Multiple Regression Analysis

Chapter 14

Learning Objectives

LO14-1 Use multiple regression analysis to describe and interpret a relationship between several independent variables and a dependent variable

LO14-2 Evaluate how well a multiple regression equation fits the data

LO14-3 Test hypothesis about the relationships inferred by a multiple regression model

LO14-4 Evaluate the assumptions of multiple regression

LO14-5 Use and interpret a qualitative, dummy variable in multiple regression

LO14-6 Include and interpret an interaction effect in multiple regression analysis

LO14-7 Apply stepwise regression to develop a multiple regression model

LO14-8 Apply multiple regression techniques to develop a linear model

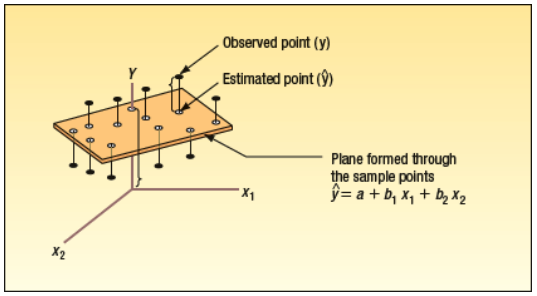
Multiple Regression Analysis

- The general form of a multiple regression formula is

GENERAL MULTIPLE REGRESSION EQUATION

$$\hat{y} = a + b_1x_1 + b_2x_2 + b_3x_3 + \cdots + b_kx_k \quad [14\text{–}1]$$

  - a is the intercept when all x's are zero
  - b refers to the sample regression coefficients
  - x_k refers to the value of the various independent variables
- When there are two independent variables, the relationship can be graphically portrayed as a plane



Multiple Regression Analysis

- The least squares criterion is used to develop the regression equation
- Example
- Suppose the selling price of a home is directly related to the number of rooms and inversely related to its age, let x_1 refer to the number of rooms, x_2 to the age of the home and $\hat{y}$ to the selling price of the home ($000)

$$\hat{y} = 21.2 + 18.7x_1 - .25x_2$$

$$\hat{y} = 21.2 + 18.7(7) - .25(30) = 144.6$$

So, a seven-room house that is 30 years old is expected to sell for $144,600

Multiple Regression Analysis Example

Salsberry Realty sells homes along the East Coast of the United States. One question frequently asked by prospective buyers is "how much can we expect to pay to heat the home in the winter"? The research department at Salsberry thinks 3 variables relate to heating costs: the mean daily outside temperature, the number of inches of insulation, and the age in years of the furnace. They conduct a random sample of 20 homes. Determine the regression equation.

Home	Heating Cost ($)	Mean Outside Temperature (°F)	Attic Insulation (inches)	Age of Furnace (years)
1	$250	35	3	6
2	360	29	4	10
3	165	36	7	3
4	43	60	6	9
5	92	65	5	6
6	200	30	5	5
7	355	10	6	7
8	290	7	10	10
9	230	21	9	11
10	120	55	2	5
11	73	54	12	4
12	205	48	5	1
13	400	20	5	15
14	320	39	4	7
15	72	60	8	6
16	272	20	5	8
17	94	58	7	3
18	190	40	8	11
19	235	27	9	8
20	139	30	7	5

y is the dependent variable

x_1 is the outside temperature

x_2 is inches of insulation

x_3 is the age of the furnace

$\hat{y} = a + b_1x_1 + b_2x_2 + b_3x_3$

$\hat{y}$ is used to estimate the value of y

Multiple Regression Analysis Example

Once we determine the regression equation, we can calculate the heating costs for January, given the mean outside temperature is 30 degrees, there are 5 inches of insulation, and the furnace is 10 years old.

$\hat{y} = a + b_1x_1 + b_2x_2 + b_3x_3$

$\hat{y} = 427.194 - 4.583x_1 - 14.831x_2 + 6.101x_3$

$\hat{y} = 427.194 - 4.583(30) - 14.831(5) + 6.101(10) = 276.56$

Thus, the estimated heating costs for January are $276.56

	A	B	C	D	E	F	G	H	I	J	K
1	Cost	Temp	Insul	Age		SUMMARY OUTPUT					
2	250	35	3	6							
3	360	29	4	10		*Regression Statistics*					
4	165	36	7	3		Multiple R	0.897				
5	43	60	6	9		R Square	0.804				
6	92	65	5	6		Adjusted R Square	0.767				
7	200	30	5	5		Standard Error	51.049				
8	355	10	6	7		Observations	20				
9	290	7	10	10							
10	230	21	9	11		ANOVA					
11	120	55	2	5			*df*	*SS*	*MS*	*F*	*Significance F*
12	73	54	12	4		Regression	3	171220.473	57073.491	21.901	0.000
13	205	48	5	1		Residual	16	41695.277	2605.955		
14	400	20	5	15		Total	19	212915.750			
15	320	39	4	7							
16	72	60	8	6			*Coefficients*	*Standard Error*	*t Stat*	*P-value*	
17	272	20	5	8		Intercept	427.194	59.601	7.168	0.000	
18	94	58	7	3		Temp	-4.583	0.772	-5.934	0.000	
19	190	40	8	11		Insul	-14.831	4.754	-3.119	0.007	
20	235	27	9	8		Age	6.101	4.012	1.521	0.148	
21	139	30	7	5							

Recall:

y is the dependent variable

x_1 is the outside temperature

x_2 is inches of insulation

x_3 is the age of the furnace

$\hat{y}$ is the estimated value of y

ANOVA Table

- An ANOVA table summarizes the multiple regression analysis
- It reports the total amount of the variation divided in two components
  - The regression, the variation in all the independent variables
  - The residual or error, the unexplained variation of y
- It reports the degrees of freedom of the independent variables, the error variation, and the total variation

Source	*df*	SS	MS	*F*
Regression	k	SSR	MSR = SSR/k	MSR/MSE
Residual or error	$n-(k+1)$	SSE	MSE = SSE/$[n-(k+1)]$	
Total	$n-1$	SS total		

Measures of Effectiveness

- There are two measures of effectiveness of the regression equation
- The multiple standard error of the estimate is similar to the standard deviation
- It is measured in the same units as the dependent variable
- It is based on squared deviations between the observed and predicted values of the dependent variable
- It ranges from 0 to plus infinity
- It is calculated from the following equation

MULTIPLE STANDARD ERROR OF ESTIMATE

$$S_{Y.123\ldots k} = \sqrt{\frac{\Sigma(y - \hat{y})^2}{n - (k + 1)}} = \sqrt{\frac{\text{SSE}}{n - (k + 1)}} \qquad [14\text{–}2]$$

ANOVA Table

	A	B	C	D	E	F	G	H	I	J	K
1	Cost	Temp	Insul	Age		SUMMARY OUTPUT					
2	250	35	3	6							
3	360	29	4	10		*Regression Statistics*					
4	165	36	7	3		Multiple R	0.897				
5	43	60	6	9		R Square	0.804				
6	92	65	5	6		Adjusted R Square	0.767				
7	200	30	5	5		Standard Error	51.049				
8	355	10	6	7		Observations	20				
9	290	7	10	10							
10	230	21	9	11		ANOVA					
11	120	55	2	5			*df*	*SS*	*MS*	*F*	*Significance F*
12	73	54	12	4		Regression	3	171220.473	57073.491	21.901	0.000
13	205	48	5	1		Residual	16	41695.277	2605.955		
14	400	20	5	15		Total	19	212915.750			
15	320	39	4	7							
16	72	60	8	6			*Coefficients*	*Standard Error*	*t Stat*	*P-value*	
17	272	20	5	8		Intercept	427.194	59.601	7.168	0.000	
18	94	58	7	3		Temp	-4.583	0.772	-5.934	0.000	
19	190	40	8	11		Insul	-14.831	4.754	-3.119	0.007	
20	235	27	9	8		Age	6.101	4.012	1.521	0.148	
21	139	30	7	5							

$\hat{y} = 427.194 - 4.583x_1 - 14.831x_2 + 6.101x_3$

$\hat{y} = 427.194 - 4.583(35) - 14.831(3) + 6.101(6) = \$258.90$

Then, $(y - \hat{y})^2 = (250 - 258.90)^2 = (8.90)^2 = 79.21$

MULTIPLE STANDARD ERROR OF ESTIMATE

$$S_{Y.123\ldots k} = \sqrt{\frac{\Sigma(y - \hat{y})^2}{n - (k + 1)}} = \sqrt{\frac{SSE}{n - (k + 1)}} \quad [14\text{–}2]$$

$$s_{Y.123\ldots k} = \sqrt{\frac{\sum (y - \hat{y})^2}{n - (k+1)}} = \sqrt{\frac{SSE}{n - (k+1)}} = \sqrt{\frac{41695.277}{16}} = \sqrt{2605.955} = 51.049$$

Measures of Effectiveness

COEFFICIENT OF MULTIPLE DETERMINATION The percent of variation in the dependent variable, y, explained by the set of independent variables, $x_1, x_2, x_3, \ldots x_k$.

- The coefficient of multiple determination
  - Is symbolized by R^2
  - It can range from 0 to 1
  - It cannot assume negative values
  - It is easy to interpret
- It is found by the following formula

COEFFICIENT OF MULTIPLE DETERMINATION

$$R^2 = \frac{\text{SSR}}{\text{SS total}} \qquad [14\text{–}3]$$

$$\text{R2} = \frac{\text{SSR}}{\text{SS total}} = \frac{171{,}220.473}{212{,}915.750} = .804$$

80.4% of the variation is explained by the 3 independent variables.

Measures of Effectiveness

- When the number of independent variables is large, we adjust the coefficient of determination for the degrees of freedom as follows

ADJUSTED COEFFICIENT OF DETERMINATION

$$R^2_{adj} = 1 - \frac{\frac{SSE}{n-(k+1)}}{\frac{SStotal}{n-1}} \qquad [14\text{–}4]$$

- For the cost of heating example, the adjusted coefficient of determination is

$$R^2_{adj} = 1 - \frac{\frac{SSE}{n-(k+1)}}{\frac{SStotal}{n-1}} = 1 - \frac{\frac{41{,}695.277}{20-(3+1)}}{\frac{212{,}915.750}{20-1}} = 1 - \frac{2{,}605.955}{11{,}206.092} = 1 - .233 = .767$$

- If we compare R^2 (0.80) to the adjusted R^2 (0.767), the difference in this case is small

Global Test

- A global test investigates whether it is possible that all the independent variables have zero regression coefficients
- The hypotheses are

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

  $H_1:$ Not all β_{is} are 0
- The test statistic is the F distribution
  - There is a family of F distributions
  - It cannot be negative
  - It is continuous
  - It is positively skewed
  - It is asymptotic

Global Test Continued

- The formula to calculate the value of the test statistic is

GLOBAL TEST

$$F = \frac{SSR/k}{SSE/[n - (k + 1)]} \qquad [14\text{–}5]$$

  - with k (the number of independent variables) degrees of freedom in the numerator
  - n – (k+1) degrees of freedom in the denominator
  - n is sample size
- We can obtain the degrees of freedom from the ANOVA table

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.897
R Square	0.804
Adjusted R Square	0.767
Standard Error	51.049
Observations	20

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	3	171220.473	57073.491	21.901	0.000
Residual	16	41695.277	2605.955		
Total	19	212915.750			

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	427.194	59.601	7.168	0.000
Temp	-4.583	0.772	-5.934	0.000
Insul	-14.831	4.754	-3.119	0.007
Age	6.101	4.012	1.521	0.148

Global Test Concluded

Step 1: State the null and the alternate hypothesis

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

H_1: Not all β_{is} are 0

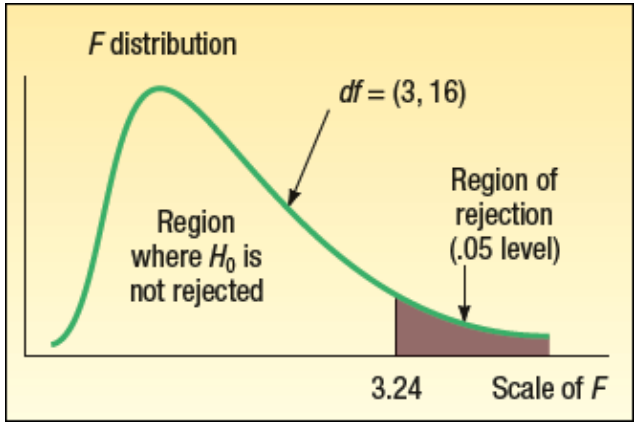
Step 2: Select the level of significance, we'll use .05

Step 3: Select the test statistic, F

Step 4: Formulate the decision rule, reject H_0 if F > 3.24

Step 5: Make decision, reject H_0, F=21.90

Step 6: Interpret, at least one of the independent variables has the ability to explain the variation in heating cost.



$$F = \frac{\text{SSR}/k}{\text{SSE}/[n-(k+1)]} = \frac{171{,}220.473/3}{41{,}695.277/[20-(3+1)]} = 21.90$$

The global test assures us that outside temperature, the amount of insulation, or the age of the furnace has a bearing on heating cost!

Test for Individual Variables

- The test for individual variables determines which independent variables have regression coefficients that differ significantly from zero
- The variables that have zero regression coefficients are usually dropped from the analysis
- The test statistic is the t distribution with n – (k +1) degrees of freedom
- The formula to calculate the value of the test statistic for the individual test is

TESTING INDIVIDUAL REGRESSION COEFFICIENTS

$$t = \frac{b_i - 0}{s_{b_i}} \qquad [14\text{–}6]$$

Evaluating Individual Regression Coefficients Example

Salsberry Realty will use three sets of hypothesis: one for temperature, one for insulation, and one for age of the furnace.
Step 1: State the null and alternate hypothesis

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

$$t = \frac{b_1 - 0}{s_{b1}} = \frac{-4.583 - 0}{0.722} = -5.937$$

For insulation
$H_0: \beta_2 = 0$
$H_1: \beta_2 \neq 0$

$$t = \frac{b_2 - 0}{s_{b2}} = \frac{-14.831 - 0}{4.754} = -3.119$$

For furnace age
$H_0: \beta_3 = 0$
$H_1: \beta_3 \neq 0$

$$t = \frac{b_3 - 0}{s_{b3}} = \frac{6.101 - 0}{4.012} = 1.521$$

Step 2: Select the level of significance, we use .05
Step 3: Select the test statistic, we'll use t
Step 4: Formulate the decision rule, reject H_0 if t < -2.120 or > 2.120
Step 5: Make decision, reject H_0 for temperature and insulation but not furnace age
Step 6: Interpret, furnace age is not a significant predictor of heating costs

Evaluating Individual Regression Coefficients Example

Salsberry Realty will rerun the regression equation using temperature and insulation.

$$\hat{y} = 490.286 - 5.150x_1 - 14.718x_2$$

The hypotheses and details of the global test are, reject the null hypothesis if F > 3.59

$H_0: \beta_1 = \beta_2 = 0$

H_1: Not all of the β_1's are equal

$$F = \frac{SSR-k}{SSE/(n-(k+1)} = \frac{165{,}194.521/2}{47{,}721.229/(20-(2+1))} = 29.424$$

For temperature

$H_0: \beta_1 = 0$

$H_1: \beta_1 \neq 0$

$$t = \frac{b_1-0}{s_{b1}} = \frac{-5.150-0}{0.702} = -7.337$$

For insulation

$H_0: \beta_2 = 0$

$H_1: \beta_2 \neq 0$

$$t = \frac{b_2-0}{s_{b2}} = \frac{-14.718-0}{4.934} = -2.983$$

Step 2: Select the level of significance, we use .05

Step 3: Select the test statistic, we'll use t

Step 4: Formulate the decision rule, reject H_0 if t < -2.110 or > 2.110

Step 5: Make decision, reject H_0 for temperature and insulation

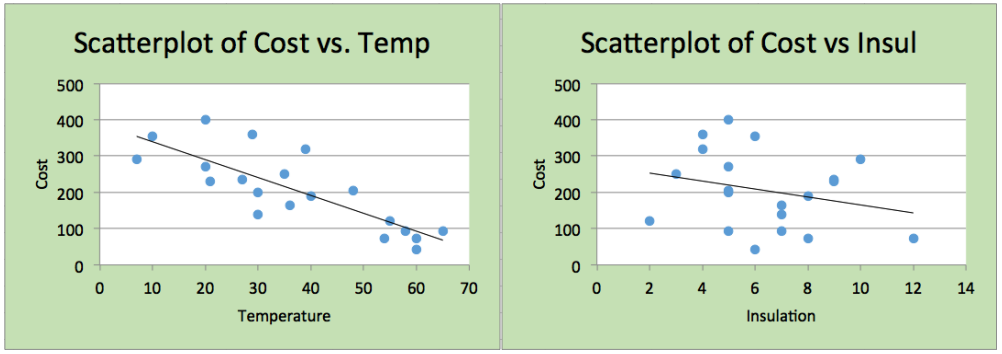
Step 6: Interpret, temperature and insulation are a significant predictor of heating costs

Multiple Regression Assumptions

- There are five assumptions to use multiple regression analysis
  1. There is a linear relationship
  2. The variation in the residuals is the same for both large and small values of $\hat{y}$
  3. The residuals follow the normal distribution
  4. The independent variable should not be correlated
  5. The residuals are independent
- Next, we'll provide a brief discussion of each of these assumptions.

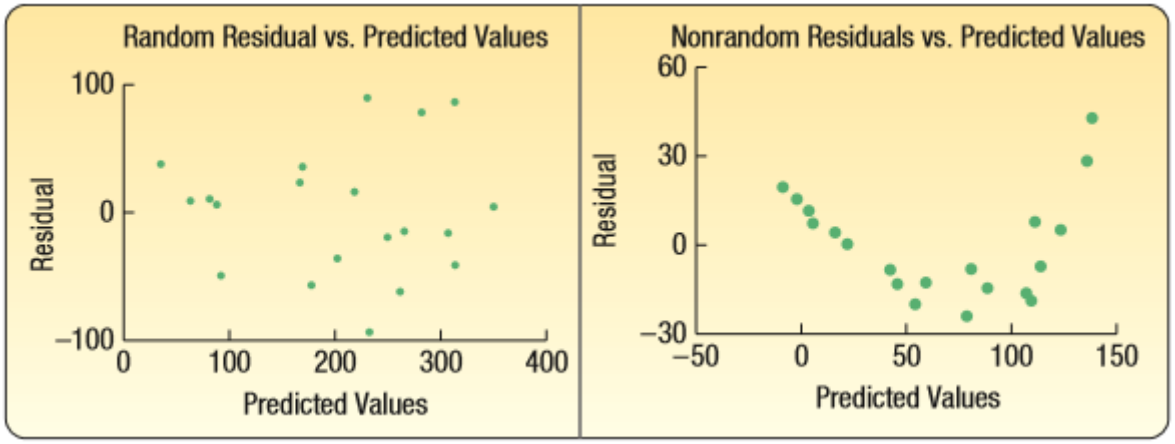
Linear Relationship Assumption

- The relationship between the dependent variable and the set of independent variables must be linear
- To verify this assumption, develop a scatter diagram and plot the dependent variable on the vertical axis and the independent variable on the horizontal axis
- The plots below indicate a fairly strong negative relationship between temperature and heating cost and negative relationship between insulation and costs



Linear Relationship Assumption Continued

- To verify the linearity assumption, we can plot the residuals to evaluate the linearity of the multiple regression equation
- The residuals are plotted on the vertical axis and are centered around zero against the predicted variable $\hat{y}$ on the horizontal axis



Variation Assumption

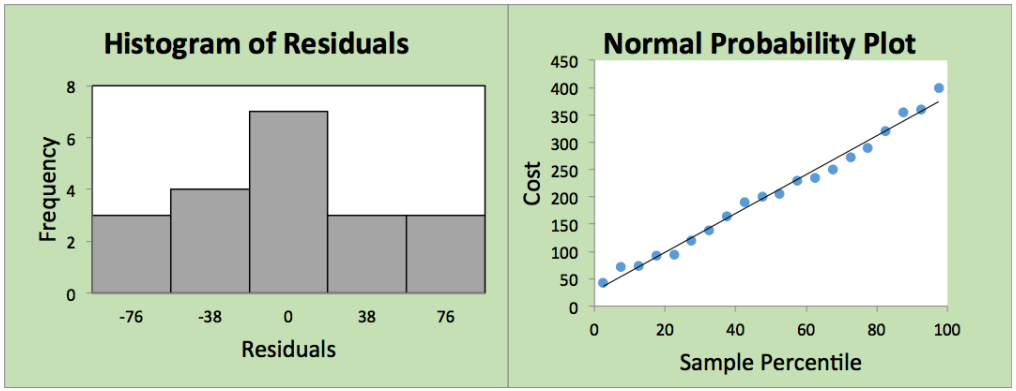
- Variation is the same for both large and small values of $\hat{y}$

HOMOSCEDASTICITY The variation around the regression equation is the same for all of the values of the independent variables.

- This condition is checked by developing a scatter diagram with the residuals on the vertical axis and the fitted values on the horizontal axis
- If there is not a pattern to the plots—that is, they appear random—the residuals meet the homoscedasticity requirement
- We did this on the previous slide and based on the scatter diagram, it is reasonable to conclude that this assumption has **NOT** been violated

Normal Probability Assumption

- The residuals follow the normal probability distribution
- This condition is checked by developing a histogram of the residuals or a normal probability plot
- The mean of the distribution of the residuals is 0
- If the plotted points are fairly close to the straight line drawn from lower left to upper right, the normal probability assumption is supported



Variables Not Correlated Assumption

- The independent variables are not correlated assumption
- A correlation matrix will show all possible correlations among independent variables
- Signs of trouble are correlations > 0.70 or < -0.70
- Signs of correlated independent variables
  - an important predictor variable is found insignificant
  - an obvious reversal occurs in signs in one or more of the independent variables
  - a variable is removed from the solution, there is a large change in the regression coefficients
- The VIF is used to identify correlated independent variables

VARIANCE INFLATION FACTOR $$VIF = \frac{1}{1 - R_j^2}$$ (14–7)

Variables Not Correlated Assumption Example

Refer to table 14-1, which relate the heating cost to the independent variables: outside temperature, amount of insulation, and age of furnace. Develop a correlation matrix for all the independent variables. Does it appear there is a problem with multicollinearity? Find and interpret the VIF for each of the independent variables.

	Cost	*Temp*	*Insul*	*Age*
Cost	1.000			
Temp	-0.812	1.000		
Insul	-0.257	-0.103	1.000	
Age	0.537	-0.486	0.064	1.000

Because all of the correlations are between -.70 and .70, we do not suspect problems with multicollinearity.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.491
R Square	0.241
Adjusted R Square	0.152
Standard Error	16.031
Observations	20

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	1390.291	695.145	2.705	0.096
Residual	17	4368.909	256.995		
Total	19	5759.200			

$$\text{VIF} = \frac{1}{1-R^2_1} = \frac{1}{1-.241} = 1.32$$

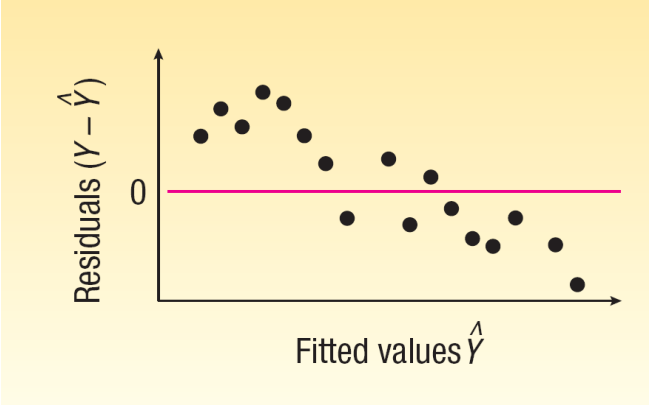
$$\text{VIF} = \frac{1}{1-R^2_2} = \frac{1}{1-.011} = 1.011$$

$$\text{VIF} = \frac{1}{1-R^2_3} = \frac{1}{1-.236} = 1.310$$

All VIFs < 10, no multicollinearity

Independent Observations Assumption

- Each residual is independent of other residuals
- Autocorrelation occurs when successive residuals are correlated
- When autocorrelation exists, the value of the standard error will be biased and will return poor results for tests of hypothesis regarding the regression coefficients



Techniques to Build a Regression Model

- Several techniques help build a regression model

DUMMY VARIABLE A variable in which there are only two possible outcomes. For analysis, one of the outcomes is coded a 1 and the other a 0.

- A dummy or qualitative independent variable can assume one of two possible outcomes, a 1 or a 0
- Use formula (14-6) to determine if the dummy variable should remain in the equation
- Example
- Suppose we are interested in estimating an executive's salary on the basis of years of experience and whether he or she graduated college, graduation will be a yes or no

Dummy Variable Example

Suppose in the Salsberry Realty example that the independent variable garage is added. For homes without a garage, 0 is used; for homes with an attached garage, 1 is used. Garage will be variable x_4. What is the effect of the garage variable?

Cost, y	Temperature, x_1	Insulation, x_2	Garage, x_4
$250	35	3	0
360	29	4	1
165	36	7	0
43	60	6	0
92	65	5	0
200	30	5	0
355	10	6	1
290	7	10	1
230	21	9	0
120	55	2	0
73	54	12	0
205	48	5	1
400	20	5	1
320	39	4	1
72	60	8	0
272	20	5	1
94	58	7	0
190	40	8	1
235	27	9	0
139	30	7	0

	A	B	C	D	E	F	G	H	I	J	K
1	Cost	Temp	Insul	Garage		SUMMARY OUTPUT					
2	250	35	3	0							
3	360	29	4	1		*Regression Statistics*					
4	165	36	7	0		Multiple R	0.933				
5	43	60	6	0		R Square	0.870				
6	92	65	5	0		Adjusted R Square	0.845				
7	200	30	5	0		Standard Error	41.618				
8	355	10	6	1		Observations	20				
9	290	7	10	1							
10	230	21	9	0		ANOVA					
11	120	55	2	0			*df*	*SS*	*MS*	*F*	*Significance F*
12	73	54	12	0		Regression	3	185202.269	61734.090	35.641	0.000
13	205	48	5	1		Residual	16	27713.481	1732.093		
14	400	20	5	1		Total	19	212915.750			
15	320	39	4	1							
16	72	60	8	0			*Coefficients*	*Standard Error*	*t Stat*	*P-value*	
17	272	20	5	1		Intercept	393.666	45.001	8.748	0.000	
18	94	58	7	0		Temp	-3.963	0.653	-6.072	0.000	
19	190	40	8	1		Insul	-11.334	4.002	-2.832	0.012	
20	235	27	9	0		Garage	77.432	22.783	3.399	0.004	

Suppose we have two houses exactly alike in Buffalo, New York. One has an attached garage (1) and the other does not (0). Both have 3 inches of insulation and the temperature is 30 degrees.

Dummy Variable Example Continued

Suppose we have two houses exactly alike in Buffalo, New York. One has an attached garage (1) and the other does not (0). Both have 3 inches of insulation and the temperature is 20 degrees.

For the house without the attached garage:

$\hat{y} = 393.666 - 3.963x_1 - 11.334x_2 + 77.432x_4$

$\hat{y} = 393.666 - 3.963(20) - 11.334\ (3) + 77.432(0) = 280.404$

For the house with the attached garage:

$\hat{y} = 393.666 - 3.963x_1 - 11.334x_2 + 77.432x_4$

$\hat{y} = 393.666 - 3.963(20) - 11.334\ (3) + 77.432(1) = 357.836$

State the null and the alternate hypothesis

H_0: $\beta_4 = 0$

H_1: $\beta_4 \neq 0$

Select the level of significance, .05

Select the test statistic, t

Formula the decision rule, reject H_0 if t < -2.120 or > 2.120

Make decision, reject H_0, t= 3.399

Interpret, the variable garage should be included in the analysis

$$t = \frac{b_4 - 0}{s_{b4}} = \frac{77.432 - 0}{22.783} = 3.399$$

Interaction Technique

- Interaction is the case in which one independent variable (such as x_2) affects the relationship with another independent variable (x_1) and the dependent variable (y)
- In regression analysis, interaction is examined as a separate independent variable, we can multiply the data values of one independent variable by the values of another independent variable

$$Y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$$

- The term $x_1 x_2$ is the interaction term
- Now develop a regression equation with the three variables and test the significance of the third

Interaction Technique Example

Refer to the heating cost example and the data in table 14-1. Is there an interaction between the outside temperature and the amount of insulation? If both variables are increased, is the effect on heating cost greater than the sum of savings from warmer temperature and the savings from increased insulation separately?

$$\hat{y} = 598.070 - 7.811x_1 - 30.161x_2 + 0.385x_1x_2$$

Cost	Temp	Insul	Temp X Insul
250	35	3	105
360	29	4	116
165	36	7	252
43	60	6	360
92	65	5	325
200	30	5	150
355	10	6	60
290	7	10	70
230	21	9	189
120	55	2	110
73	54	12	648
205	48	5	240
400	20	5	100
320	39	4	156
72	60	8	480
272	20	5	100
94	58	7	406
190	40	8	320
235	27	9	243
139	30	7	210

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.893
R Square	0.798
Adjusted R Square	0.760
Standard Error	51.846
Observations	20

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	3	169908.452	56636.151	21.070	0.000
Residual	16	43007.298	2687.956		
Total	19	212915.750			

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	598.070	92.265	6.482	0.000
Temp	-7.811	2.124	-3.678	0.002
Insul	-30.161	12.621	-2.390	0.030
Temp X Insul	0.385	0.291	1.324	0.204

1. It is possible to have a three-way interaction
2. It is possible to have an interaction where one variable is nominal scale

$H_0: \beta_{1X2} = 0$
$H_1: \beta_{1X2} \neq 0$
The level of significance is .05
The test statistic is t
The decision rule is reject H_0 if $r < -2.120$ or > 2.120
Make decision, do not reject H_0
Interpret, there is not a significant interaction between temperature and insulation.

$$t = \frac{b_{1x2} - 0}{s_{b_{1x2}}} = \frac{0.385 - 0}{0.291} = 1.324$$

Stepwise Regression

STEPWISE REGRESSION A step-by-step method to determine a regression equation that begins with a single independent variable and adds or deletes independent variables one by one. Only independent variables with nonzero regression coefficients are included in the regression equation.

- Advantages to the stepwise method
  - Only independent variables with significant regression coefficients are entered into the equation
  - The steps involved are clear
  - It is efficient
  - The changes in the multiple standard error of estimate and the coefficient of determination are shown

Stepwise Regression Technique

The stepwise procedure selects the independent variable temperature first. Temperature explains 65.85% of the variation in heating cost.

$$\hat{y} = 388.8 - 4.93x_1$$

The next independent variable selected is garage. Now the coefficient of determination is 80.46%.

$$\hat{y} = 300.3 - 3.56x_1 + 93.0x_2$$

Next, the procedure selects insulation and stops. At this point 86.98% of the variation is explained.

$$\hat{y} = 393.7 - 3.96x_1 + 77.0x_2 - 11.3\,x_3$$

Worksheet 1 ***

	C1	C2	C3	C4
	Cost	Temp	Insul	Garage
1	250	35	3	0
2	360	29	4	1
3	165	36	7	0
4	43	60	6	0
5	92	65	5	0
6	200	30	5	0
7	355	10	6	1
8	290	7	10	1
9	230	21	9	0
10	120	55	2	0
11	73	54	12	0
12	205	48	5	1
13	400	20	5	1
14	320	39	4	1
15	72	60	8	0
16	272	20	5	1
17	94	58	7	0
18	190	40	8	1
19	235	27	9	0
20	139	30	7	0

Session

Stepwise Regression: Cost versus Temp, Insul, Garage

Alpha-to-Enter: 0.15 Alpha-to-Remove: 0.15

Response is Cost on 3 predictors, with N = 20

Step	1	2	3
Constant	388.8	300.3	393.7
Temp	-4.93	-3.56	-3.96
T-Value	-5.89	-4.70	-6.07
P-Value	0.000	0.000	0.000
Garage		93	77
T-Value		3.56	3.40
P-Value		0.002	0.004
Insul			-11.3
T-Value			-2.83
P-Value			0.012
S	63.6	49.5	41.6
R-Sq	65.85	80.46	86.98
R-Sq(adj)	63.96	78.16	84.54
Mallows Cp	26.0	10.0	4.0

This is the same regression equation we developed before!