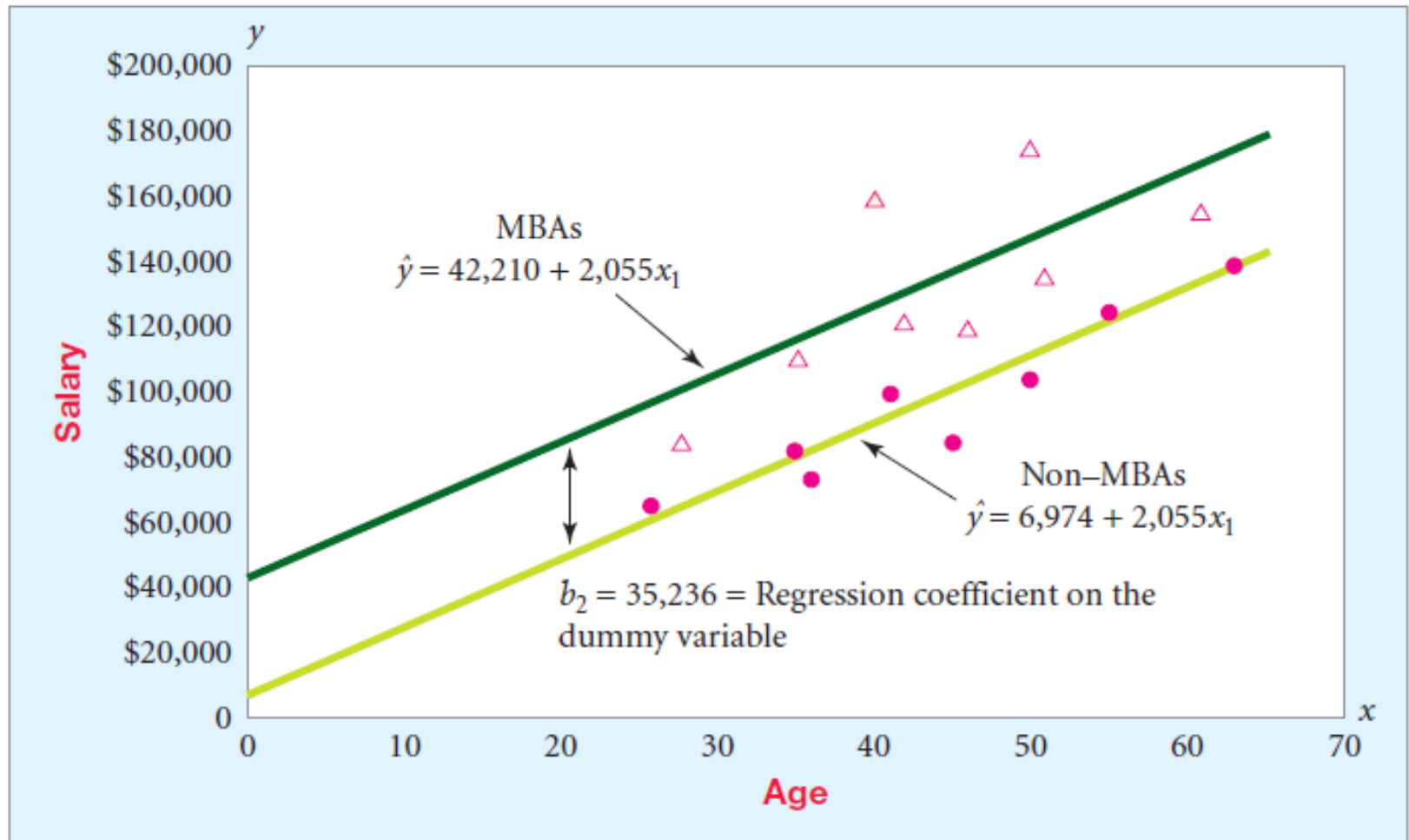
$$\hat{y} = 42,210 + 2,055x_1$$

and when  $x_2 = 0$  (respondent who did not holds MBA degree),

$$\hat{y} = 6,974 + 2,055x_1 + 35,236(0)$$

$$\hat{y} = 6,974 + 2,055x_1$$

# USING QUALITATIVE INDEPENDENT VARIABLE



# USING QUALITATIVE INDEPENDENT VARIABLE

Example: FIRST CITY REAL ESTATE COMPANY (continue)

The regression model developed (slide #17) showed potential because the overall model was statistically significant. The model explained nearly 82% of the variation in sales prices for the home in the sample (slide #18 and slide #22). All the independent variables were significant, given that the other independent variables were in the model (slide #24). However, the standard error of the estimate is quite high at \$27,350 (slide #26). The manager have decided to improve the model. They decided to add new variable: area. This variable is a categorical variable with two possible outcome either foothills or not foothills. The revised data is on sheet 2 name **Homes-Sample 2**. Perform model building and model diagnosis for the sales price with incorporating new variable.

$$x_6(\text{area}) = 1 \text{ if foothills, } 0 \text{ if not}$$



# WORKING WITH NONLINEAR RELATIONSHIP

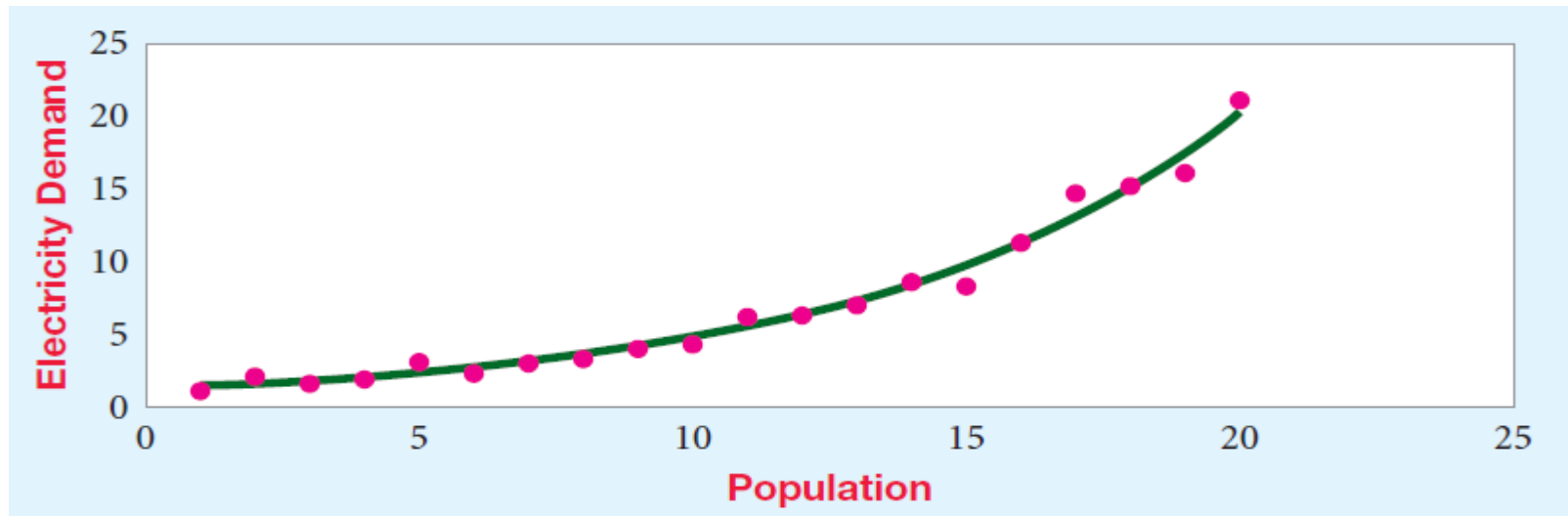
---

# NONLINEAR RELATIONSHIP

- There are also many instances in which the relationship between two variables will be curvilinear, rather than linear.

Example:

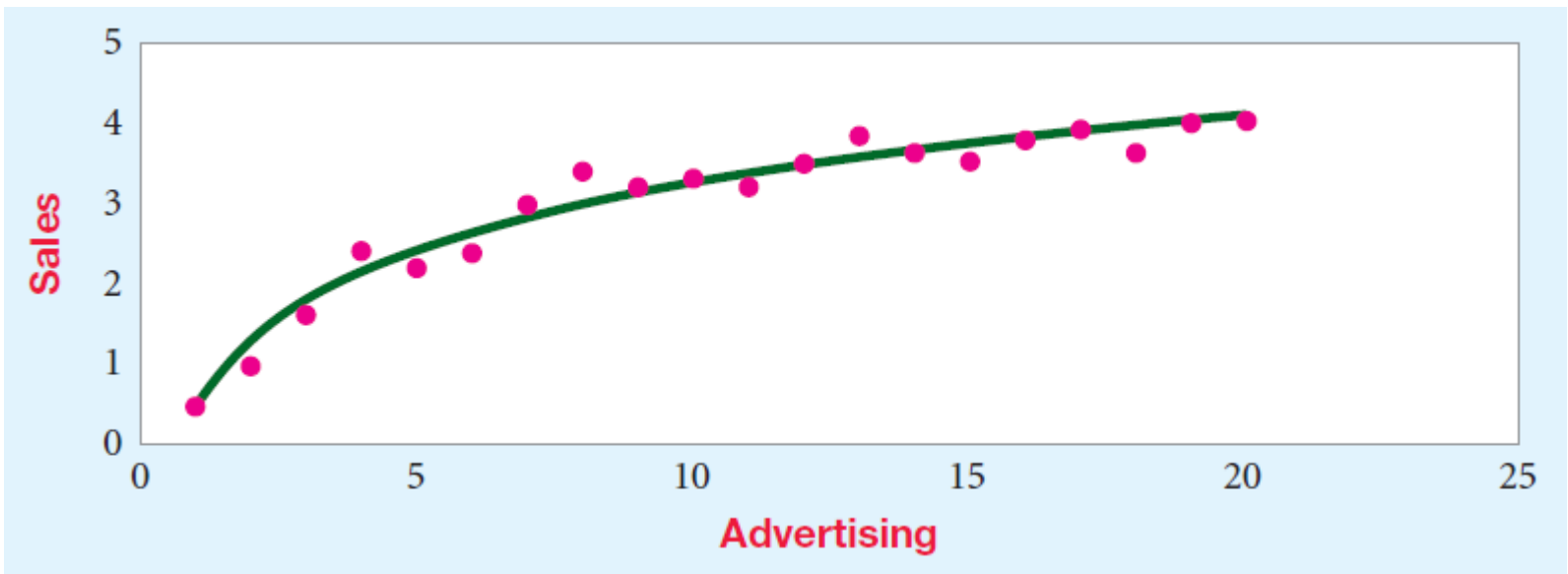
- Demand for electricity has grown at an almost exponential rate relative to the population growth in some areas.



# NONLINEAR RELATIONSHIP

## Example:

- Advertisers believe that a diminishing returns relationship will occur between sales and advertising if advertising is allowed to grow too large.



# NONLINEAR RELATIONSHIP

- To model such curvilinear relationships, we must incorporate terms into the multiple regression model that will create “*curves*” in the model we are building.
- The model which possesses the curvilinear is refer as a *polynomial model*. The general equation for a polynomial with one independent variable is given as below

$$y = \beta_0 + \beta_1x + \beta_2x^2 + \cdots + \beta_px^p + \varepsilon$$

where:

$\beta_0$  = Population regression's constant

$\beta_j$  = Population's regression coefficient for variable  $x^j$ ;  $j = 1, 2, \dots, p$

$p$  = Order (or degree) of the polynomial

$\varepsilon$  = Model error

# NONLINEAR RELATIONSHIP

$$y = \beta_0 + \beta_1x + \beta_2x^2 + \cdots + \beta_px^p + \varepsilon$$

where:

$\beta_0$  = Population regression's constant

$\beta_j$  = Population's regression coefficient for variable  $x^j$ ;  $j = 1, 2, \dots, p$

$p$  = Order (or degree) of the polynomial

$\varepsilon$  = Model error

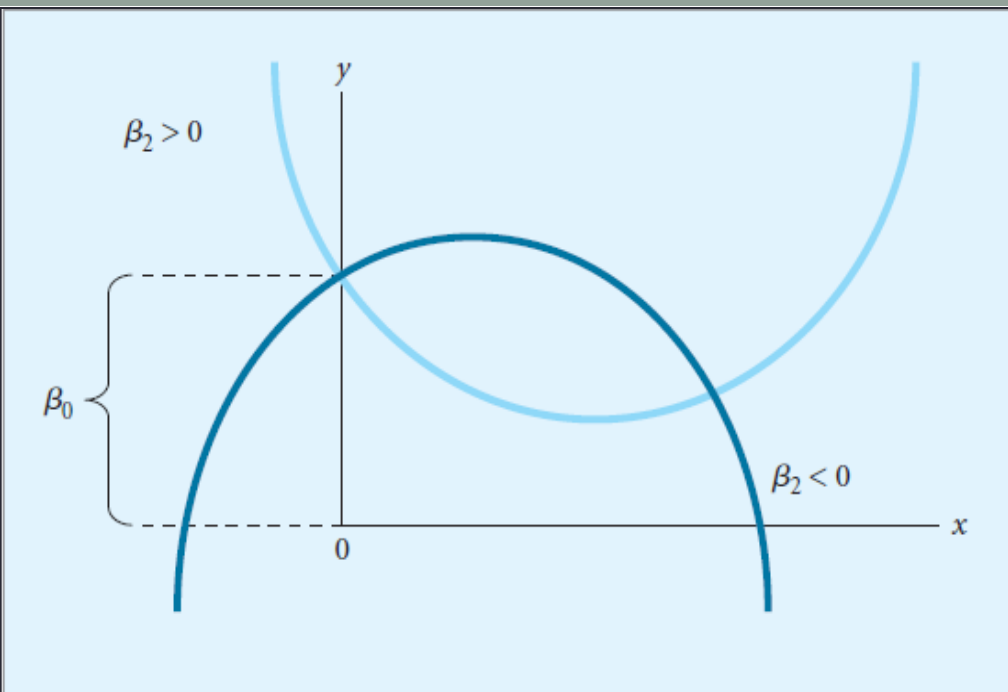
The order/degree of the model is determined by the largest exponent of the independent variable in the model.

$$y = \beta_0 + \beta_1x + \beta_2x^2 + \varepsilon$$

The model above is a second-order polynomial because the largest exponent in any term of the polynomial is 2.

As more curves appear in the data, the order of the polynomial must be increased. Example of third-order polynomial

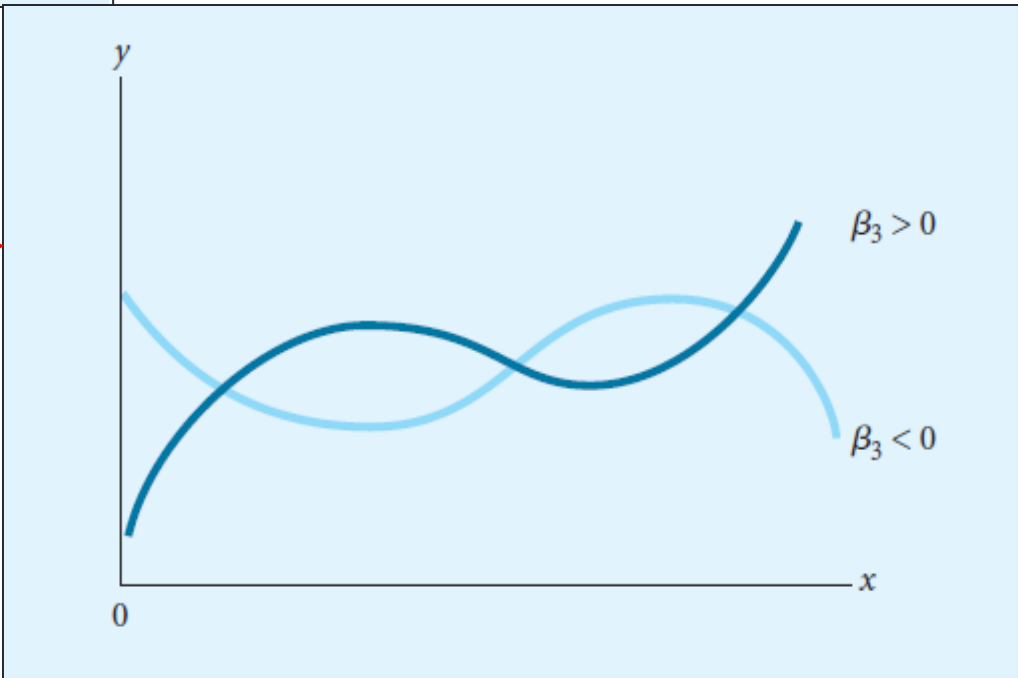
$$y = \beta_0 + \beta_1x + \beta_2x^2 + \beta_3x^3 + \varepsilon$$



Second-order Regression Model



Third-order Regression Model



# NONLINEAR RELATIONSHIP

Example:

Ashley Investment Services was severely shaken by the downturn in the stock market during the summer and fall of 2008. To maintain profitability and save as many jobs as possible, since then everyone has been extra busy analyzing new investment opportunities. The director of personnel has noticed an increased number of people suffering from “burnout,” in which physical and emotional fatigue hurt job performance. Although he cannot change the job’s pressures, he has read that the more time a person spends socializing with coworkers away from the job, the more likely there is to be a higher degree of burnout. With the help of the human resources lab at the local university, the personnel director has administered a questionnaire to company employees. A burnout index has been computed from the responses to the survey. Likewise, the survey responses are used to determine quantitative measures of socialization. Sample data from questionnaires are contained in the file **Ashley**.

# NONLINEAR RELATIONSHIP

To model the relationship between the socialization index and the burnout index for Ashley employees these steps can be followed:

**Step 1: Specify the model by determining the dependent and potential independent variables.**

The dependent variable is the burnout index. The company wishes to explain the variation in burnout level. One potential independent variable is the socialization index.

**Step 2: Formulate the model.**

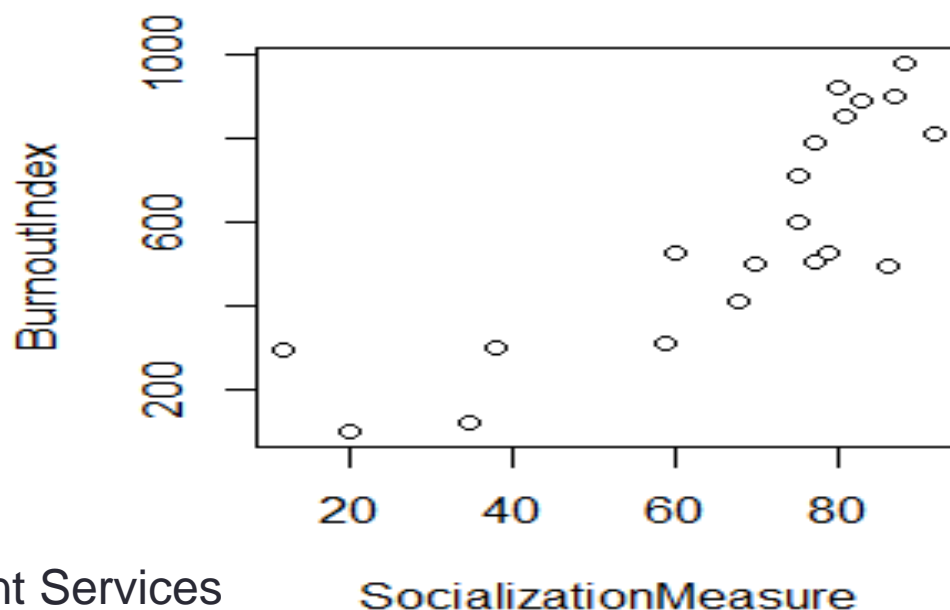
Begin by proposing that a linear relationship exists between the two variables.



```
> library(readxl)
> Ashley <- read_excel("D:/AIS_STUFF/SUBJEK PENGAJARAN/MANB1123_Business Stat for DS/excel_data/Ashley.xlsx")
> view(Ashley)
> plot(BurnoutIndex~SocializationMeasure,data = Ashley)
> with(Ashley,cor.test(BurnoutIndex,SocializationMeasure))
```

Pearson's product-moment correlation

data: BurnoutIndex and SocializationMeasure  
t = 6.0357, df = 18, p-value = 1.048e-05  
alternative hypothesis: true correlation is not equal to 0  
95 percent confidence interval:  
0.5887197 0.9255510  
sample estimates:  
cor  
0.8181067



Scatter plot for Ashley Investment Services

## Develop the estimate regression equation/model for Ashley Investment Services

```
> ashelyModel = lm(BurnoutIndex~SocializationMeasure,data = Ashley)
> summary(ashelyModel)
```

```
Call:
lm(formula = BurnoutIndex ~ SocializationMeasure, data = Ashley)
```

```
Residuals:
    Min       1Q   Median       3Q      Max
-265.480 -153.175  -2.113  135.065  247.097
```

```
Coefficients:
```

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	-66.164	112.444	-0.588	0.564
SocializationMeasure	9.589	1.589	6.036	1.05e-05 ***

---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

```
Residual standard error: 160 on 18 degrees of freedom
Multiple R-squared:  0.6693,    Adjusted R-squared:  0.6509
F-statistic: 36.43 on 1 and 18 DF,  p-value: 1.048e-05
```

```
> options(show.signif.stars = F)
```

```
> anova(ashelyModel)
```

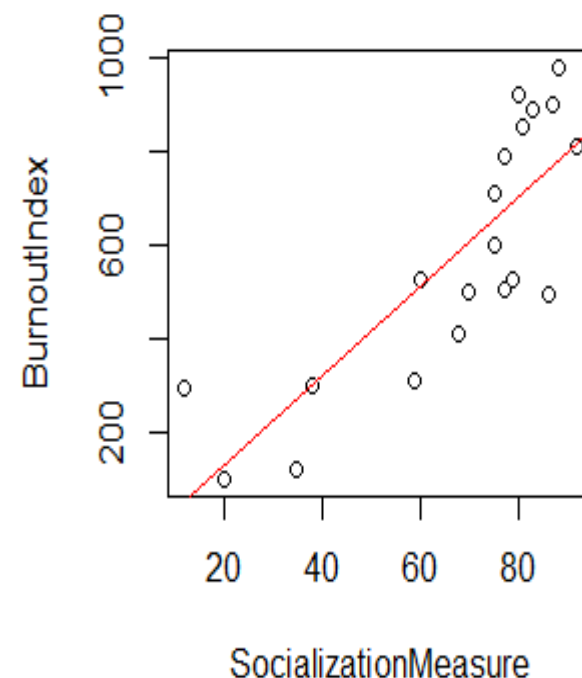
```
Analysis of Variance Table
```

```
Response: BurnoutIndex
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
SocializationMeasure	1	932504	932504	36.43	1.048e-05
Residuals	18	460752	25597		

```
> |
```

**Regression Line Plot**



$$\hat{y} = -66.14 + 9.589x$$

# NONLINEAR RELATIONSHIP

- **From the regression line plot:** The line appears to fit the data. However, a closer inspection reveals instances where several consecutive points lie above or below the line. The points are not randomly dispersed around the regression line

## **Step 3: Perform Diagnosis check on the model.**

Can use an  $F$ -test to test whether a regression model explains a significant amount of variation in the dependent variable.

$$H_0: \rho^2 = 0$$

$$H_A: \rho^2 \neq 0$$

```
> ashelyModel = lm(BurnoutIndex~SocializationMeasure,data = Ashley)
> summary(ashelyModel)
```

Call:

```
lm(formula = BurnoutIndex ~ SocializationMeasure, data = Ashley)
```

Residuals:

	Min	1Q	Median	3Q	Max
	-265.480	-153.175	-2.113	135.065	247.097

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	-66.164	112.444	-0.588	0.564
SocializationMeasure	9.589	1.589	6.036	1.05e-05 ***

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 160 on 18 degrees of freedom

Multiple R-squared: 0.6693, Adjusted R-squared: 0.6509

F-statistic: 36.43 on 1 and 18 DF, p-value: 1.048e-05

```
> options(show.signif.stars = F)
```

```
> anova(ashelyModel)
```

Analysis of Variance Table

Response: BurnoutIndex

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
SocializationMeasure	1	932504	932504	36.43	1.048e-05
Residuals	18	460752	25597		

```
> |
```

We conclude that the simple linear model is statistically significant. However, we should also examine the data to determine if any curvilinear relationships may be present.

# NONLINEAR RELATIONSHIP

## Step 4: Model the curvilinear relationship

One possible approach to model the curvilinear is with the use of polynomials. The second-order polynomial (quadratic model) will be as follow:

$$y = \beta_0 + \beta_1x + \beta_2x^2 + \varepsilon$$

From the computer output (slide #46), the estimated regression equation is:

$$\hat{y} = 265.68 - 6.837x + 0.154x^2$$

```
> attach(Ashley)
The following objects are masked from Ashley (pos = 3):

    BurnoutIndex, SocializationMeasure

> SocializationMeasure2=SocializationMeasure^2
> AshleyNModel=lm(BurnoutIndex~SocializationMeasure+SocializationMeasure2)
> summary(AshleyNModel)
```

```
Call:
lm(formula = BurnoutIndex ~ SocializationMeasure + SocializationMeasure2)
```

```
Residuals:
```

Min	1Q	Median	3Q	Max
-322.01	-96.53	23.67	118.25	217.13

```
Coefficients:
```

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	265.6805	184.3081	1.442	0.1676
SocializationMeasure	-6.8366	7.7210	-0.885	0.3883
SocializationMeasure2	0.1538	0.0710	2.166	0.0448

```
Residual standard error: 145.7 on 17 degrees of freedom
```

```
Multiple R-squared: 0.7408, Adjusted R-squared: 0.7103
```

```
F-statistic: 24.29 on 2 and 17 DF, p-value: 1.037e-05
```

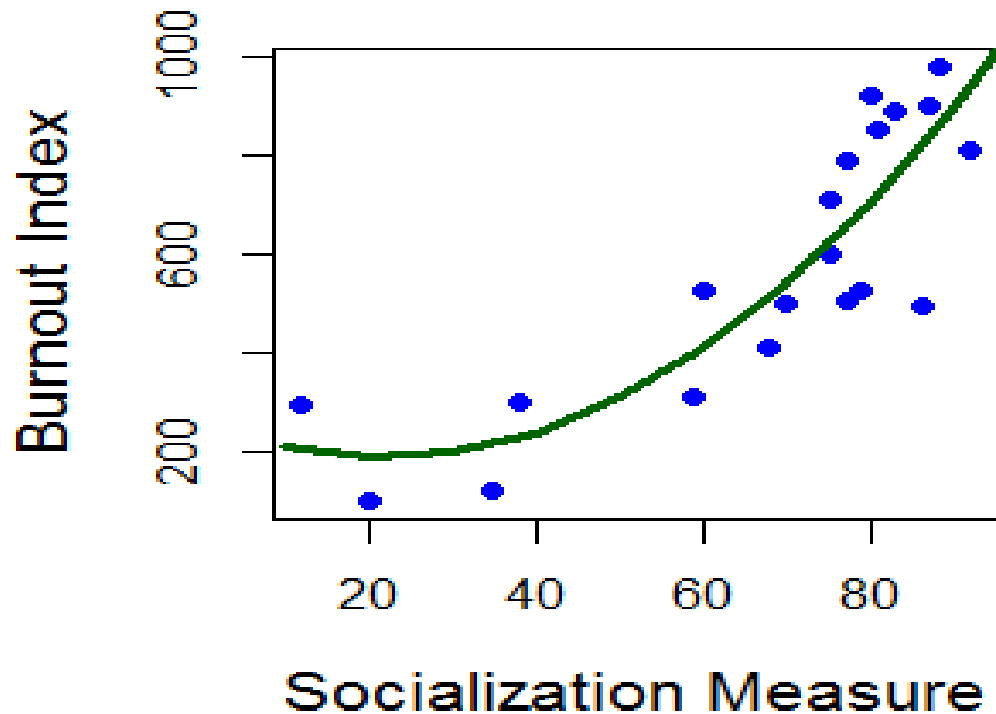
# NONLINEAR RELATIONSHIP

## **Step 5: Perform diagnosis on the revised curvilinear model**

From step 2, based on simple linear regression model, the  $R^2$  value is 0.6693 or 67%. However, based on revised model (when considering non-linear relationship), the  $R^2$  value is 0.7408 or 74%, which is more higher (see output on slide #46).

The second-order polynomial plot is depicted on the following slide (slide #48)

```
> timevalues <- seq(10, 100, 10)
> predictedcounts <- predict(AshleyNModel, list(SocializationMeasure=timevalues, SocializationMeasure2=timevalues^2))
> plot(SocializationMeasure, BurnoutIndex, pch=16, xlab = "Socialization Measure", ylab = "Burnout Index", cex.lab = 1.3, col
= "blue")
> lines(timevalues, predictedcounts, col = "darkgreen", lwd = 3)
>
```





# THE END

---

HAVE A FURTHER READING ON NON-LINEAR  
REGRESSION 😊