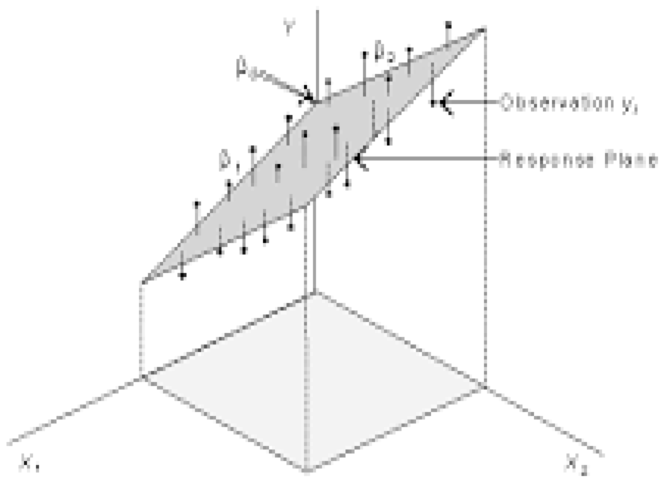
MULTIPLE REGRESSION ANALYSIS AND MODEL BUILDING



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

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon$$

where:

β_0 = Population's regression constant

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

k = Number of independent variables

ε = Model error

Estimated Multiple Regression Model

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_k x_k$$

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

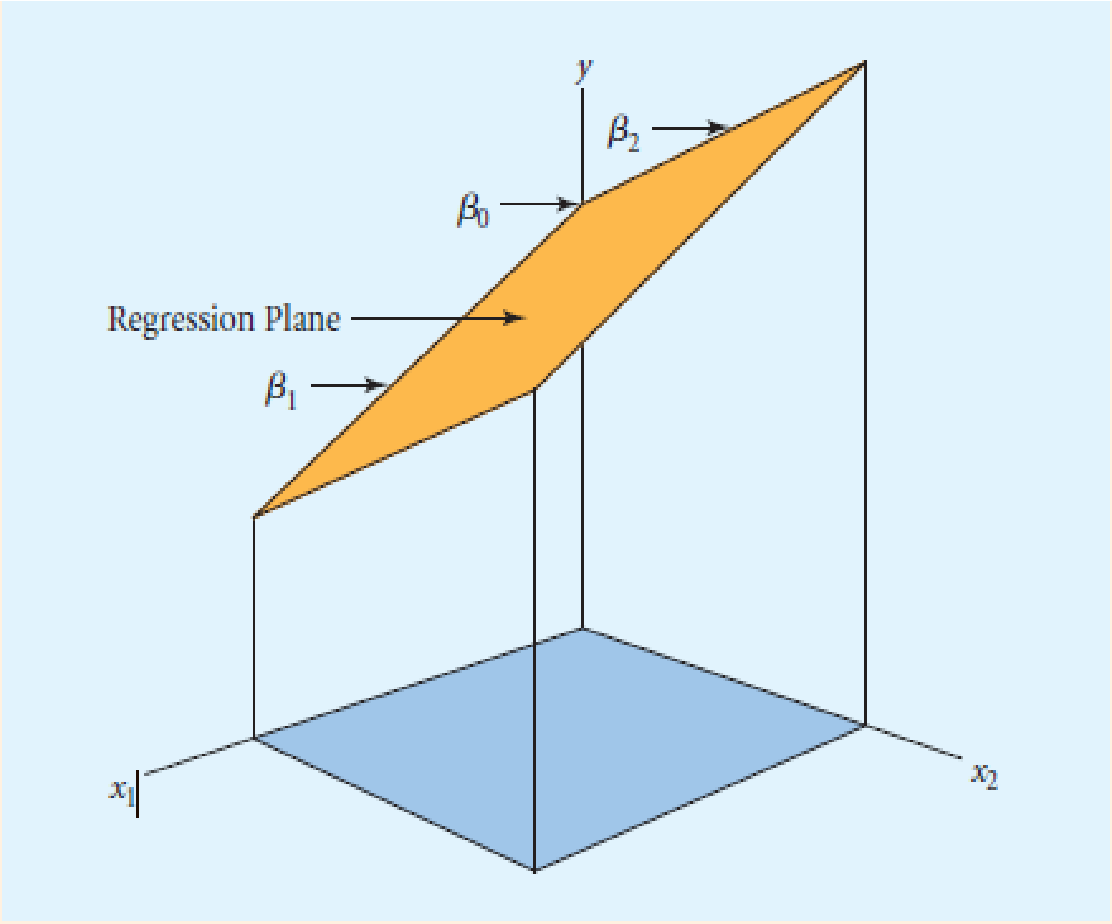
Multiple Regression Model

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

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

1. 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.

2. 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.
- We must decide which variables to include in the model and which to exclude.
  - One tool that will aid us in this decision is the *correlation matrix*.

Correlation Matrix

- What it is?

  shows the linear correlation between each pair of variables under consideration in a multiple regression model.

- How to choose the explanatory (independent) variables for the regression model?

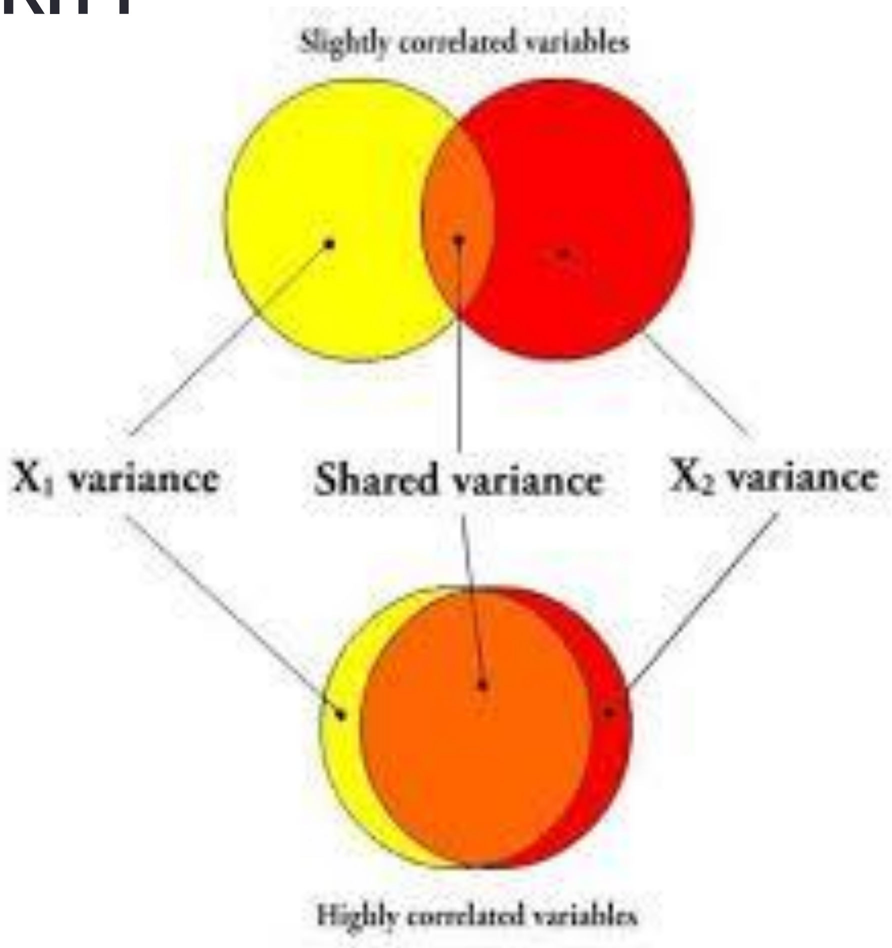
  -- choose the ones that have a high linear correlation with the response variable.

- **Caution!!!** avoid explanatory variables that are highly correlated among themselves. --- Multicollinearity
- Multicollinearity: it exists between two explanatory variables if they have a high linear correlation.

- If exists explanatory variables that are highly correlated among themselves watch out for strange results in the regression output.

- Example of strange results:
  - Getting estimates of slope coefficients that are the opposite sign of what we would expect or
  - Obtaining estimates of slope coefficients that are not as large (or small) as we would expect.

MULTICOLLINEARITY



A general rule: linear correlation between two explanatory variables less than -0.7 or greater than 0.7 may be cause for concern.

Example of multicollinearity situation:

Problem 1: Prediction analysis on sales of lemonade for XYZ Café. Variables that might help to explain lemonade sales are, outside temperature and air-conditioning bills.

If the researcher includes both explanatory variables in the model, he may get results that are a little strange, because the two explanatory variables are themselves highly correlated.

As temperatures increase, so do air-conditioning bills. It would be meaningless to include both variables in the model because they are both doing the same job when it comes to explaining lemonade sales.

x_1	x_2	y
11.4	−9.7	14.7
12.5	−11.5	38.8
16.4	−15.9	42.9
14.4	−13.9	45.7
15.3	−14.2	52.3
18	−18.5	55.9
19.5	−21.2	60.1
25.2	−27.2	72.6

Example: Refer to the dataset given and answer the questions:

(a) Find the correlation matrix among all three variables.

(b) Find the least-squares regression model using both x_1 and x_2 as explanatory variables.

(c) Comment on the effect that including both x_1 and x_2 has on the *t*-test statistics.

Solution:

(a) The correlation matrix is as below.

	x1	x2
x2	-0.99607612	
y	0.89091722	-0.89445738

An extremely high correlation exists between *x*1 and *x*2, so multicollinearity exists between the two variables.

(b) Find the least-squares regression model using both x_1 and x_2 as explanatory variables.

The *P*-value for the *F*-test statistic is 0.0179, indicating that at least one of the slope coefficients is different from zero. However, if we look at each individual *t*-test statistic, we see that each has a very high *P*-value indicating that neither coefficient is different from zero.

Parameter estimates:

Parameter	Estimate	Std. Err.	Alternative	DF	T-Stat	P-value
Intercept	3.1966894	35.467992	≠ 0	5	0.090128851	0.9317
x1	-0.015176864	8.8287255	≠ 0	5	-0.0017190323	0.9987
x2	-2.7209724	6.8440747	≠ 0	5	-0.39756615	0.7074

Analysis of variance table for multiple regression model:

Source	DF	SS	MS	F-stat	P-value
Model	2	1645.8514	822.92568	10.003384	0.0179
Error	5	411.32364	82.264729		
Total	7	2057.175			

Summary of fit:
Root MSE: 9.0699906
R-squared: 0.8001
R-squared (adjusted): 0.7201

(c) Comment on the effect that including both x_1 and x_2 has on the *t*-test statistics.

The contradictory results of the regression output occur because both x_1 and x_2 are related to the response variable y, as indicated by the correlation matrix.

However, x_1 and x_2 are also related to each other. So, with x_1 in the model, x_2 adds little explanation. Likewise, with x_2 in the model, x_1 adds little explanation.

The solution is to use only one explanatory variable. Which explanatory variable we choose is up to you. We can choose either the explanatory variable with the lower *P* value or the explanatory variable that has the higher correlation with the response.

3. 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.

- The objective of model diagnosis is to help you make better decisions.
  - Sophisticated model does not necessary will produce an 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 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) → 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**.

	A	B	C	D	E	F
1	Price	Sq. Feet	Age	Bedrooms	Bathrooms	Garage
2	$110,000	1,000	28	3	1	1
3	$133,500	1,400	23	3	1	1
4	$112,500	1,248	58	3	4	1
5	$141,750	1,106	12	2	1	1
6	$195,250	2,112	78	2	6	2
7	$132,250	1,078	33	2	1	1
8	$136,000	952	13	2	3	2
9	$162,750	1,100	1	2	1	2
10	$148,500	1,040	17	3	1	2
11	$123,500	1,416	27	4	2	1
12	$142,250	1,150	25	3	2	2
13	$145,500	1,220	17	3	2	2
14	$155,250	1,464	28	3	2	2
15	$150,750	1,228	15	3	2	2
16	$150,900	1,132	1	3	4	2
17	$144,000	1,132	1	3	4	2
18	$151,900	1,132	1	3	4	2
19	$161,500	1,464	29	3	3	2
20	$155,750	1,270	1	4	3	2
21	$157,250	1,362	23	3	4	2
22	$152,900	1,120	1	3	3	2
23	$145,250	1,025	1	3	5	2

y = Prices sales for residential property
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)

EXAMPLE: Developing Multiple Regression Model

2. Model Building:

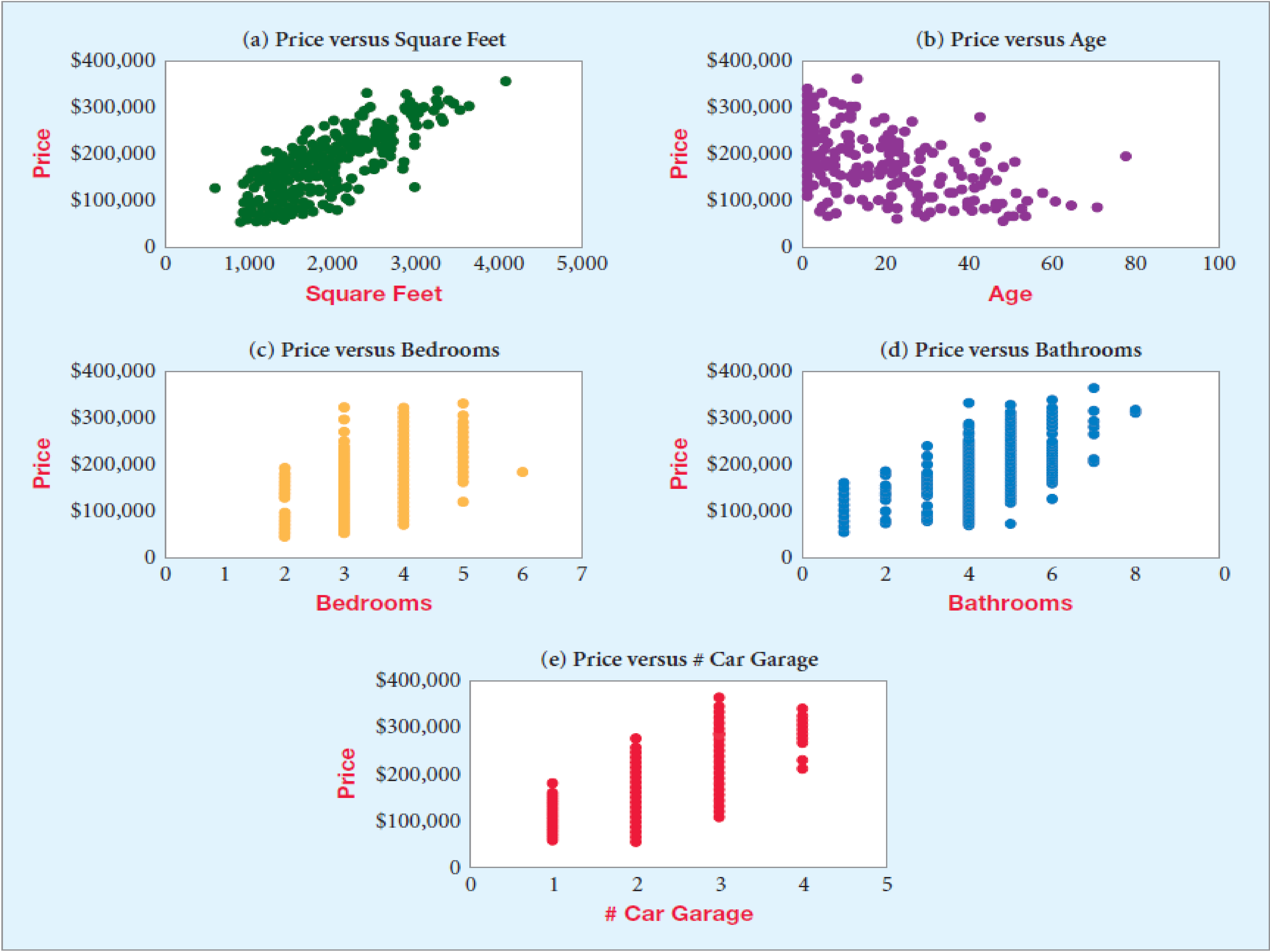
There is **NO WAY** to determine whether an independent variable will be a good predictor variable by analyzing the individual variable's descriptive statistics.

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  1.0000000  0.5996403  0.3120343
Bathrooms  0.6655043  0.62928955 -0.38710488  0.5996403  1.0000000  0.4646015
Garage     0.6935385  0.41626129 -0.43737948  0.3120343  0.4646015  1.0000000
>
```

Correlation Matrix:

Do you find any multicollinearity among the explanatory variable?



(a) Price versus Square Feet
$400,000
$300,000
$200,000
$100,000
0
Price
0
1,000
2,000
3,000
4,000
5,000
Square Feet
(b) Price versus Age
$400,000
$300,000
$200,000
$100,000
0
Price
0
20
40
60
80
100
Age
(c) Price versus Bedrooms
$400,000
$300,000
$200,000
$100,000
0
Price
0
1
2
3
4
5
6
7
Bedrooms
(d) Price versus Bathrooms
$400,000
$300,000
$200,000
$100,000
0
Price
0
2
4
6
8
0
Bathrooms
(e) Price versus # Car Garage
$400,000
$300,000
$200,000
$100,000
0
Price
0
1
2
3
4
5
Car Garage

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
>
```

Regression Coefficient

Sum of squares error (SSE)

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
>
```

R^2 value = 0.8161 or 82%
Adjusted R^2 value = 0.8131 or 81%

Do you know the difference between R^2 value with adjusted R^2 value? When can we use them?

What is the difference between R^2 and adjusted R^2?

Coefficient of determination, $\boldsymbol{R^2}$, measures the percentage of total variation in the response variable that is explained by the least squares regression line.

Adjusted $\boldsymbol{R^2}$ is the adjusted coefficient of determination based on the sample size, *n*, and the number of explanatory variables, *k*.

When to use R^2 and adjusted R^2?

It is recommended that the adjusted R^2 be used when working with least squares regression models with two or more explanatory variables.

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

R^2 value = 0.8161 or 82%

Adjusted R^2 value = 0.8131 or 81%

Based on the Adjusted R^2 value, about 81% 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.

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 the correlation coefficient and identify multicollinearity
ii. Analyzing R-squared (R^2)- multiple coefficient of determination
iii. Analyzing adjusted R-Squared
iv. 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

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 α = 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

a) Is the regression 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^2 value is 81.3%, only slightly less than R^2 value 81.6%.

EXAMPLE: Developing Multiple Regression Model

b) Are the individual variables significant? (Testing the significance of individual predictor variables)

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, α. 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 $\pm \mathbf{2}\boldsymbol{s_e}$

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) = ± \$54,700.
Thus, the rough prediction range of the standard deviation for the model error for the price of an individual home is ± \$54,700.

USING QUALITATIVE INDEPENDENT VARIABLES

USING QUALITATIVE INDEPENDENT VARIABLE

- There is situation 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) is describe in the table. The objective of the problem is to determine whether a model can be generated to explain the variation in annual salary for business executives given the explanatory variable of age and the qualitative variable (x_2) of had a master of business administration (MBA) degree. The dummy variable is hold with this indication:

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),

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

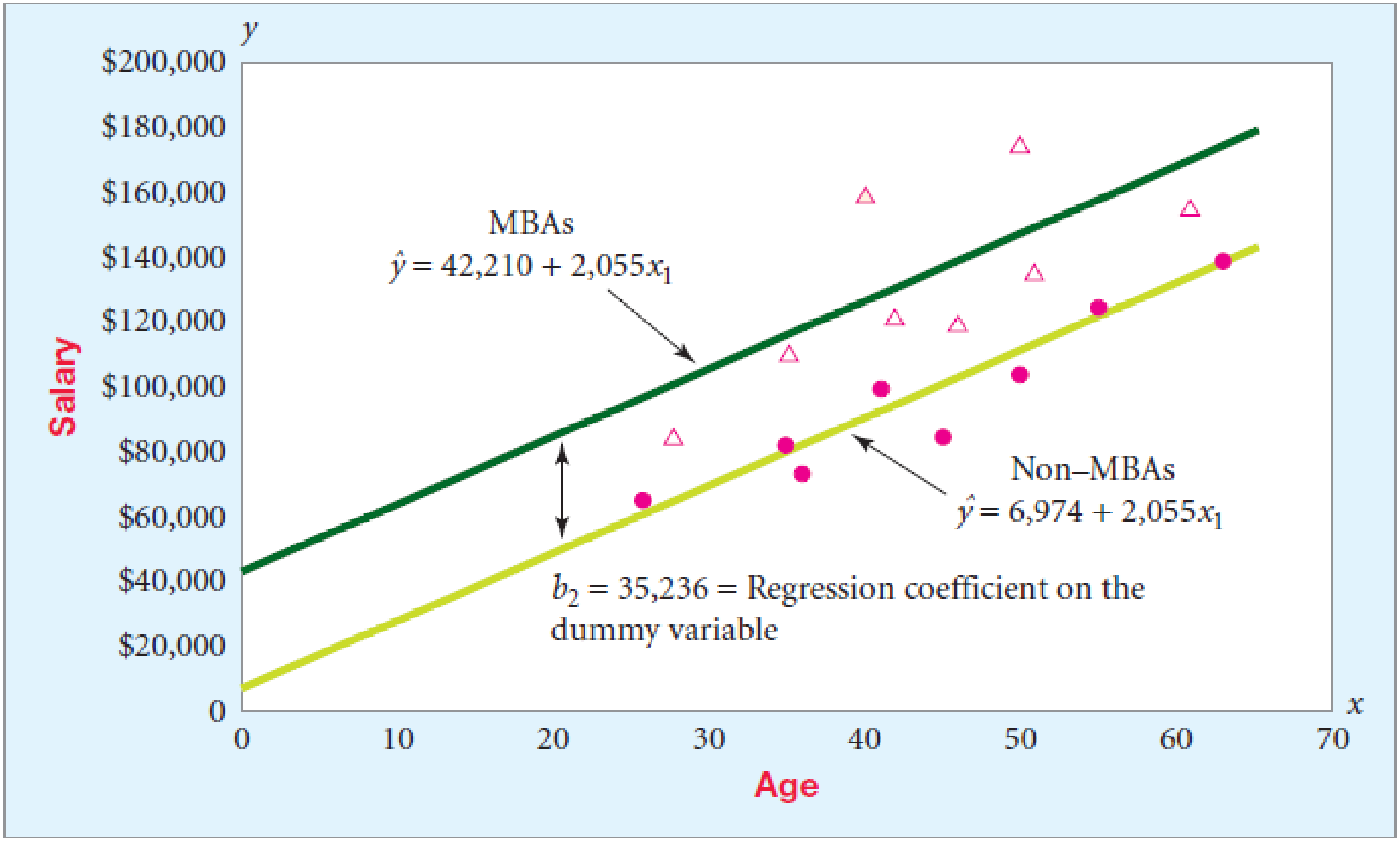
$$\hat{y} = 42{,}210 + 2{,}055x_1$$

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 showed potential because the overall model was statistically significant (slide #21). The model explained nearly 82% of the variation in sales prices for the home in the sample (slide #24). All the independent variables were significant, given that the other independent variables were in the model (slide #32). However, the standard error of the estimate is quite high at $27,350 (slide #34).

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(area) = 1 if foothills, 0 if not

Price	Sq. Feet	Age	Bedrooms	Bathrooms	Garage	Area
$110,000	1000	28	3	1	1	1
$133,500	1400	23	3	1	1	1
$112,500	1248	58	3	4	1	1
$141,750	1106	12	2	1	1	1
$195,250	2112	78	2	6	2	1
$132,250	1078	33	2	1	1	1
$136,000	952	13	2	3	2	1
$162,750	1100	1	2	1	2	1
$148,500	1040	17	3	1	2	1
$123,500	1416	27	4	2	1	1
$142,250	1150	25	3	2	2	1
$145,500	1220	17	3	2	2	1
$155,250	1464	28	3	2	2	1
$150,750	1228	15	3	2	2	1
$150,900	1132	1	3	4	2	1
$144,000	1132	1	3	4	2	1
$151,900	1132	1	3	4	2	1
$161,500	1464	29	3	3	2	1
$155,750	1270	1	4	3	2	1
$157,250	1362	23	3	4	2	1
$152,900	1120	1	3	3	2	1
$145,250	1025	1	3	5	2	1

y = Prices sales for residential property

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)

x_6 = House area (either foothills or not)

```
> library(readxl)
> FirstCityNew <- read_excel("D:/AIS_STUFF/SUBJEK PENGAJARAN/MANB1123_Business Stat for DS/excel_data/FirstCityNew.xlsx",
+     sheet = "Sheet2")
> View(FirstCityNew)
> FirstModel = lm(Price~Sq.Feet+Age+Bedrooms+Bathrooms+Garage+Area,data = FirstCityNew)
> summary(FirstModel)

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

Residuals:
   Min     1Q Median     3Q    Max
-97212 -10810   2133  12010  53857

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -6817.339  7273.961  -0.937 0.349368
Sq.Feet        63.333     2.912  21.747  < 2e-16 ***
Age          -333.836    94.883  -3.518 0.000499 ***
Bedrooms    -8444.831  2176.762  -3.880 0.000128 ***
Bathrooms    -949.195  1176.549  -0.807 0.420418
Garage      26246.435  2075.752  12.644  < 2e-16 ***
Area        62040.983  3684.608  16.838  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 19830 on 312 degrees of freedom
Multiple R-squared:  0.9036,    Adjusted R-squared:  0.9018
F-statistic: 487.7 on 6 and 312 DF,  p-value: < 2.2e-16

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

Response: Price
           Df     Sum Sq    Mean Sq  F value    Pr(>F)
Sq.Feet     1 7.1172e+11 7.1172e+11 1810.228 < 2.2e-16
Age         1 2.3744e+11 2.3744e+11  603.921 < 2.2e-16
Bedrooms    1 7.4847e+09 7.4847e+09   19.037 1.745e-05
Bathrooms   1 9.4429e+09 9.4429e+09   24.018 1.537e-06
Garage      1 7.2811e+10 7.2811e+10  185.191 < 2.2e-16
Area        1 1.1147e+11 1.1147e+11  283.514 < 2.2e-16
Residuals 312 1.2267e+11 3.9317e+08
>
```

Regression Coefficient

Standard error

Improved R^2 and adjusted R^2

Significance test on model

```
> library(readxl)
> FirstCityNew <- read_excel("D:/AIS_STUFF/SUBJEK PENGAJARAN/MANB1123_Business Stat for DS/excel_data/FirstCityNew.xlsx",
+     sheet = "Sheet2")
> View(FirstCityNew)
> FirstModel = lm(Price~Sq.Feet+Age+Bedrooms+Bathrooms+Garage+Area,data = FirstCityNew)
> summary(FirstModel)

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

Residuals:
   Min     1Q Median     3Q    Max 
-97212 -10810   2133  12010  53857 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -6817.339   7273.961  -0.937 0.349368    
Sq.Feet        63.333      2.912  21.747  < 2e-16 ***
Age          -333.836     94.883  -3.518 0.000499 ***
Bedrooms    -8444.831   2176.762  -3.880 0.000128 ***
Bathrooms    -949.195   1176.549  -0.807 0.420418    
Garage      26246.435   2075.752  12.644  < 2e-16 ***
Area        62040.983   3684.608  16.838  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 19830 on 312 degrees of freedom
Multiple R-squared:  0.9036,    Adjusted R-squared:  0.9018 
F-statistic: 487.7 on 6 and 312 DF,  p-value: < 2.2e-16
```

** signal of multicollinearity problem (sign of negative slope)

** need to be excluded from the model

** more work to be done before the model is completed

We could start by identifying possible problems:
1. We maybe missing useful independent variables.
2. Independent variables may have been included that should not have been included.

There is no sure way of determining the correct model specification. However, a recommended approach is for the decision maker to try adding variables or removing variables from the model.

```
> FirstModel = lm(Price~Sq.Feet+Age+Garage+Area,data = FirstCityNew)
> summary(FirstModel)

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

Residuals:
    Min      1Q  Median      3Q     Max 
-101248   -9585    1376   11633   57750 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -25617.326   5878.261  -4.358 1.78e-05
Sq.Feet         54.832      2.051  26.737  < 2e-16
Age           -261.297     94.917  -2.753  0.00625
Garage       26753.303   2106.618  12.700  < 2e-16
Area         60578.045   3674.322  16.487  < 2e-16

Residual standard error: 20330 on 314 degrees of freedom
Multiple R-squared:  0.8981,    Adjusted R-squared:  0.8968 
F-statistic: 691.9 on 4 and 314 DF,  p-value: < 2.2e-16
```

** what can you conclude with the revised model?

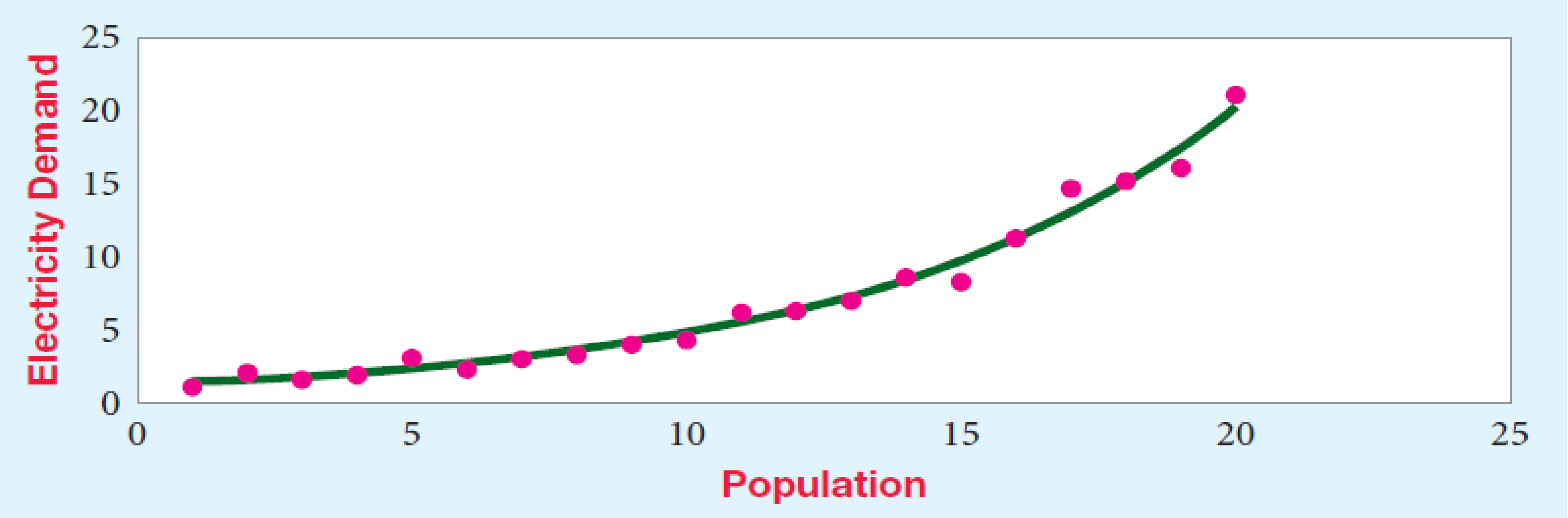
WORKING WITH NONLINEAR RELATIONSHIP

NONLINEAR RELATIONSHIP

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

Example:

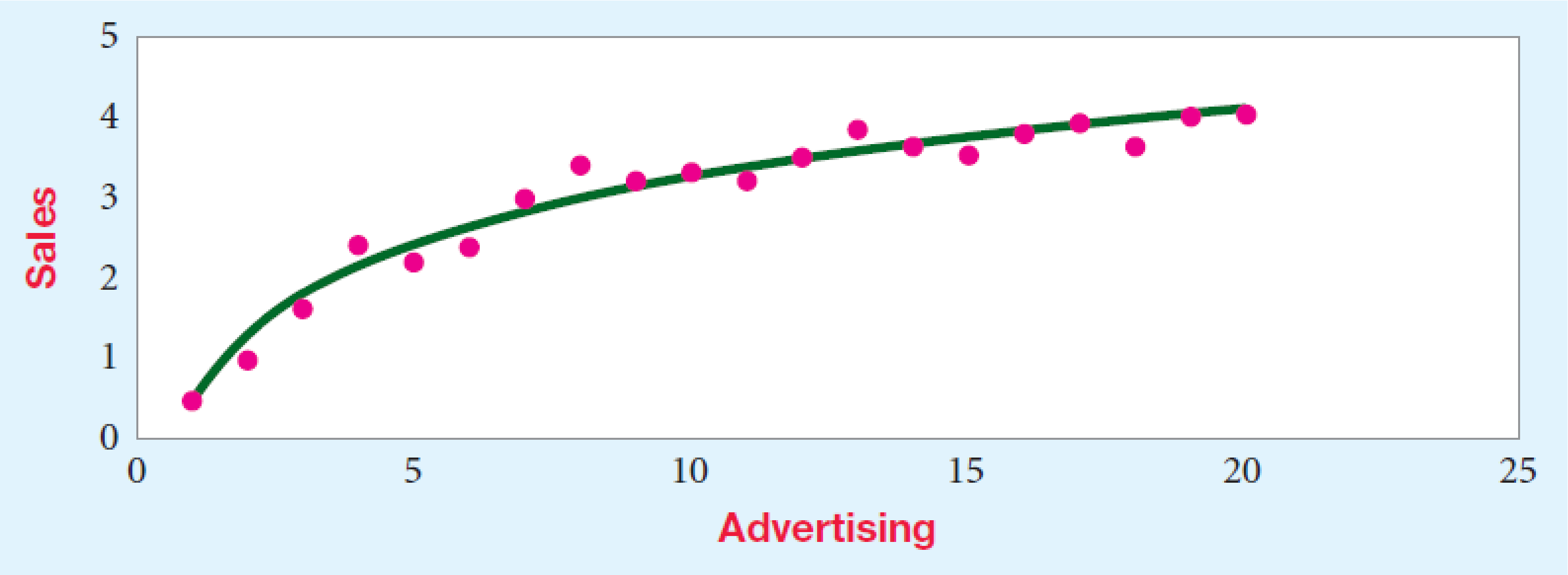
- Demand for electricity has grown at an almost exponential rate relative to the population growth in some areas.



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

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_p x^p + \varepsilon$$

where:

β_0 = Population regression's constant

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

p = Order (or degree) of the polynomial

ε = Model error

NONLINEAR RELATIONSHIP

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_p x^p + \varepsilon$$

where:

β_0 = Population regression's constant
β_j = Population's regression coefficient for variable x^j; $j = 1, 2, \ldots, p$
p = Order (or degree) of the polynomial
ε = Model error

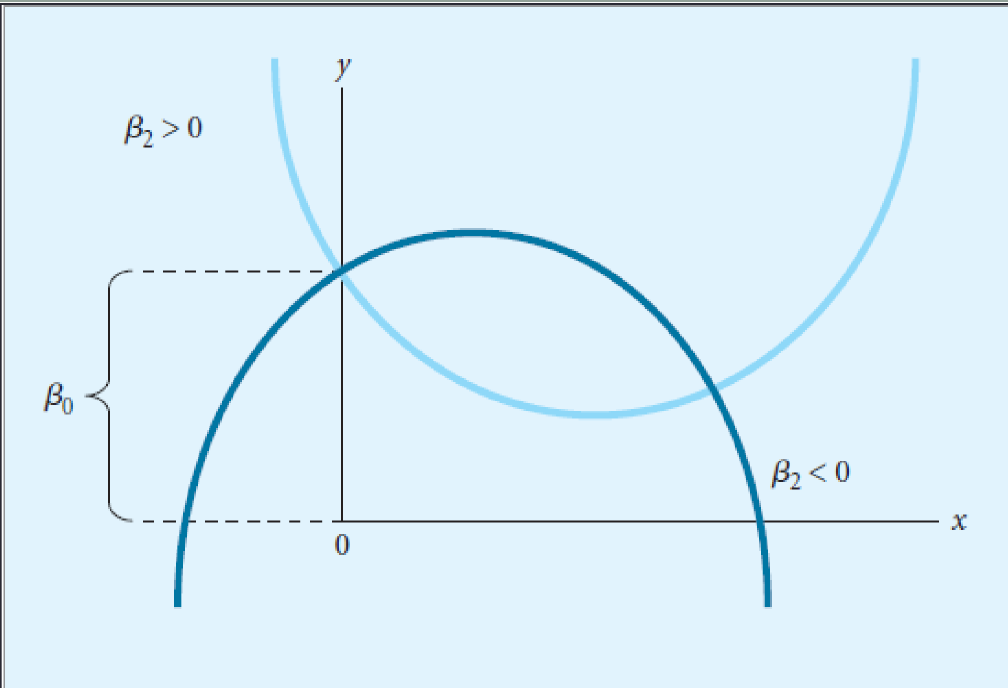
The order/degree of the model is determined by the largest exponent of the independent variable in the model.

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon$$

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

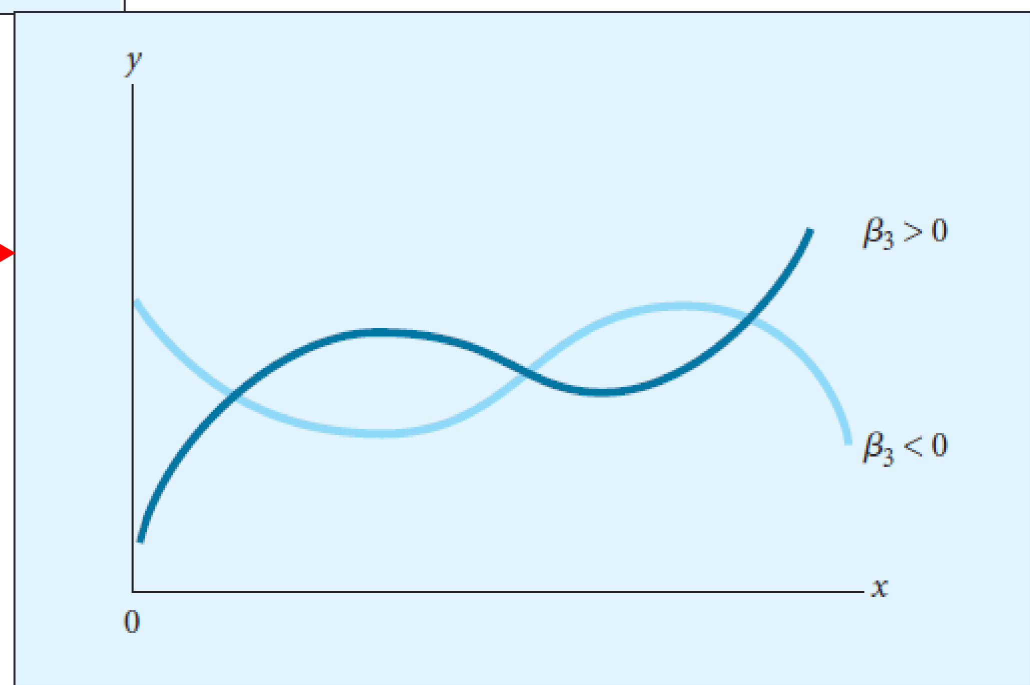
As more curves appear in the data, the order of the polynomial must be increased. Example of third-order polynomial

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 \varepsilon$$



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**.

	A	B
1	SocializationMeasure	BurnoutIndex
2	20	100
3	60	525
4	38	300
5	88	980
6	59	310
7	87	900
8	68	410
9	12	296
10	35	120
11	70	501
12	80	920
13	92	810
14	77	506
15	86	493
16	83	892
17	79	527
18	75	600
19	81	855
20	75	709
21	77	791
22		

y = Burn out Index

x_1 = Socialization Measure

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.**

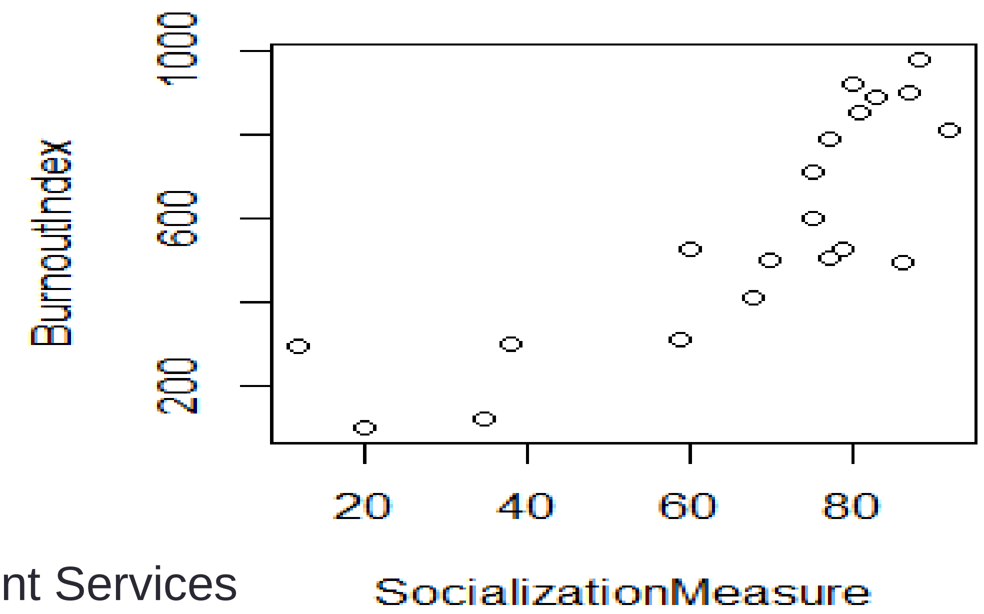
The dependent variable is the burnout index. The company wishes to explain the variation in burnout level. One potential independent variable is the socialization index.

Step 2: **Formulate the model.**

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

Console ~/

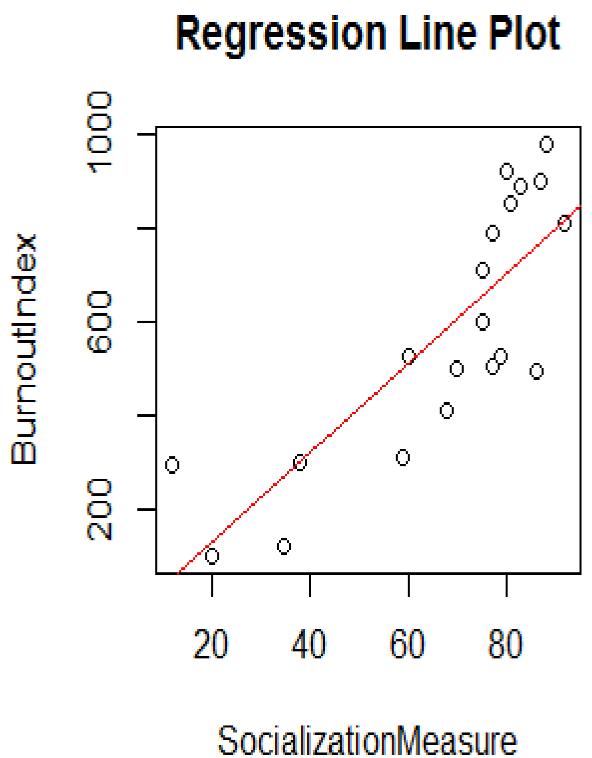
```
> library(readxl)
> Ashley <- read_excel("D:/AIS_STUFF/SUBJEK PENGAJARAN/MANB1123_Business Stat for DS/excel_data/Ashley.xlsx")
> View(Ashley)
> plot(BurnoutIndex~SocializationMeasure,data = Ashley)
> with(Ashley,cor.test(BurnoutIndex,SocializationMeasure))

        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                  
> 
```



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

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:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon$$

From the computer output (slide #62), the estimated regression equation is:

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

	A	B	C
1	SocializationMeasure	BurnoutIndex	SocializationMeasureSq
2	20	100	400
3	60	525	3600
4	38	300	1444
5	88	980	7744
6	59	310	3481
7	87	900	7569
8	68	410	4624
9	12	296	144
10	35	120	1225
11	70	501	4900
12	80	920	6400
13	92	810	8464
14	77	506	5929
15	86	493	7396
16	83	892	6889
17	79	527	6241
18	75	600	5625
19	81	855	6561
20	75	709	5625
21	77	791	5929
22			
23			

y = Burn out Index
x_1 = Socialization Measure
x_2 = Socialization Measure square
(to add the quadratic/polynomial)

Console ~/

```
> 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
```

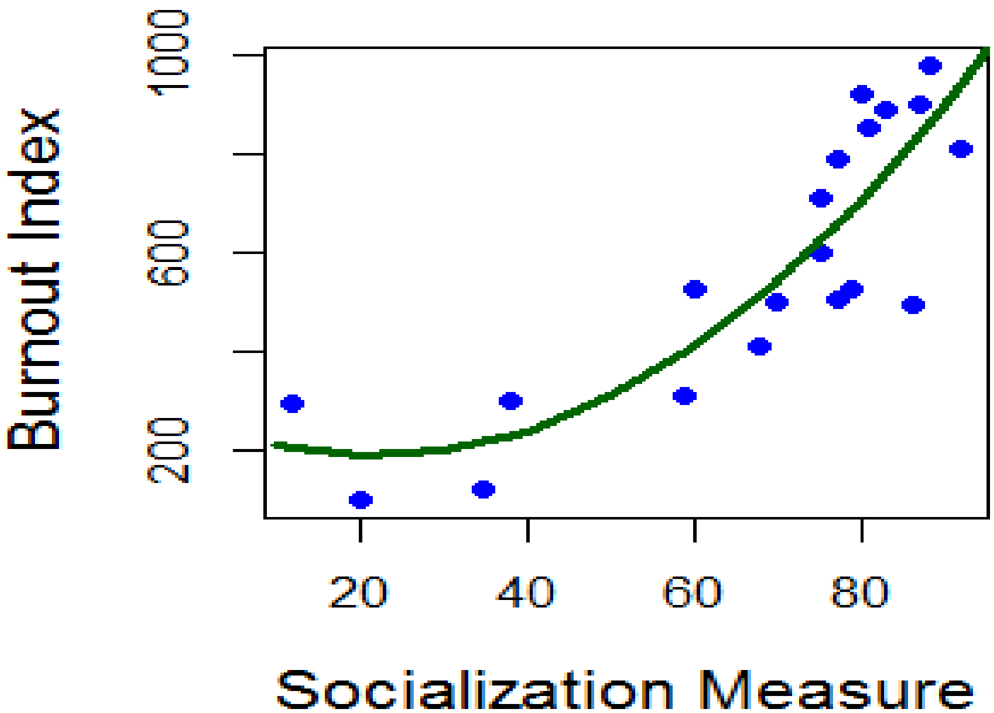
NONLINEAR RELATIONSHIP

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% (or Adjusted R^2 = 71%) , which is more higher (see output on slide #62).

The second-order polynomial plot is depicted on the following slide (slide #64)

```
> 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 ☺