Chapter 4

Regression Models

To accompany
***Quantitative Analysis for Management*, *Eleventh Edition*,**
by Render, Stair, and Hanna
Power Point slides created by Brian Peterson

Learning Objectives

After completing this chapter, students will be able to:

1. **Identify variables and use them in a regression model.**
2. **Develop simple linear regression equations. from sample data and interpret the slope and intercept.**
3. **Compute the coefficient of determination and the coefficient of correlation and interpret their meanings.**
4. **Interpret the F-test in a linear regression model.**
5. **List the assumptions used in regression and use residual plots to identify problems.**

Learning Objectives

After completing this chapter, students will be able to:

6. Develop a multiple regression model and use it for prediction purposes.
7. Use dummy variables to model categorical data.
8. Determine which variables should be included in a multiple regression model.
9. Transform a nonlinear function into a linear one for use in regression.
10. Understand and avoid common mistakes made in the use of regression analysis.

Chapter Outline

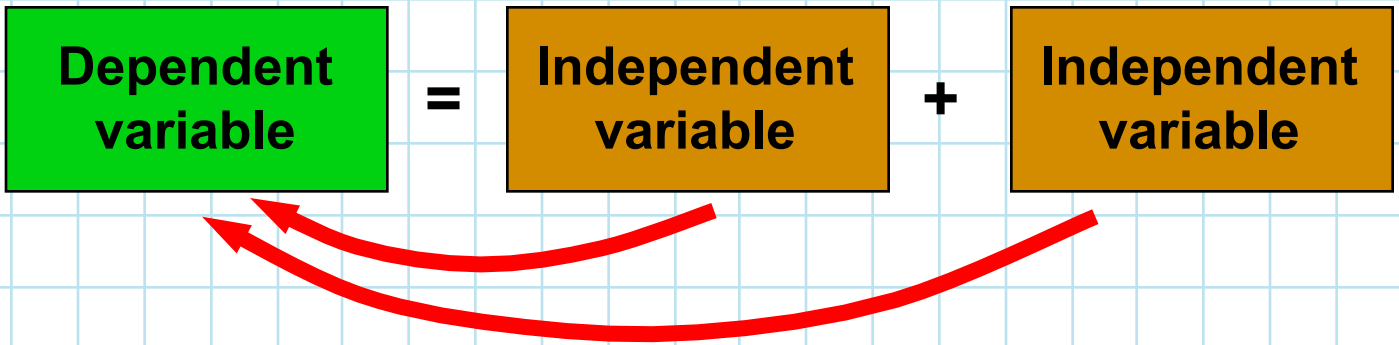
Chapter Outline

Introduction

- *Regression analysis* is a very valuable tool for a manager.
- Regression can be used to:
  - Understand the relationship between variables.
  - Predict the value of one variable based on another variable.
- Simple linear regression models have only two variables.
- Multiple regression models have more variables.

Introduction

- The variable to be predicted is called the *dependent variable*.
  - This is sometimes called the *response variable.*
- The value of this variable depends on the value of the *independent variable.*
  - This is sometimes called the *explanatory* or *predictor variable.*



Scatter Diagram

- A *scatter diagram* or *scatter plot* is often used to investigate the relationship between variables.
- The independent variable is normally plotted on the X axis.
- The dependent variable is normally plotted on the Y axis.

Triple A Construction

- Triple A Construction renovates old homes.
- Managers have found that the dollar volume of renovation work is dependent on the area payroll.

TRIPLE A'S SALES ($100,000s)	LOCAL PAYROLL ($100,000,000s)
6	3
8	4
9	6
5	4
4.5	2
9.5	5

Table 4.1

Triple A Construction

Scatter Diagram of Triple A Construction Company Data

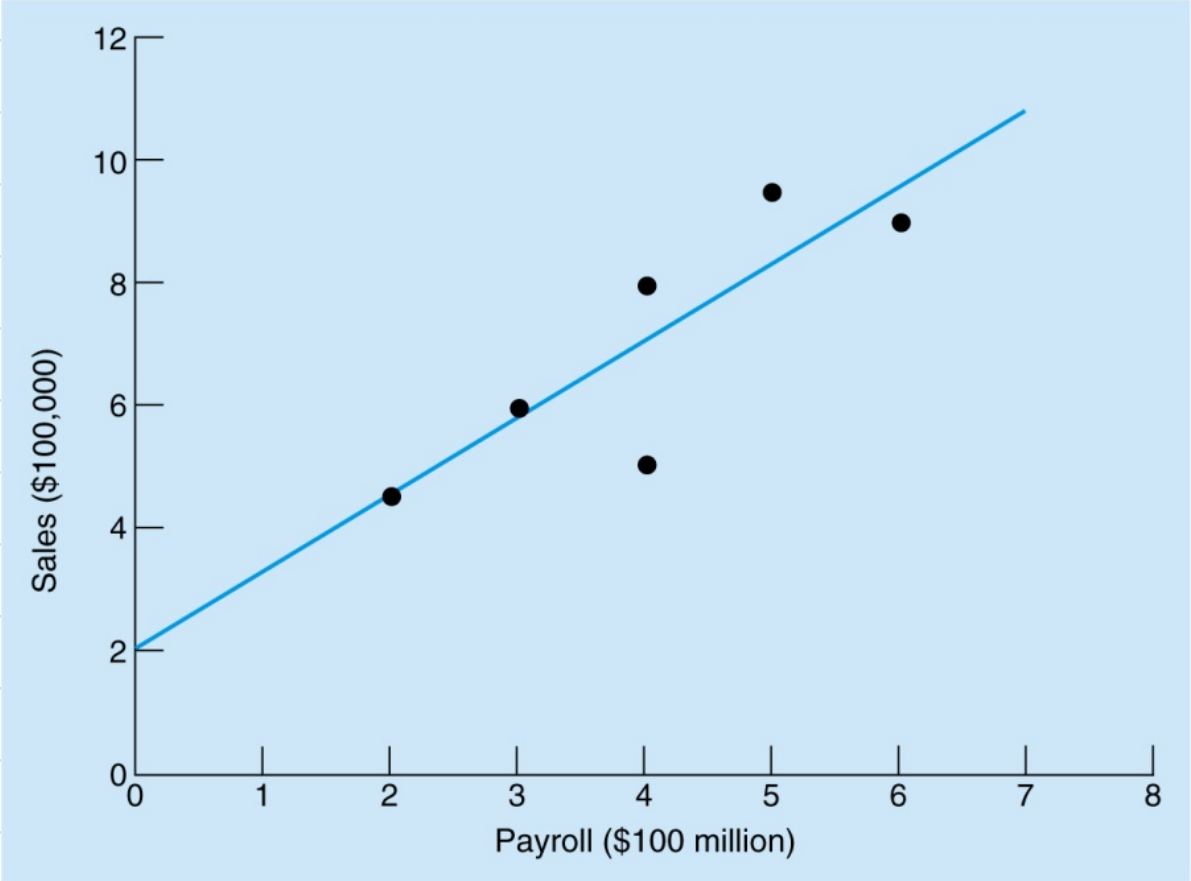


Figure 4.1

Simple Linear Regression

- **Regression models are used to test if there is a relationship between variables.**
- **There is some random error that cannot be predicted.**

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

where

Y = dependent variable (response)

X = independent variable (predictor or explanatory)

β_0 = intercept (value of Y when $X = 0$)

β_1 = slope of the regression line

ε = random error

Simple Linear Regression

- True values for the slope and intercept are not known so they are estimated using sample data.

$$\hat{Y} = b_0 + b_1 X$$

where

$\hat{Y}$ = predicted value of Y

b_0 = estimate of β_0, based on sample results

b_1 = estimate of β_1, based on sample results

Triple A Construction

Triple A Construction is trying to predict sales based on area payroll.

Y = Sales

X = Area payroll

The line chosen in Figure 4.1 is the one that minimizes the errors.

Error = (Actual value) – (Predicted value)

$$e = Y - \hat{Y}$$

Triple A Construction

For the simple linear regression model, the values of the intercept and slope can be calculated using the formulas below.

$$\hat{Y} = b_0 + b_1 X$$

$$\overline{X} = \frac{\sum X}{n} = \text{average (mean) of } X \text{ values}$$

$$\overline{Y} = \frac{\sum Y}{n} = \text{average (mean) of } Y \text{ values}$$

$$b_1 = \frac{\sum (X - \overline{X})(Y - \overline{Y})}{\sum (X - \overline{X})^2}$$

$$b_0 = \overline{Y} - b_1 \overline{X}$$

Triple A Construction

Regression calculations for Triple A Construction

Y	X	$(X - \bar{X})^2$	$(X - \bar{X})(Y - \bar{Y})$
6	3	$(3 - 4)^2 = 1$	$(3 - 4)(6 - 7) = 1$
8	4	$(4 - 4)^2 = 0$	$(4 - 4)(8 - 7) = 0$
9	6	$(6 - 4)^2 = 4$	$(6 - 4)(9 - 7) = 4$
5	4	$(4 - 4)^2 = 0$	$(4 - 4)(5 - 7) = 0$
4.5	2	$(2 - 4)^2 = 4$	$(2 - 4)(4.5 - 7) = 5$
9.5	5	$(5 - 4)^2 = 1$	$(5 - 4)(9.5 - 7) = 2.5$
$\Sigma Y = 42$	$\Sigma X = 24$	$\Sigma(X - \bar{X})^2 = 10$	$\Sigma(X - \bar{X})(Y - \bar{Y}) = 12.5$
$\bar{Y} = 42/6 = 7$	$\bar{X} = 24/6 = 4$		

Table 4.2

Triple A Construction

Regression calculations

$$\bar{X} = \frac{\sum X}{6} = \frac{24}{6} = 4$$

$$\bar{Y} = \frac{\sum Y}{6} = \frac{42}{6} = 7$$

$$b_1 = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{\sum (X - \bar{X})^2} = \frac{12.5}{10} = 1.25$$

$$b_0 = \bar{Y} - b_1 \bar{X} = 7 - (1.25)(4) = 2$$

Therefore $\hat{Y} = 2 + 1.25X$

Triple A Construction

Regression calculations

$$\overline{X} = \frac{\sum X}{6} = \frac{24}{6}$$

$$\overline{Y} = \frac{\sum Y}{6} = \frac{42}{6}$$

$$b_1 = \frac{\sum (X - \overline{X})}{\sum (X -$$

$$b_0 = \overline{Y} - b_1 \overline{X} = 7 - (1.25)(4) = 2$$

Therefore $\hat{Y} = 2 + 1.25X$

sales = 2 + 1.25(payroll)

If the payroll next year is $600 million

$$\hat{Y} = 2 + 1.25(6) = 9.5 \text{ or } \$\,950{,}000$$

Measuring the Fit of the Regression Model

- Regression models can be developed for any variables X and Y.
- How do we know the model is actually helpful in predicting Y based on X?
  - We could just take the average error, but the positive and negative errors would cancel each other out.
- Three measures of variability are:
  - *SST* – Total variability about the mean.
  - *SSE* – Variability about the regression line.
  - *SSR* – Total variability that is explained by the model.

Measuring the Fit of the Regression Model

- **Sum of the squares total**:

$$SST = \sum (Y - \bar{Y})^2$$

- **Sum of the squared error**:

$$SSE = \sum e^2 = \sum (Y - \hat{Y})^2$$

- **Sum of squares due to regression**:

$$SSR = \sum (\hat{Y} - \bar{Y})^2$$

- **An important relationship**:

$$SST = SSR + SSE$$

Measuring the Fit of the Regression Model

Sum of Squares for Triple A Construction

Y	X	$(Y - \bar{Y})^2$	$\hat{Y}$	$(Y - \hat{Y})^2$	$(\hat{Y} - \bar{Y})^2$
6	3	$(6 - 7)^2 = 1$	$2 + 1.25(3) = 5.75$	0.0625	1.563
8	4	$(8 - 7)^2 = 1$	$2 + 1.25(4) = 7.00$	1	0
9	6	$(9 - 7)^2 = 4$	$2 + 1.25(6) = 9.50$	0.25	6.25
5	4	$(5 - 7)^2 = 4$	$2 + 1.25(4) = 7.00$	4	0
4.5	2	$(4.5 - 7)^2 = 6.25$	$2 + 1.25(2) = 4.50$	0	6.25
9.5	5	$(9.5 - 7)^2 = 6.25$	$2 + 1.25(5) = 8.25$	1.5625	1.563
		$\sum(Y - \bar{Y})^2 = 22.5$		$\sum(Y - \hat{Y})^2 = 6.875$	$\sum(\hat{Y} - \bar{Y})^2 = 15.625$
$\bar{Y} = 7$		$SST = 22.5$		$SSE = 6.875$	$SSR = 15.625$

Table 4.3

Measuring the Fit of the Regression Model

- Sum of the square

$$SST$$

- Sum of the square

$$SSE = \sum$$

For Triple A Construction

SST = 22.5
SSE = 6.875
SSR = 15.625

- Sum of squares due to regression

$$SSR = \sum (\hat{Y} - \bar{Y})^2$$

- An important relationship

$$SST = SSR + SSE$$

Measuring the Fit of the Regression Model

Deviations from the Regression Line and from the Mean

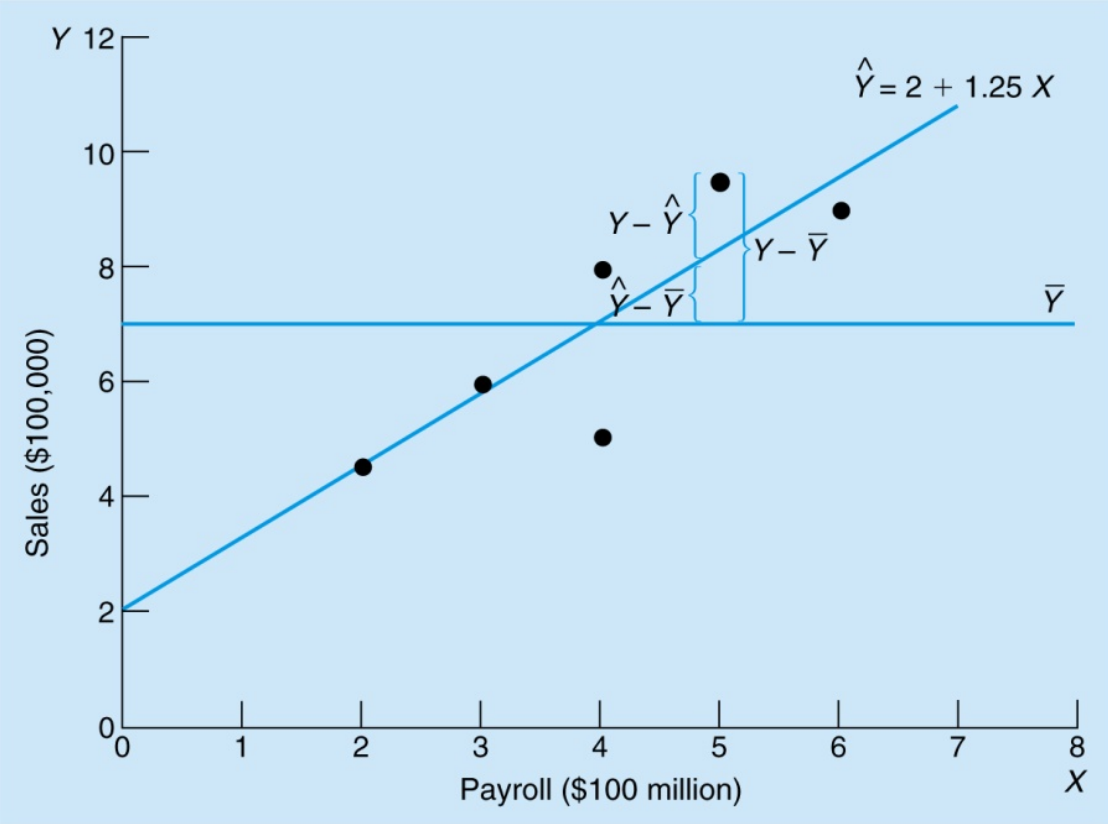


Figure 4.2

Coefficient of Determination

- The proportion of the variability in Y explained by the regression equation is called the *coefficient of determination.*
- The coefficient of determination is r^2.

$$r^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

- For Triple A Construction:

$$r^2 = \frac{15.625}{22.5} = 0.6944$$

- About 69% of the variability in Y is explained by the equation based on payroll (X).

Correlation Coefficient

- The *correlation coefficient* is an expression of the strength of the linear relationship.
- It will always be between +1 and –1.
- The correlation coefficient is r.

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

- For Triple A Construction:

$$r = \sqrt{0.6944} = 0.8333$$

Four Values of the Correlation Coefficient

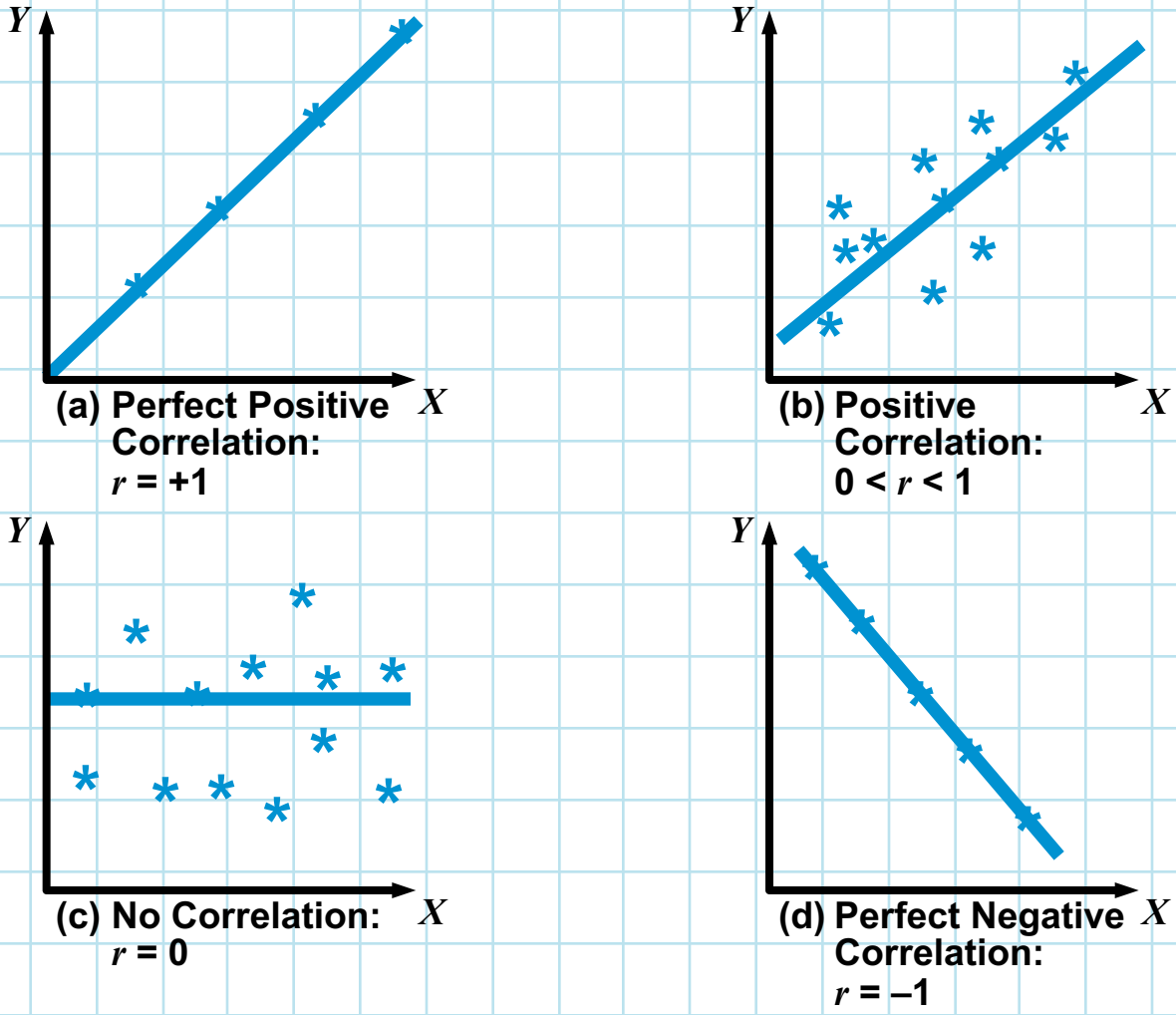
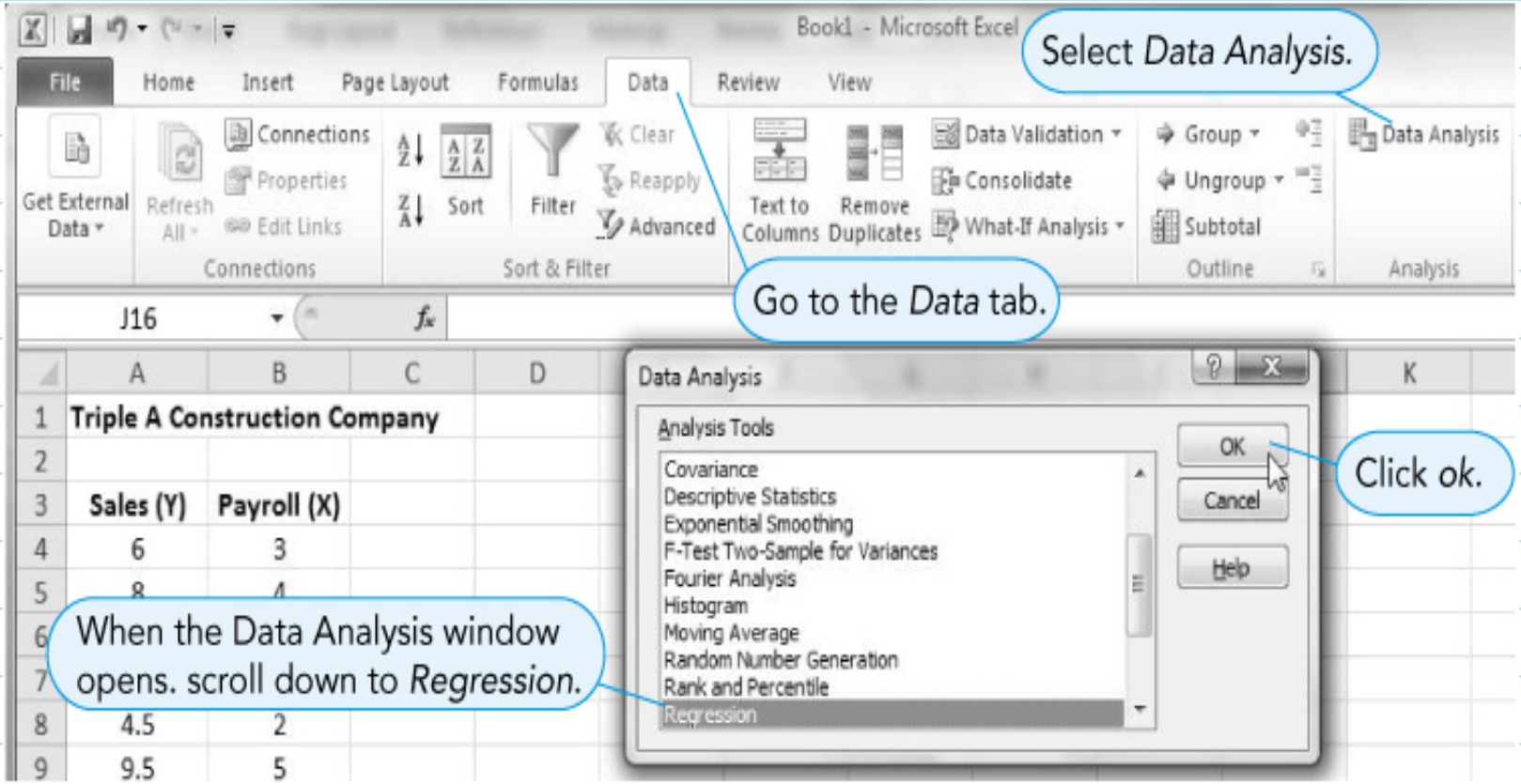


Figure 4.3

Using Computer Software for Regression

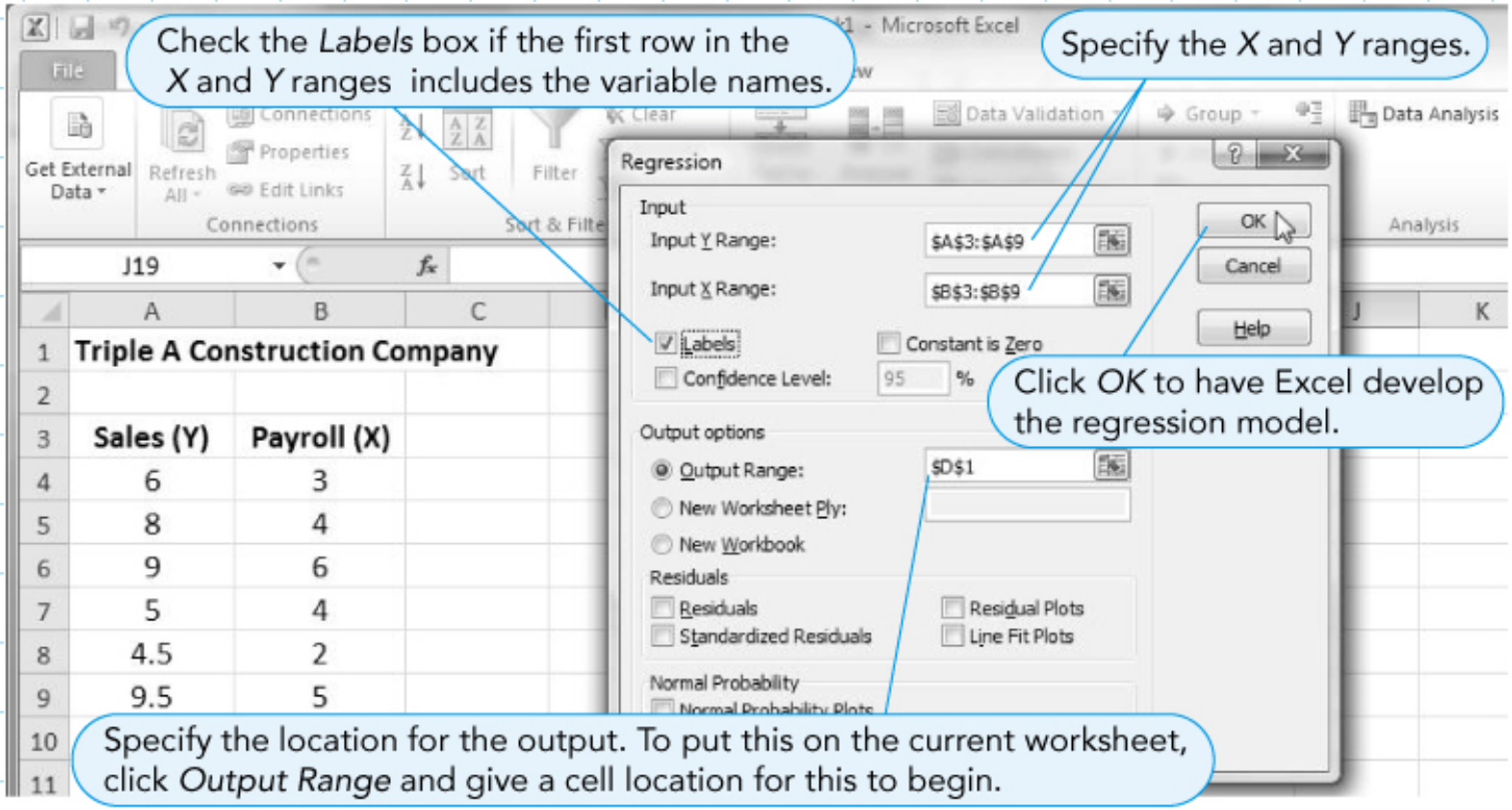
Accessing the Regression Option in Excel 2010



Program 4.1A

Using Computer Software for Regression

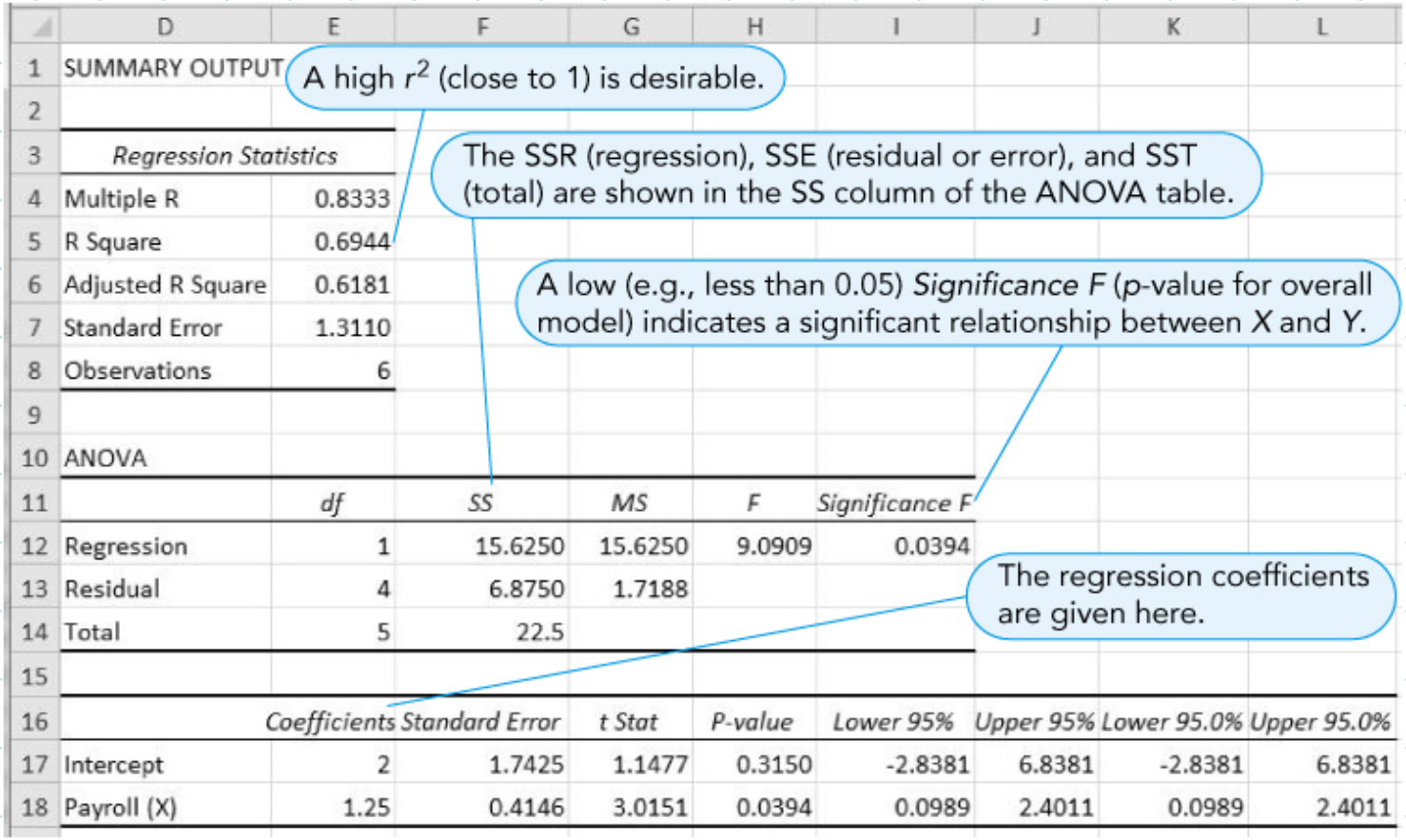
Data Input for Regression in Excel



Program 4.1B

Using Computer Software for Regression

Excel Output for the Triple A Construction Example



	D	E	F	G	H	I	J	K	L
1	SUMMARY OUTPUT								
2									
3	*Regression Statistics*								
4	Multiple R	0.8333							
5	R Square	0.6944							
6	Adjusted R Square	0.6181							
7	Standard Error	1.3110							
8	Observations	6							
9									
10	ANOVA								
11		*df*	*SS*	*MS*	*F*	*Significance F*			
12	Regression	1	15.6250	15.6250	9.0909	0.0394			
13	Residual	4	6.8750	1.7188					
14	Total	5	22.5						
15									
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
17	Intercept	2	1.7425	1.1477	0.3150	-2.8381	6.8381	-2.8381	6.8381
18	Payroll (X)	1.25	0.4146	3.0151	0.0394	0.0989	2.4011	0.0989	2.4011

Program 4.1C

Assumptions of the Regression Model

- **If we make certain assumptions about the errors in a regression model, we can perform statistical tests to determine if the model is useful.**

  1. **Errors are independent.**
  2. **Errors are normally distributed.**
  3. **Errors have a mean of zero.**
  4. **Errors have a constant variance.**

- **A plot of the residuals (errors) will often highlight any glaring violations of the assumption.**

Residual Plots

Pattern of Errors Indicating Randomness

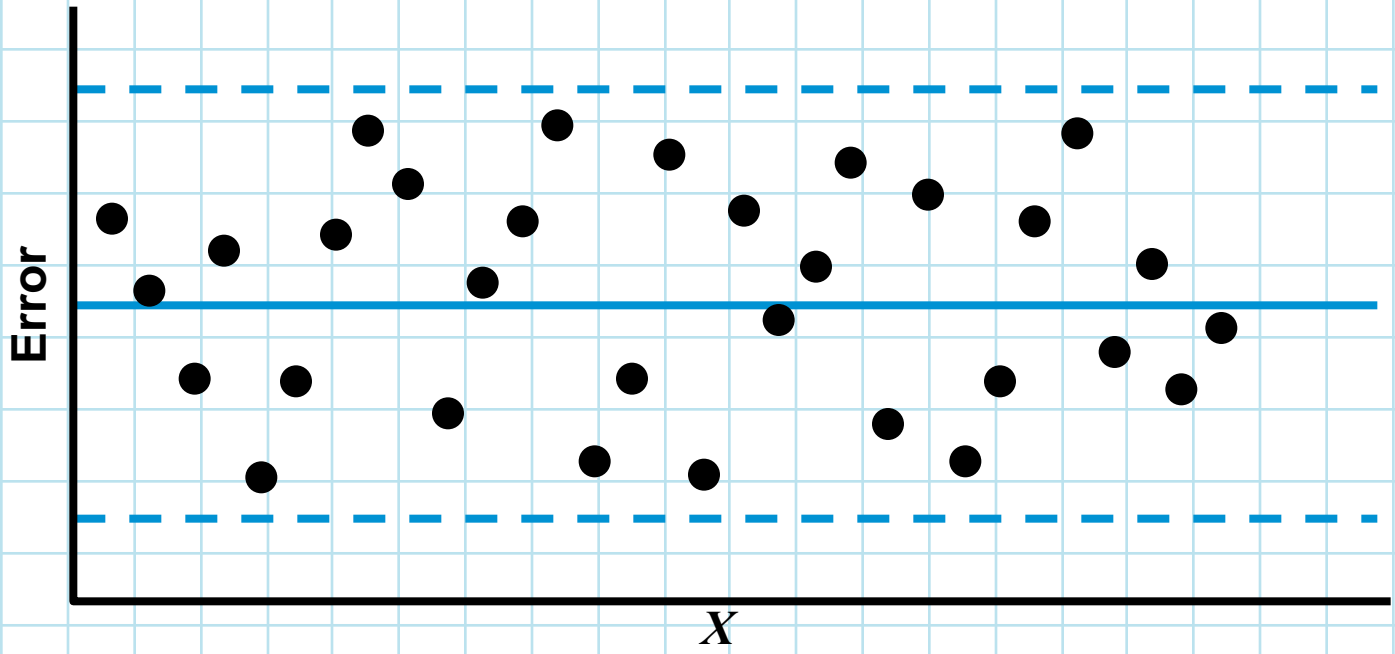


Figure 4.4A

Residual Plots

Nonconstant error variance

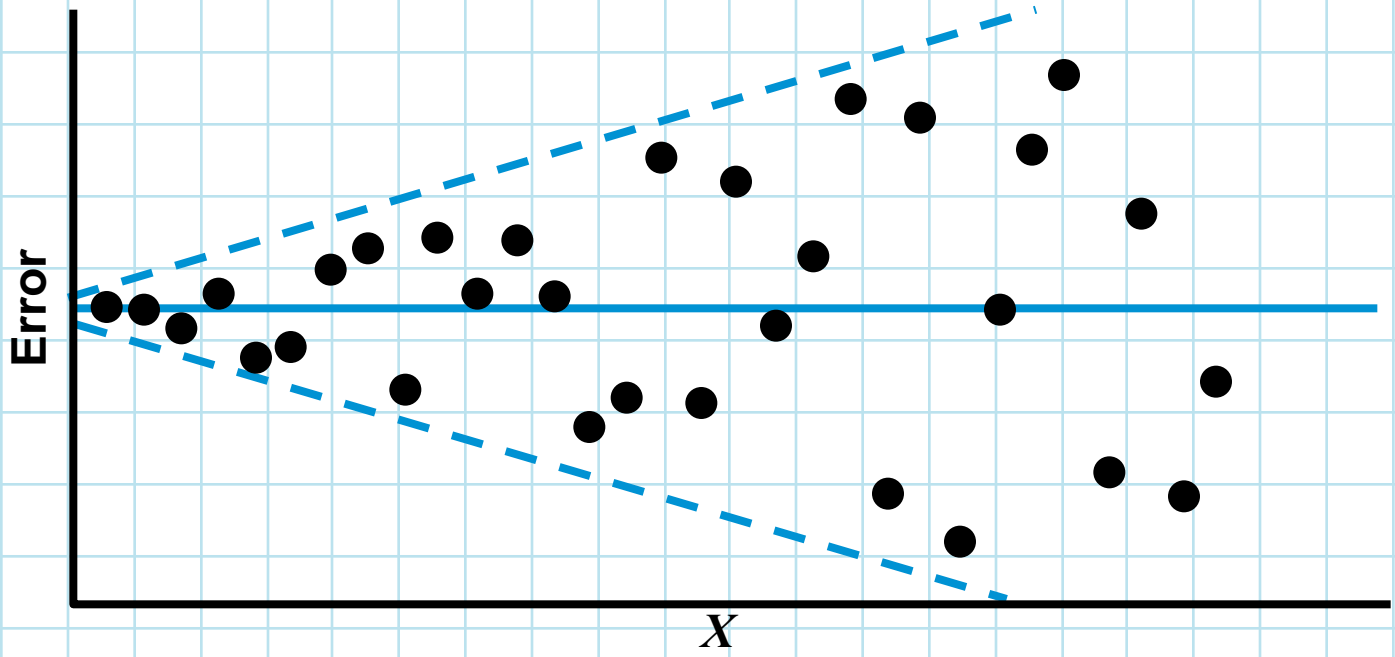


Figure 4.4B

Residual Plots

Errors Indicate Relationship is not Linear

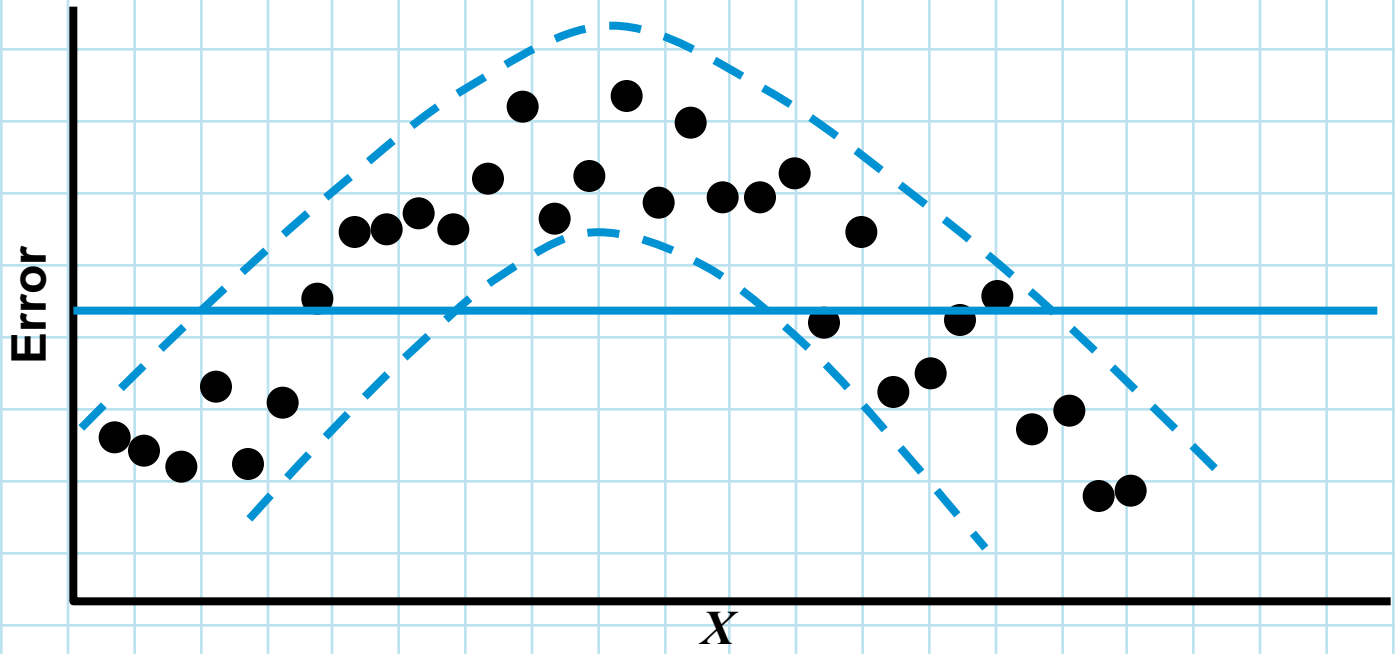


Figure 4.4C

Estimating the Variance

- Errors are assumed to have a constant variance (σ^2), but we usually don't know this.
- It can be estimated using the ***mean squared error*** (***MSE***), s^2.

$$s^2 = MSE = \frac{SSE}{n - k - 1}$$

where

n = number of observations in the sample
k = number of independent variables

Estimating the Variance

- For Triple A Construction:

$$s^2 = MSE = \frac{SSE}{n-k-1} = \frac{6.8750}{6-1-1} = \frac{6.8750}{4} = 1.7188$$

- We can estimate the standard deviation, s.
- This is also called the ***standard error of the estimate*** or the ***standard deviation of the regression.***

$$s = \sqrt{MSE} = \sqrt{1.7188} = 1.31$$

Testing the Model for Significance

- When the sample size is too small, you can get good values for MSE and r^2 even if there is no relationship between the variables.
- Testing the model for significance helps determine if the values are meaningful.
- We do this by performing a statistical hypothesis test.

Testing the Model for Significance

- We start with the general linear model

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

- If $\beta_1 = 0$, the null hypothesis is that there is *no* relationship between X and Y.
- The alternate hypothesis is that there *is* a linear relationship ($\beta_1 \neq 0$).
- If the null hypothesis can be rejected, we have proven there is a relationship.
- We use the F statistic for this test.

Testing the Model for Significance

- The F statistic is based on the MSE and MSR:

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

where

k = number of independent variables in the model

- The F statistic is:

$$F = \frac{MSR}{MSE}$$

- This describes an F distribution with:

degrees of freedom for the numerator = $df_1 = k$

degrees of freedom for the denominator = $df_2 = n - k - 1$

Testing the Model for Significance

- If there is very little error, the MSE would be small and the F-statistic would be large indicating the model is useful.
- If the F-statistic is large, the significance level (p-value) will be low, indicating it is unlikely this would have occurred by chance.
- So when the F-value is large, we can reject the null hypothesis and accept that there is a linear relationship between X and Y and the values of the MSE and r^2 are meaningful.

Steps in a Hypothesis Test

1. Specify null and alternative hypotheses:

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0$$

2. Select the level of significance (α). Common values are 0.01 and 0.05.
3. Calculate the value of the test statistic using the formula:

$$F = \frac{MSR}{MSE}$$

Steps in a Hypothesis Test

4. Make a decision using one of the following methods:

   a) Reject the null hypothesis if the test statistic is greater than the F-value from the table in Appendix D. Otherwise, do not reject the null hypothesis:

$$\text{Reject if } F_{calculated} > F_{\alpha, df_1, df_2}$$
$$df_1 = k$$
$$df_2 = n - k - 1$$

   b) Reject the null hypothesis if the observed significance level, or p-value, is less than the level of significance (α). Otherwise, do not reject the null hypothesis:

$$p\text{-value} = P(F > \text{calculated test statistic})$$
$$\text{Reject if } p\text{-value} < \alpha$$

Triple A Construction

Step 1.

H_0: $\beta_1 = 0$ (no linear relationship between X and Y)

H_1: $\beta_1 \neq 0$ (linear relationship exists between X and Y)

Step 2.

Select $\alpha = 0.05$

Step 3.

Calculate the value of the test statistic.

$$MSR = \frac{SSR}{k} = \frac{15.6250}{1} = 15.6250$$

$$F = \frac{MSR}{MSE} = \frac{15.6250}{1.7188} = 9.09$$

Triple A Construction

Step 4.

Reject the null hypothesis if the test statistic is greater than the F-value in Appendix D.

$$df_1 = k = 1$$
$$df_2 = n - k - 1 = 6 - 1 - 1 = 4$$

The value of F associated with a 5% level of significance and with degrees of freedom 1 and 4 is found in Appendix D.

$$F_{0.05,1,4} = 7.71$$
$$F_{\text{calculated}} = 9.09$$

Reject H_0 because 9.09 > 7.71

Triple A Construction

- We can conclude there is a statistically significant relationship between X and Y.
- The r^2 value of 0.69 means about 69% of the variability in sales (Y) is explained by local payroll (X).

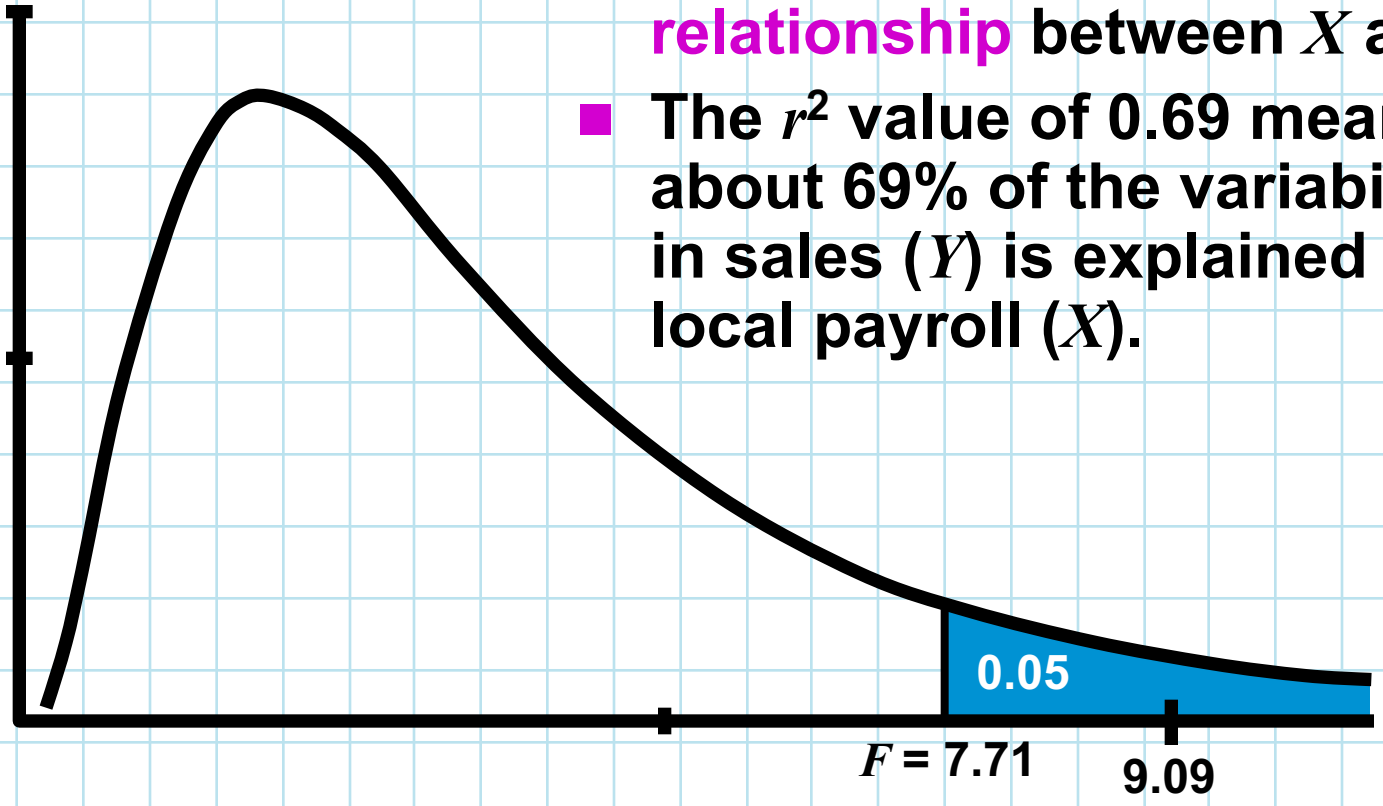


Figure 4.5

Analysis of Variance (ANOVA) Table

- When software is used to develop a regression model, an ANOVA table is typically created that shows the observed significance level (p-value) for the calculated F value.
- This can be compared to the level of significance (α) to make a decision.

	DF	*SS*	*MS*	*F*	SIGNIFICANCE
Regression	k	SSR	$MSR = SSR/k$	MSR/MSE	$P(F > MSR/MSE)$
Residual	$n - k - 1$	SSE	$MSE = SSE/(n - k - 1)$		
Total	$n - 1$	SST			

Table 4.4

ANOVA for Triple A Construction

	D	E	F	G	H	I	J
1	SUMMARY OUTPUT						
2							
3	*Regression Statistics*						
4	Multiple R	0.8333					
5	R Square	0.6944					
6	Adjusted R Square	0.6181					
7	Standard Error	1.3110					
8	Observations	6					
9							
10	ANOVA						
11		*df*	*SS*	*MS*	*F*	*Significance F*	
12	Regression	1	15.6250	15.6250	9.0909	0.0394	
13	Residual	4	6.8750	1.7188			
14	Total	5	22.5				
15							

Program 4.1C (partial)

$$P(F > 9.0909) = 0.0394$$

Because this probability is less than 0.05, we reject the null hypothesis of no linear relationship and conclude there is a linear relationship between X and Y.

Multiple Regression Analysis

- ***Multiple regression models*** **are extensions to the simple linear model and allow the creation of models with more than one independent variable.**

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \varepsilon$$

where

Y = dependent variable (response variable)

X_i = i^{th} independent variable (predictor or explanatory variable)

β_0 = intercept (value of Y when all $X_i = 0$)

β_i = coefficient of the i^{th} independent variable

k = number of independent variables

ε = random error

Multiple Regression Analysis

To estimate these values, a sample is taken the following equation developed

$$\hat{Y} = b_0 + b_1X_1 + b_2X_2 + ... + b_kX_k$$

where

$\hat{Y}$ = predicted value of Y

b_0 = sample intercept (and is an estimate of β_0)

b_i = sample coefficient of the *ith* variable (and is an estimate of β_i)

Jenny Wilson Realty

Jenny Wilson wants to develop a model to determine the suggested listing price for houses based on the size and age of the house.

$$\hat{Y} = b_0 + b_1 X_1 + b_2 X_2$$

where

$\hat{Y}$ = predicted value of dependent variable (selling price)

b_0 = Y intercept

X_1 and X_2 = value of the two independent variables (square footage and age) respectively

b_1 and b_2 = slopes for X_1 and X_2 respectively

She selects a sample of houses that have sold recently and records the data shown in Table 4.5

Jenny Wilson Real Estate Data

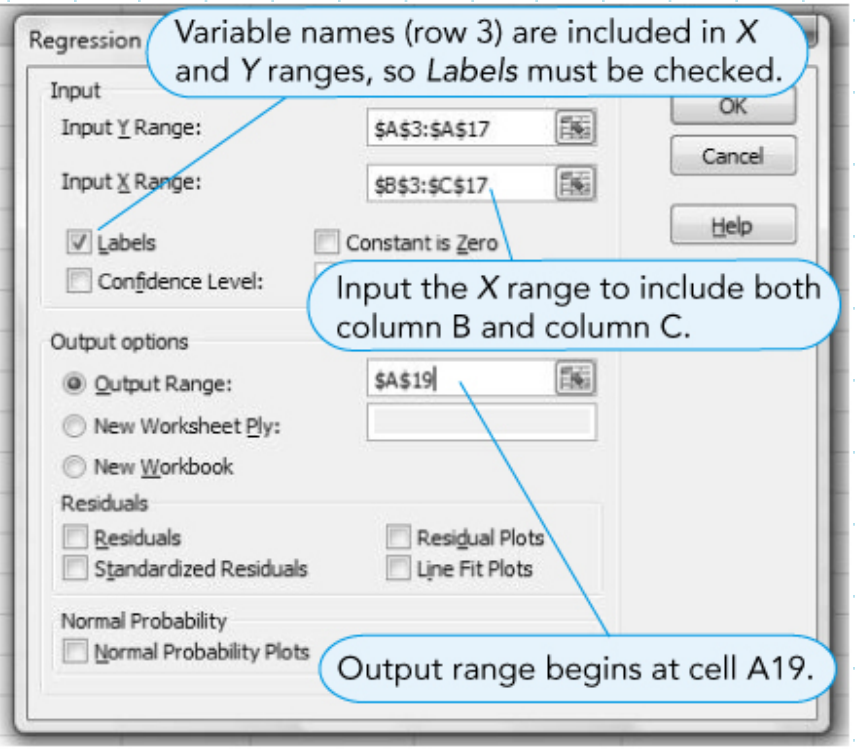
SELLING PRICE ($)	SQUARE FOOTAGE	AGE	CONDITION
95,000	1,926	30	Good
119,000	2,069	40	Excellent
124,800	1,720	30	Excellent
135,000	1,396	15	Good
142,000	1,706	32	Mint
145,000	1,847	38	Mint
159,000	1,950	27	Mint
165,000	2,323	30	Excellent
182,000	2,285	26	Mint
183,000	3,752	35	Good
200,000	2,300	18	Good
211,000	2,525	17	Good
215,000	3,800	40	Excellent
219,000	1,740	12	Mint

Table 4.5

Jenny Wilson Realty

Input Screen for the Jenny Wilson Realty Multiple Regression Example

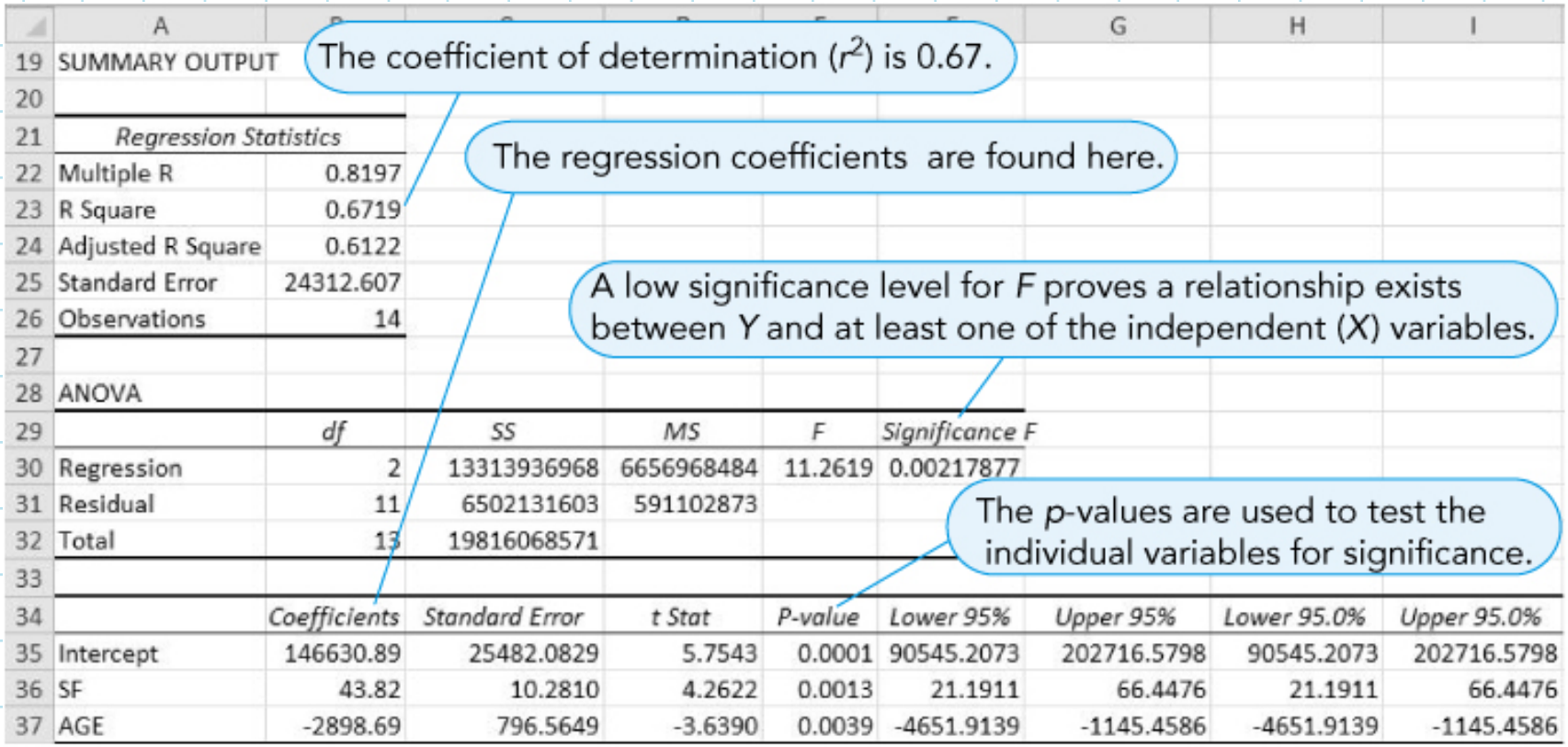
	A	B	C
1	Jenny Wilson Realty		
2			
3	SELL PRICE	SF	AGE
4	95000	1926	30
5	119000	2069	40
6	124800	1720	30
7	135000	1396	15
8	142800	1706	32
9	145000	1847	38
10	159000	1950	27
11	165000	2323	30
12	182000	2285	26
13	183000	3752	35
14	200000	2300	18
15	211000	2525	17
16	215000	3800	40
17	219000	1740	12



Program 4.2A

Jenny Wilson Realty

Output for the Jenny Wilson Realty Multiple Regression Example



	A	B	C	D	E	F	G	H	I
19	SUMMARY OUTPUT								
20									
21	*Regression Statistics*								
22	Multiple R	0.8197							
23	R Square	0.6719							
24	Adjusted R Square	0.6122							
25	Standard Error	24312.607							
26	Observations	14							
27									
28	ANOVA								
29		*df*	*SS*	*MS*	*F*	*Significance F*			
30	Regression	2	13313936968	6656968484	11.2619	0.00217877			
31	Residual	11	6502131603	591102873					
32	Total	13	19816068571						
33									
34		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
35	Intercept	146630.89	25482.0829	5.7543	0.0001	90545.2073	202716.5798	90545.2073	202716.5798
36	SF	43.82	10.2810	4.2622	0.0013	21.1911	66.4476	21.1911	66.4476
37	AGE	-2898.69	796.5649	-3.6390	0.0039	-4651.9139	-1145.4586	-4651.9139	-1145.4586

Program 4.2B

Evaluating Multiple Regression Models

- Evaluation is similar to simple linear regression models.
  - The p-value for the F-test and r^2 are interpreted the same.
- The hypothesis is different because there is more than one independent variable.
  - The F-test is investigating whether all the coefficients are equal to 0 at the same time.

Evaluating Multiple Regression Models

- To determine which independent variables are significant, tests are performed for each variable.

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0$$

- The test statistic is calculated and if the p-value is lower than the level of significance (α), the null hypothesis is rejected.

Jenny Wilson Realty

- The model is statistically significant
  - The *p*-value for the F-test is 0.002.
  - r^2 = 0.6719 so the model explains about 67% of the variation in selling price (Y).
- But the F-test is for the entire model and we can't tell if one or both of the independent variables are significant.
- By calculating the p-value of each variable, we can assess the significance of the individual variables.
- Since the p-value for X_1 (square footage) and X_2 (age) are both less than the significance level of 0.05, both null hypotheses can be rejected.

Binary or Dummy Variables

- *Binary* (or *dummy* or *indicator*) variables are special variables created for qualitative data.
- A dummy variable is assigned a value of 1 if a particular condition is met and a value of 0 otherwise.
- The number of dummy variables must equal one less than the number of categories of the qualitative variable.

Jenny Wilson Realty

- Jenny believes a better model can be developed if she includes information about the condition of the property.

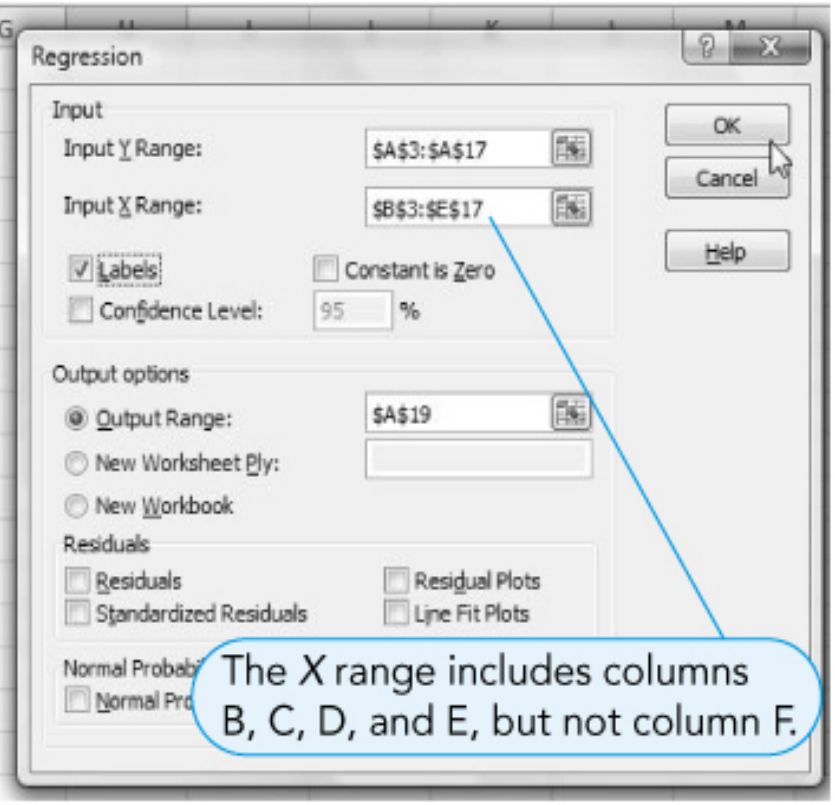
X_3 = 1 if house is in excellent condition
= 0 otherwise
X_4 = 1 if house is in mint condition
= 0 otherwise

- Two dummy variables are used to describe the three categories of condition.
- No variable is needed for "good" condition since if both X_3 and X_4 = 0, the house must be in good condition.

Jenny Wilson Realty

Input Screen for the Jenny Wilson Realty Example with Dummy Variables

	A	B	C	D	E	F
1	Jenny Wilson Realty					
2						
3	SELL PRICE	SF	AGE	X3 (Exc.)	X4 (Mint)	Condition
4	95000	1926	30	0	0	Good
5	119000	2069	40	1	0	Excellent
6	124800	1720	30	1	0	Excellent
7	135000	1396	15	0	0	Good
8	142800	1706	32	0	1	Mint
9	145000	1847	38	0	1	Mint
10	159000	1950	27	0	1	Mint
11	165000	2323	30	1	0	Excellent
12	182000	2285	26	0	1	Mint
13	183000	3752	35	0	0	Good
14	200000	2300	18	0	0	Good
15	211000	2525	17	0	0	Good
16	215000	3800	40	1	0	Excellent
17	219000	1740	12	0	1	Mint



Program 4.3A

Jenny Wilson Realty

Output for the Jenny Wilson Realty Example with Dummy Variables

	A	B	C	D	E	F	G	H	I
19	SUMMARY OUTPUT								
20									
21	*Regression Statistics*								
22	Multiple R	0.9476							
23	R Square	0.8980							
24	Adjusted R Squ	0.8526							
25	Standard Error	14987.5545							
26	Observations	14							
27									
28	ANOVA								
29		*df*	*SS*	*MS*	*F*	*Significance F*			
30	Regression	4	17794427451	4.449E+09	19.804436	0.000174			
31	Residual	9	2021641120	224626791					
32	Total	13	19816068571						
33									
34		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*Upper 95.0%*
35	Intercept	121658.45	17426.61	6.981	0.000	82236.71	161080.19	82236.71	161080.19
36	SF	56.43	6.95	8.122	0.000	40.71	72.14	40.71	72.14
37	AGE	-3962.82	596.03	-6.649	0.000	-5311.13	-2614.51	-5311.13	-2614.51
38	X3 (Exc.)	33162.65	12179.62	2.723	0.023	5610.43	60714.87	5610.43	60714.87
39	X4 (Mint)	47369.25	10649.27	4.448	0.002	23278.93	71459.57	23278.93	71459.57

The coefficient of age is negative, indicating that the price decreases as a house gets older.

The overall model is helpful because the significance *F* probability is low (much less than 5%).

Each of the variables individually is helpful because the *p*-values for each of them is low (much less than 5%).

Program 4.3B

Model Building

- The best model is a statistically significant model with a high r^2 and few variables.
- As more variables are added to the model, the r^2-value usually increases.
- For this reason, the ***adjusted*** r^2 value is often used to determine the usefulness of an additional variable.
- The adjusted r^2 takes into account the number of independent variables in the model.

Model Building

- The formula for r^2

$$r^2 = \frac{\text{SSR}}{\text{SST}} = 1 - \frac{\text{SSE}}{\text{SST}}$$

- The formula for adjusted r^2

$$\text{Adjusted } r^2 = 1 - \frac{\text{SSE}/(n-k-1)}{\text{SST}/(n-1)}$$

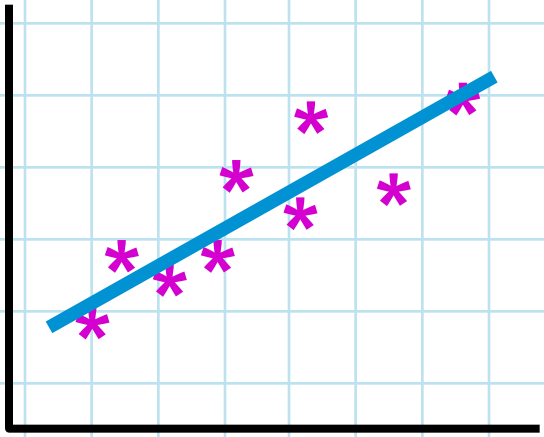
- As the number of variables increases, the adjusted r^2 gets smaller unless the increase due to the new variable is large enough to offset the change in k.

Model Building

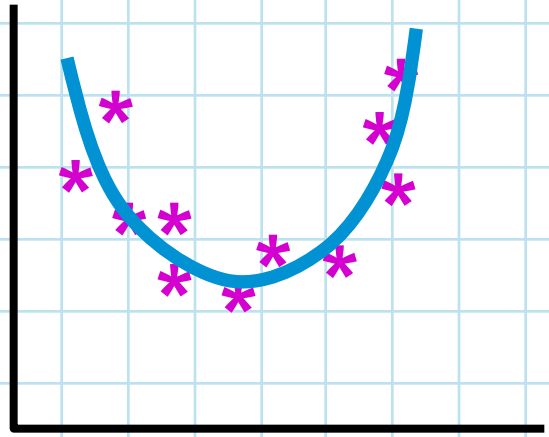
- In general, if a new variable increases the adjusted r^2, it should probably be included in the model.
- In some cases, variables contain duplicate information.
- When two independent variables are correlated, they are said to be *collinear.*
- When more than two independent variables are correlated, *multicollinearity* exists.
- When multicollinearity is present, hypothesis tests for the individual coefficients are not valid but the model may still be useful.

Nonlinear Regression

- In some situations, variables are not linear.
- Transformations may be used to turn a nonlinear model into a linear model.



Linear relationship



Nonlinear relationship

Colonel Motors

- Engineers at Colonel Motors want to use regression analysis to improve fuel efficiency.
- They have been asked to study the impact of weight on miles per gallon (MPG).

MPG	WEIGHT (1,000 LBS.)	MPG	WEIGHT (1,000 LBS.)
12	4.58	20	3.18
13	4.66	23	2.68
15	4.02	24	2.65
18	2.53	33	1.70
19	3.09	36	1.95
19	3.11	42	1.92

Table 4.6

Colonel Motors

Linear Model for MPG Data

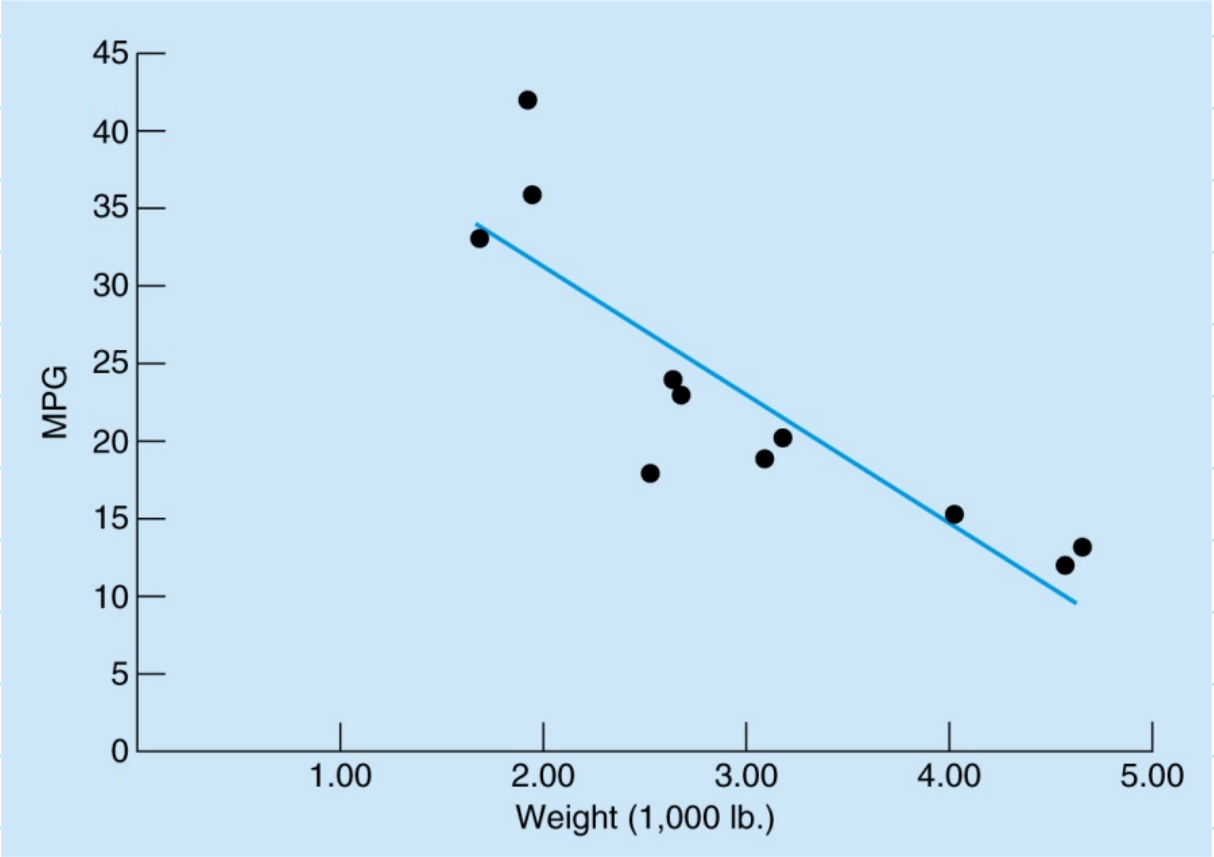


Figure 4.6A

Colonel Motors

Excel Output for Linear Regression Model with MPG Data

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	**Automobile Weight vs. MPG**			SUMMARY OUTPUT									
2													
3	**MPG (Y)**	**Weight (X1)**		*Regression Statistics*									
4	12	4.58		Multiple R	0.8629								
5	13	4.66		R Square	0.7446								
6	15	4.02		Adjusted R Sc	0.7190								
7	18	2.53		Standard Erro	5.0076								
8	19	3.09		Observations	12								
9	19	3.11											
10	20	3.18		ANOVA									
11	23	2.68			*df*	*SS*	*MS*	*F*	*Significance F*				
12	24	2.65		Regression	1	730.9090	730.9090	29.1480	0.0003				
13	33	1.70		Residual	10	250.7577	25.0758						
14	36	1.95		Total	11	981.6667							
15	42	1.92											
16					*Coefficients*	*Standard Err*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0*	*Upper 95.0%*	
17				Intercept	47.6193	4.8132	9.8936	0.0000	36.8950	58.3437	36.8950	58.3437	
18				Weight	-8.2460	1.5273	-5.3989	0.0003	-11.6491	-4.8428	-11.6491	-4.8428	

Program 4.4

This is a useful model with a small F-test for significance and a good r^2 value.

Colonel Motors

Nonlinear Model for MPG Data

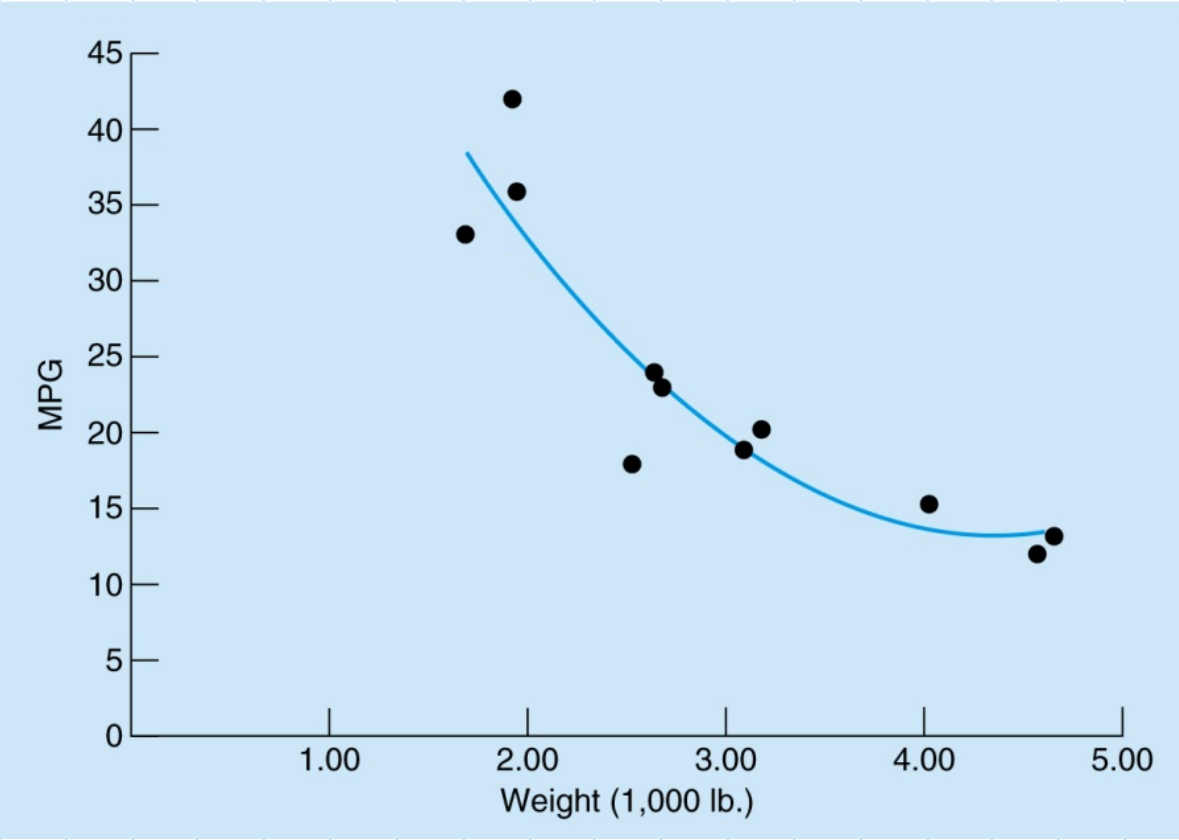


Figure 4.6B

Colonel Motors

- The nonlinear model is a quadratic model.
- The easiest way to work with this model is to develop a new variable.

$$X_2 = (\text{weight})^2$$

- This gives us a model that can be solved with linear regression software:

$$\hat{Y} = b_0 + b_1 X_1 + b_2 X_2$$

Colonel Motors

	A	B	C	D	E	F	G	H	I	J	K	L
1	Automobile Weight vs. MPG			SUMMARY OUTPUT								
2												
3	MPG (Y)	Weight (X1)	WeightSq. (X2)	*Regression Statistics*								
4	12	4.58	20.98	Multiple R	0.9208							
5	13	4.66	21.72	R Square	0.8478							
6	15	4.02	16.16	Adjusted R Sc	0.8140							
7	18	2.53	6.40	Standard Erro	4.0745							
8	19	3.09	9.55	Observations	12							
9	19	3.11	9.67									
10	20	3.18	10.11	ANOVA								
11	23	2.68	7.18		*df*	*SS*	*MS*	*F*	*Significance F*			
12	24	2.65	7.02	Regression	2	832.2557	416.1278	25.0661	0.0002			
13	33	1.70	2.89	Residual	9	149.4110	16.6012					
14	36	1.95	3.80	Total	11	981.6667						
15	42	1.92	3.69									
16					*Coefficient*	*Standard Er*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0*	*Upper 95.0%*
17				Intercept	79.7888	13.5962	5.8685	0.0002	49.0321	110.5454	49.0321	110.5454
18				Weight	-30.2224	8.9809	-3.3652	0.0083	-50.5386	-9.9061	-50.5386	-9.9061
19				Weight2	3.4124	1.3811	2.4708	0.0355	0.2881	6.5367	0.2881	6.5367

$$\hat{Y} = 79.8 - 30.2X_1 + 3.4X_2$$

Program 4.5

A better model with a smaller F-test for significance and a larger adjusted r^2 value

Cautions and Pitfalls

- If the assumptions are not met, the statistical test may not be valid.
- Correlation does not necessarily mean causation.
- Multicollinearity makes interpreting coefficients problematic, but the model may still be good.
- Using a regression model beyond the range of X is questionable, as the relationship may not hold outside the sample data.

Cautions and Pitfalls

- *A* t-test for the intercept (b_0) may be ignored as this point is often outside the range of the model.
- A linear relationship may not be the best relationship, even if the F-test returns an acceptable value.
- A nonlinear relationship can exist even if a linear relationship does not.
- Even though a relationship is statistically significant it may not have any practical value.

Copyright

This work is protected by United States copyright laws and is provided solely for the use of instructors in teaching their courses and assessing student learning. Dissemination or sale of any part of this work (including on the World Wide Web) will destroy the integrity of the work and is not permitted. The work and materials from it should never be made available to students except by instructors using the accompanying text in their classes. All recipients of this work are expected to abide by these restrictions and to honor the intended pedagogical purposes and the needs of other instructors who rely on these materials.