Basic Business Statistics
11th Edition

Chapter 14

Introduction to Multiple Regression

Basic Business Statistics, 11e © 2009 Prentice-Hall, Inc.

Learning Objectives

In this chapter, you learn:

- How to develop a multiple regression model
- How to interpret the regression coefficients
- How to determine which independent variables to include in the regression model
- How to determine which independent variables are more important in predicting a dependent variable
- How to use categorical variables in a regression model
- How to predict a categorical dependent variable using logistic regression

Basic Business Statistics, 11e © 2009 Prentice-Hall, Inc..

 © 2006 Prentice Hall, Inc.

The Multiple Regression Model

Idea: Examine the linear relationship between 1 dependent (Y) & 2 or more independent variables (X_i)

Multiple Regression Model with k Independent Variables:

Y-intercept

Population slopes

Random Error

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + \varepsilon_i$$

Multiple Regression Equation

The coefficients of the multiple regression model are estimated using sample data

Multiple regression equation with k independent variables:

Estimated (or predicted) value of Y

Estimated intercept

Estimated slope coefficients

$$\hat{Y}_i = b_0 + b_1 X_{1i} + b_2 X_{2i} + \cdots + b_k X_{ki}$$

In this chapter we will use Excel or Minitab to obtain the regression slope coefficients and other regression summary measures.

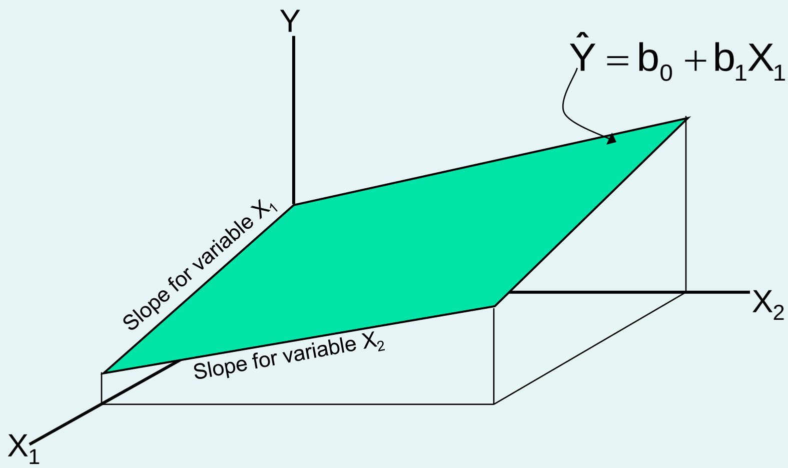
© 2006 Prentice Hall, Inc.

Multiple Regression Equation

(continued)

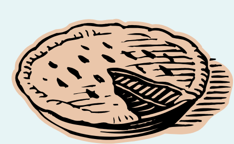
Two variable model

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



Example: 2 Independent Variables

- A distributor of frozen dessert pies wants to evaluate factors thought to influence demand
  - Dependent variable: Pie sales (units per week)
  - Independent variables: Price (in $); Advertising ($100's)
- Data are collected for 15 weeks



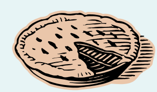
 © 2006 Prentice Hall, Inc.

Pie Sales Example

Week	Pie Sales	Price ($)	Advertising ($100s)
1	350	5.50	3.3
2	460	7.50	3.3
3	350	8.00	3.0
4	430	8.00	4.5
5	350	6.80	3.0
6	380	7.50	4.0
7	430	4.50	3.0
8	470	6.40	3.7
9	450	7.00	3.5
10	490	5.00	4.0
11	340	7.20	3.5
12	300	7.90	3.2
13	440	5.90	4.0
14	450	5.00	3.5
15	300	7.00	2.7

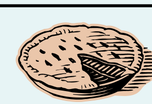
Multiple regression equation:

$$\widehat{Sales} = b_0 + b_1\,(Price) + b_2\,(Advertising)$$



Basic Business Statistics, 11e © 2009 Prentice-Hall, Inc.. Chap 14-7

Excel Multiple Regression Output



Regression Statistics	
Multiple R	0.72213
R Square	0.52148
Adjusted R Square	0.44172
Standard Error	47.46341
Observations	15

$$\widehat{Sales} = 306.526 - 24.975(Price) + 74.131(Advertising)$$

ANOVA	*df*	*SS*	*MS*	*F*	*Significance F*
Regression	2	29460.027	14730.013	6.53861	0.01201
Residual	12	27033.306	2252.776		
Total	14	56493.333			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	306.52619	114.25389	2.68285	0.01993	57.58835	555.46404
Price	-24.97509	10.83213	-2.30565	0.03979	-48.57626	-1.37392
Advertising	74.13096	25.96732	2.85478	0.01449	17.55303	130.70888

Basic Business Statistics, 11e © 2009 Prentice-Hall, Inc.. Chap 14-8

© 2006 Prentice Hall, Inc.

Minitab Multiple Regression Output

$$\text{Sales} = 306.526 - 24.975(\text{Price}) + 74.131(\text{Advertising})$$

```
The regression equation is
Sales = 307 - 25.0 Price + 74.1 Advertising

Predictor     Coef  SE Coef      T      P
Constant    306.50   114.30   2.68  0.020
Price       -24.98    10.83  -2.31  0.040
Advertising  74.13    25.97   2.85  0.014

S = 47.4634   R-Sq = 52.1%   R-Sq(adj) = 44.2%

Analysis of Variance

Source          DF     SS     MS     F      P
Regression       2  29460  14730  6.54  0.012
Residual Error  12  27033   2253
Total           14  56493
```

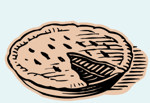

The Multiple Regression Equation

$$\widehat{\text{Sales}} = 306.526 - 24.975(\text{Price}) + 74.131(\text{Advertising})$$

where
- Sales is in number of pies per week
- Price is in $
- Advertising is in $100's.

b_1 = -24.975: sales will decrease, on average, by 24.975 pies per week for each $1 increase in selling price, net of the effects of changes due to advertising

b_2 = 74.131: sales will increase, on average, by 74.131 pies per week for each $100 increase in advertising, net of the effects of changes due to price



© 2006 Prentice Hall, Inc.

Using The Equation to Make Predictions

Predict sales for a week in which the selling price is $5.50 and advertising is $350:

$$\widehat{\text{Sales}} = 306.526 - 24.975(\text{Price}) + 74.131(\text{Advertising})$$
$$= 306.526 - 24.975(5.50) + 74.131(3.5)$$
$$= 428.62$$

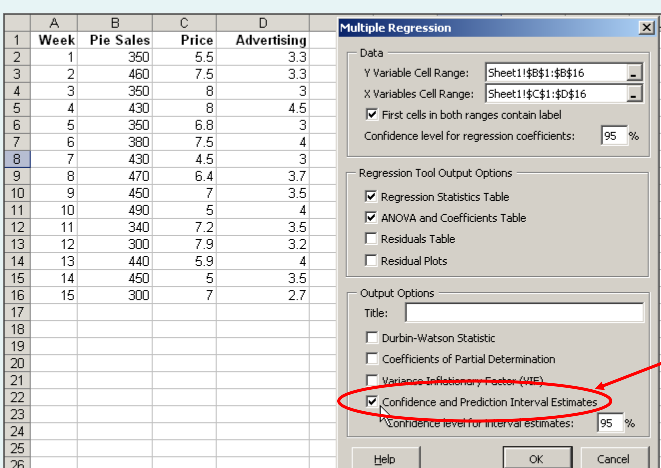
Predicted sales is 428.62 pies

Note that Advertising is in $100's, so $350 means that $X_2 = 3.5$

Basic Business Statistics, 11e © 2009 Prentice-Hall, Inc.. Chap 14-11

Predictions in Excel using PHStat

- PHStat | regression | multiple regression …



Week	Pie Sales	Price	Advertising
1	350	5.5	3.3
2	460	7.5	3.3
3	350	8	3
4	430	8	4.5
5	350	6.8	3
6	380	7.5	4
7	430	4.5	3
8	470	6.4	3.7
9	450	7	3.5
10	490	5	4
11	340	7.2	3.5
12	300	7.9	3.2
13	440	5.9	4
14	450	5	3.5
15	300	7	2.7

Check the "confidence and prediction interval estimates" box

Basic Business Statistics, 11e © 2009 Prentice-Hall, Inc.. Chap 14-12

© 2006 Prentice Hall, Inc.

Predictions in PHStat

(continued)

	A	B
1	Confidence and Prediction Estimate Intervals	
2		
3	Data	
4	Confidence Level	95%
5		
6	Price given value	5.5
7	Advertising given value	3.5
8		
20	t Statistic	2.178813
21	Predicted Y (YHat)	428.6216
22		
23	For Average Predicted Y (Yhat)	
24	Interval Half Width	37.50306
25	Confidence Interval Lower Limit	391.1185
26	Confidence Interval Upper Limit	466.1246
27		
28	For Individual Response Y	
29	Interval Half Width	110.0041
30	Prediction Interval Lower Limit	318.6174
31	Prediction Interval Upper Limit	538.6257

Input values

Predicted $\hat{Y}$ value

Confidence interval for the mean value of Y, given these X values

Prediction interval for an individual Y value, given these X values

Basic Business Statistics, 11e © 2009 Prentice-Hall, Inc.. Chap 14-13

Predictions in Minitab

Confidence interval for the mean value of Y, given these X values

```
Predicted Values for New Observations

New
Obs    Fit  SE Fit      95% CI            95% PI
  1  428.6    17.2  (391.1, 466.1)  (318.6, 538.6)

Values of Predictors for New Observations

New
Obs  Price  Advertising
  1   5.50         3.50
```

Predicted $\hat{Y}$ value

Input values

Prediction interval for an individual Y value, given these X values

Basic Business Statistics, 11e © 2009 Prentice-Hall, Inc.. Chap 14-14

 © 2006 Prentice Hall, Inc.

Coefficient of Multiple Determination

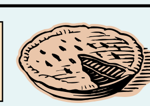
- Reports the proportion of total variation in Y explained by all X variables taken together

$$r^2 = \frac{SSR}{SST} = \frac{\text{regression sum of squares}}{\text{total sum of squares}}$$

Multiple Coefficient of Determination In Excel

Regression Statistics	
Multiple R	0.72213
R Square	0.52148
Adjusted R Square	0.44172
Standard Error	47.46341
Observations	15

$$r^2 = \frac{SSR}{SST} = \frac{29460.0}{56493.3} = .52148$$



52.1% of the variation in pie sales is explained by the variation in price and advertising

ANOVA	*df*	*SS*	*MS*	*F*	*Significance F*
Regression	2	29460.027	14730.013	6.53861	0.01201
Residual	12	27033.306	2252.776		
Total	14	56493.333			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	306.52619	114.25389	2.68285	0.01993	57.58835	555.46404
Price	-24.97509	10.83213	-2.30565	0.03979	-48.57626	-1.37392
Advertising	74.13096	25.96732	2.85478	0.01449	17.55303	130.70888

© 2006 Prentice Hall, Inc.

Multiple Coefficient of Determination In Minitab

```
The regression equation is
Sales = 307 - 25.0 Price + 74.1 Advertising

Predictor     Coef  SE Coef      T      P
Constant    306.50   114.30   2.68  0.020
Price       -24.98    10.83  -2.31  0.040
Advertising  74.13    25.97   2.85  0.014

S = 47.4634   R-Sq = 52.1%   R-Sq(adj) = 44.2%

Analysis of Variance

Source          DF     SS     MS     F      P
Regression       2  29460  14730  6.54  0.012
Residual Error  12  27033   2253
Total           14  56493
```

$$r^2 = \frac{SSR}{SST} = \frac{29460.0}{56493.3} = .52148$$

52.1% of the variation in pie sales is explained by the variation in price and advertising

Adjusted r^2

- r^2 never decreases when a new X variable is added to the model
  - This can be a disadvantage when comparing models
- What is the net effect of adding a new variable?
  - We lose a degree of freedom when a new X variable is added
  - Did the new X variable add enough explanatory power to offset the loss of one degree of freedom?

 © 2006 Prentice Hall, Inc.

Adjusted r^2

(continued)

- Shows the proportion of variation in Y explained by all X variables adjusted for the number of X variables used

$$r_{adj}^2 = 1 - \left[(1 - r^2) \left(\frac{n-1}{n-k-1} \right) \right]$$

(where n = sample size, k = number of independent variables)

- Penalize excessive use of unimportant independent variables
- Smaller than r^2
- Useful in comparing among models

Adjusted r^2 in Excel

Regression Statistics	
Multiple R	0.72213
R Square	0.52148
Adjusted R Square	0.44172
Standard Error	47.46341
Observations	15

$$r_{adj}^2 = .44172$$

44.2% of the variation in pie sales is explained by the variation in price and advertising, taking into account the sample size and number of independent variables

ANOVA	*df*	*SS*	*MS*	*F*	*Significance F*
Regression	2	29460.027	14730.013	6.53861	0.01201
Residual	12	27033.306	2252.776		
Total	14	56493.333			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	306.52619	114.25389	2.68285	0.01993	57.58835	555.46404
Price	-24.97509	10.83213	-2.30565	0.03979	-48.57626	-1.37392
Advertising	74.13096	25.96732	2.85478	0.01449	17.55303	130.70888

 © 2006 Prentice Hall, Inc.

Adjusted r^2 in Minitab

The regression equation is
Sales = 307 - 25.0 Price + 74.1 Advertising

Predictor	Coef	SE Coef	T	P
Constant	306.50	114.30	2.68	0.020
Price	-24.98	10.83	-2.31	0.040
Advertising	74.13	25.97	2.85	0.014

S = 47.4634 R-Sq = 52.1% R-Sq(adj) = 44.2%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	2	29460	14730	6.54	0.012
Residual Error	12	27033	2253		
Total	14	56493			

$$r^2_{adj} = .44172$$

44.2% of the variation in pie sales is explained by the variation in price and advertising, taking into account the sample size and number of independent variables

Is the Model Significant?

- F Test for Overall Significance of the Model
- Shows if there is a linear relationship between all of the X variables considered together and Y
- Use F-test statistic
- Hypotheses:

H_0: $\beta_1 = \beta_2 = \cdots = \beta_k = 0$ (no linear relationship)

H_1: at least one $\beta_i \neq 0$ (at least one independent variable affects Y)

F Test for Overall Significance

- Test statistic:

$$F_{STAT} = \frac{MSR}{MSE} = \frac{\dfrac{SSR}{k}}{\dfrac{SSE}{n-k-1}}$$

where F_{STAT} has numerator d.f. = k and denominator d.f. = $(n - k - 1)$

F Test for Overall Significance In Excel

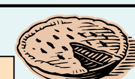
(continued)

Regression Statistics	
Multiple R	**0.72213**
R Square	**0.52148**
Adjusted R Square	**0.44172**
Standard Error	**47.46341**
Observations	**15**

$$F_{STAT} = \frac{MSR}{MSE} = \frac{14730.0}{2252.8} = 6.5386$$

With 2 and 12 degrees of freedom

P-value for the F Test



ANOVA	*df*	*SS*	*MS*	*F*	*Significance F*
Regression	**2**	**29460.027**	**14730.013**	**6.53861**	**0.01201**
Residual	**12**	**27033.306**	**2252.776**		
Total	**14**	**56493.333**			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	**306.52619**	**114.25389**	**2.68285**	**0.01993**	**57.58835**	**555.46404**
Price	**-24.97509**	**10.83213**	**-2.30565**	**0.03979**	**-48.57626**	**-1.37392**
Advertising	**74.13096**	**25.96732**	**2.85478**	**0.01449**	**17.55303**	**130.70888**

F Test for Overall Significance In Minitab

```
The regression equation is
Sales = 307 - 25.0 Price + 74.1 Advertising

Predictor      Coef  SE Coef     T      P
Constant     306.50   114.30  2.68  0.020
Price        -24.98    10.83 -2.31  0.040
Advertising   74.13    25.97  2.85  0.014

S = 47.4634   R-Sq = 52.1%   R-Sq(adj) = 44.2%

Analysis of Variance

Source          DF     SS     MS     F      P
Regression       2  29460  14730  6.54  0.012
Residual Error  12  27033   2253
Total           14  56493
```

$$F_{STAT} = \frac{MSR}{MSE} = \frac{14730.0}{2252.8} = 6.5386$$

With 2 and 12 degrees of freedom

P-value for the F Test

Basic Business Statistics, 11e © 2009 Prentice-Hall, Inc.. Chap 14-25

F Test for Overall Significance

(continued)

H_0: $\beta_1 = \beta_2 = 0$

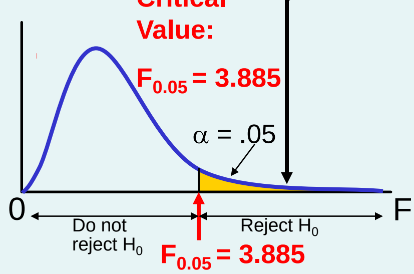
H_1: β_1 and β_2 not both zero

$\alpha = .05$

$df_1 = 2$ $df_2 = 12$

Critical Value:

$F_{0.05} = 3.885$



Test Statistic:

$$F_{STAT} = \frac{MSR}{MSE} = 6.5386$$

Decision:

Since F_{STAT} test statistic is in the rejection region (p-value < .05), reject H_0

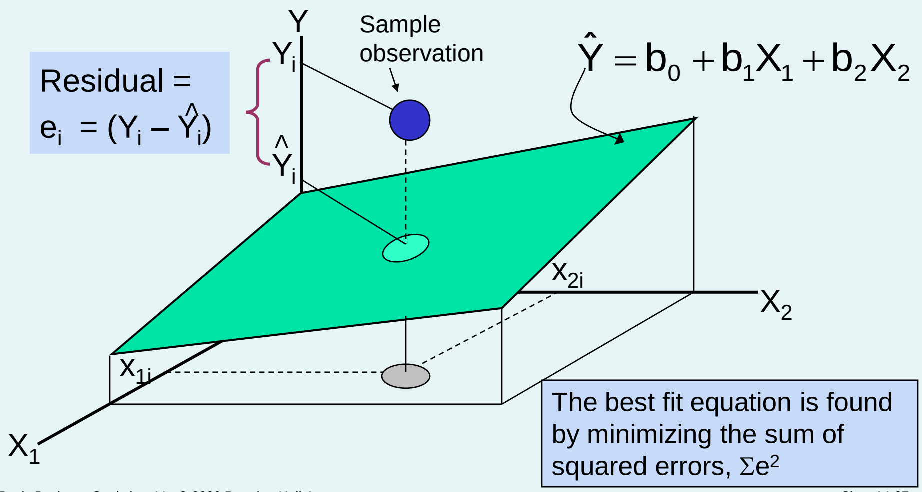
Conclusion:

There is evidence that at least one independent variable affects Y

Basic Business Statistics, 11e © 2009 Prentice-Hall, Inc.. Chap 14-26

Residuals in Multiple Regression

Two variable model



Multiple Regression Assumptions

Errors (residuals) from the regression model:

$$e_i = (Y_i - \hat{Y}_i)$$

Assumptions:

- The errors are normally distributed
- Errors have a constant variance
- The model errors are independent

 © 2006 Prentice Hall, Inc.

Residual Plots Used in Multiple Regression

- These residual plots are used in multiple regression:
  - Residuals vs. $\hat{Y}_i$
  - Residuals vs. X_{1i}
  - Residuals vs. X_{2i}
  - Residuals vs. time (if time series data)

Use the residual plots to check for violations of regression assumptions

Are Individual Variables Significant?

- Use t tests of individual variable slopes
- Shows if there is a linear relationship between the variable X_j and Y holding constant the effects of other X variables
- Hypotheses:
  - H_0: $\beta_j = 0$ (no linear relationship)
  - H_1: $\beta_j \neq 0$ (linear relationship does exist between X_j and Y)

 © 2006 Prentice Hall, Inc.

Are Individual Variables Significant?

(continued)

H_0: $\beta_j = 0$ (no linear relationship)

H_1: $\beta_j \neq 0$ (linear relationship does exist between X_j and Y)

Test Statistic:

$$t_{STAT} = \frac{b_j - 0}{S_{b_j}}$$

(df = n – k – 1)

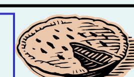
Are Individual Variables Significant? Excel Output

(continued)

Regression Statistics	
Multiple R	0.72213
R Square	0.52148
Adjusted R Square	0.44172
Standard Error	47.46341
Observations	15

t Stat for Price is t_{STAT} = -2.306, with p-value .0398

t Stat for Advertising is t_{STAT} = 2.855, with p-value .0145



ANOVA	*df*	*SS*	*MS*	*F*	*Significance F*
Regression	2	29460.027	14730.013	6.53861	0.01201
Residual	12	27033.306	2252.776		
Total	14	56493.333			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	306.52619	114.25389	2.68285	0.01993	57.58835	555.46404
Price	-24.97509	10.83213	-2.30565	0.03979	-48.57626	-1.37392
Advertising	74.13096	25.96732	2.85478	0.01449	17.55303	130.70888

Are Individual Variables Significant? Minitab Output

The regression equation is
Sales = 307 - 25.0 Price + 74.1 Advertising

Predictor	Coef	SE Coef	T	P
Constant	306.50	114.30	2.68	0.020
Price	-24.98	10.83	-2.31	0.040
Advertising	74.13	25.97	2.85	0.014

S = 47.4634 R-Sq = 52.1% R-Sq(adj) = 44.2%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	2	29460	14730	6.54	0.012
Residual Error	12	27033	2253		
Total	14	56493			

t Stat for Price is t_{STAT} = -2.306, with p-value .0398

t Stat for Advertising is t_{STAT} = 2.855, with p-value .0145

Inferences about the Slope: t Test Example

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

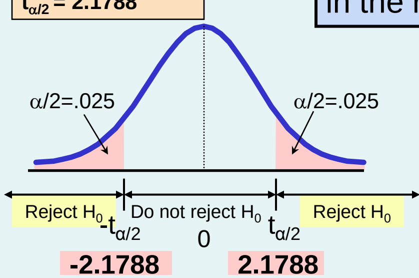
d.f. = 15-2-1 = 12
α = .05
$t_{\alpha/2}$ = 2.1788

From the Excel and Minitab output:

For Price t_{STAT} = -2.306, with p-value .0398

For Advertising t_{STAT} = 2.855, with p-value .0145

The test statistic for each variable falls in the rejection region (p-values < .05)



Decision:

Reject H_0 for each variable

Conclusion:

There is evidence that both Price and Advertising affect pie sales at α = .05

 © 2006 Prentice Hall, Inc.

Confidence Interval Estimate for the Slope

Confidence interval for the population slope β_j

$$\boxed{b_j \pm t_{\alpha/2} S_{b_j}}$$

where t has (n – k – 1) d.f.

	Coefficients	*Standard Error*
Intercept	**306.52619**	**114.25389**
Price	**-24.97509**	**10.83213**
Advertising	**74.13096**	**25.96732**

Here, t has (15 – 2 – 1) = 12 d.f.

Example: Form a 95% confidence interval for the effect of changes in price (X_1) on pie sales:

-24.975 ± (2.1788)(10.832)

So the interval is (-48.576 , -1.374)

(This interval does not contain zero, so price has a significant effect on sales)

Confidence Interval Estimate for the Slope

(continued)

Confidence interval for the population slope β_j

	Coefficients	*Standard Error*	…	*Lower 95%*	*Upper 95%*
Intercept	**306.52619**	**114.25389**	…	**57.58835**	**555.46404**
Price	**-24.97509**	**10.83213**	…	**-48.57626**	**-1.37392**
Advertising	**74.13096**	**25.96732**	…	**17.55303**	**130.70888**

Example: Excel output also reports these interval endpoints:

Weekly sales are estimated to be reduced by between 1.37 to 48.58 pies for each increase of $1 in the selling price, holding the effect of price constant

Testing Portions of the Multiple Regression Model

- Contribution of a Single Independent Variable X_j

SSR(X_j | all variables except X_j)
= SSR (all variables) – SSR(all variables except X_j)

- Measures the contribution of X_j in explaining the total variation in Y (SST)

Basic Business Statistics, 11e © 2009 Prentice-Hall, Inc.. Chap 14-37

Testing Portions of the Multiple Regression Model

(continued)

Contribution of a Single Independent Variable X_j, assuming all other variables are already included (consider here a 2-variable model):

SSR(X_1 | X_2)
= SSR (all variables) – SSR(X_2)

From ANOVA section of regression for

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

From ANOVA section of regression for

$$\hat{Y} = b_0 + b_2X_2$$

Measures the contribution of X_1 in explaining SST

Basic Business Statistics, 11e © 2009 Prentice-Hall, Inc.. Chap 14-38

© 2006 Prentice Hall, Inc.

The Partial F-Test Statistic

- Consider the hypothesis test:

H_0: variable X_j does not significantly improve the model after all other variables are included

H_1: variable X_j significantly improves the model after all other variables are included

- Test using the F-test statistic:

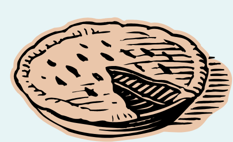
(with 1 and n-k-1 d.f.)

$$F_{STAT} = \frac{SSR\ (X_j \mid \text{all variables except } j)}{MSE}$$

Testing Portions of Model: Example

Example: Frozen dessert pies

Test at the α = .05 level to determine whether the price variable significantly improves the model given that advertising is included



Testing Portions of Model: Example

(continued)

H_0: X_1 (price) does not improve the model with X_2 (advertising) included

H_1: X_1 does improve model

$\alpha = .05$, df = 1 and 12

$F_{0.05} = 4.75$

(For X_1 and X_2)

ANOVA			
	df	*SS*	*MS*
Regression	2	**29460.02687**	14730.01343
Residual	12	27033.30647	**2252.775539**
Total	14	56493.33333	

(For X_2 only)

ANOVA		
	df	*SS*
Regression	1	**17484.22249**
Residual	13	39009.11085
Total	14	56493.33333

Basic Business Statistics, 11e © 2009 Prentice-Hall, Inc.. Chap 14-41

Testing Portions of Model: Example

(continued)

(For X_1 and X_2)

ANOVA			
	df	*SS*	*MS*
Regression	2	**29460.02687**	14730.01343
Residual	12	27033.30647	**2252.775539**
Total	14	56493.33333	

(For X_2 only)

ANOVA		
	df	*SS*
Regression	1	**17484.22249**
Residual	13	39009.11085
Total	14	56493.33333

$$F_{STAT} = \frac{\text{SSR }(X_1 | X_2)}{\text{MSE(all)}} = \frac{29{,}460.03 - 17{,}484.22}{2252.78} = 5.316$$

Conclusion: Since $F_{STAT} = 5.316 > F_{0.05} = 4.75$ Reject H_0; Adding X_1 does improve model

Basic Business Statistics, 11e © 2009 Prentice-Hall, Inc.. Chap 14-42

Relationship Between Test Statistics

- The partial *F* test statistic developed in this section and the *t* test statistic are both used to determine the contribution of an independent variable to a multiple regression model.
- The hypothesis tests associated with these two statistics always result in the same decision (that is, the *p*-values are identical).

$$t_a^2 = F_{1,a}$$

Where a = degrees of freedom

Coefficient of Partial Determination for k variable model

$$r^2_{Yj.(\text{all variables except j})} = \frac{SSR(X_j \mid \text{all variables except j})}{SST - SSR(\text{all variables}) + SSR(X_j \mid \text{all variables except j})}$$

- Measures the proportion of variation in the dependent variable that is explained by X_j while controlling for (holding constant) the other independent variables

© 2006 Prentice Hall, Inc.

Coefficient of Partial Determination in Excel

- Coefficients of Partial Determination can be found using Excel:
  - PHStat | regression | multiple regression …
    - Check the "coefficient of partial determination" box

Regression Analysis
Coefficients of Partial Determination

Intermediate Calculations			
SSR(X1,X2)	29460.02687		
SST	56493.33333		
SSR(X2)	17484.22249	SSR(X1 | X2)	11975.80438
SSR(X1)	11100.43803	SSR(X2 | X1)	18359.58884

Coefficients	
r2 Y1.2	0.307000188
r2 Y2.1	0.404459524

Using Dummy Variables

- A dummy variable is a categorical independent variable with two levels:
  - yes or no, on or off, male or female
  - coded as 0 or 1
- Assumes the slopes associated with numerical independent variables do not change with the value for the categorical variable
- If more than two levels, the number of dummy variables needed is (number of levels - 1)

Dummy-Variable Example (with 2 Levels)

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

Let:

Y = pie sales

X_1 = price

X_2 = holiday ($X_2 = 1$ if a holiday occurred during the week)
($X_2 = 0$ if there was no holiday that week)

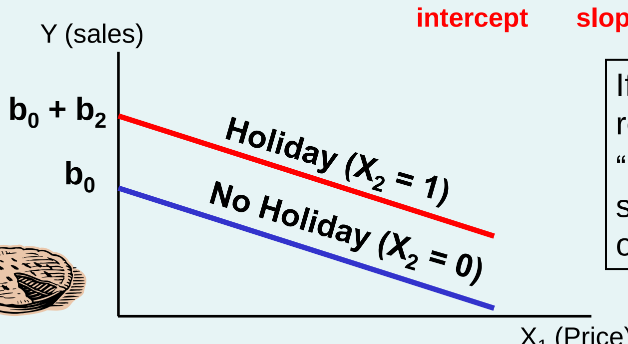
Dummy-Variable Example (with 2 Levels)

(continued)

$$\hat{Y} = b_0 + b_1 X_1 + b_2(1) = \boxed{(b_0 + b_2)} + \boxed{b_1} X_1 \quad \textbf{Holiday}$$

$$\hat{Y} = b_0 + b_1 X_1 + b_2(0) = \boxed{b_0} + \boxed{b_1} X_1 \quad \textbf{No Holiday}$$

Different intercept **Same slope**



If H_0: $\beta_2 = 0$ is rejected, then "Holiday" has a significant effect on pie sales

 © 2006 Prentice Hall, Inc.

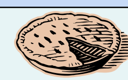
Interpreting the Dummy Variable Coefficient (with 2 Levels)

Example: $\widehat{Sales} = 300 - 30(Price) + 15(Holiday)$

Sales: number of pies sold per week
Price: pie price in $

$$\text{Holiday:} \begin{cases} 1 & \text{If a holiday occurred during the week} \\ 0 & \text{If no holiday occurred} \end{cases}$$

b_2 = 15: on average, sales were 15 pies greater in weeks with a holiday than in weeks without a holiday, given the same price



Basic Business Statistics, 11e © 2009 Prentice-Hall, Inc.. Chap 14-49

Dummy-Variable Models (more than 2 Levels)

- The number of dummy variables is **one less than the number of levels**
- Example:

  Y = house price ; X_1 = square feet

- If style of the house is also thought to matter:

  Style = $\underbrace{\text{ranch, split level, colonial}}$

Three levels, so two dummy variables are needed

Basic Business Statistics, 11e © 2009 Prentice-Hall, Inc.. Chap 14-50

Dummy-Variable Models (more than 2 Levels)

(continued)

- Example: Let "colonial" be the default category, and let X_2 and X_3 be used for the other two categories:

Y = house price

X_1 = square feet

X_2 = 1 if ranch, 0 otherwise

X_3 = 1 if split level, 0 otherwise

The multiple regression equation is:

$$\hat{Y} = b_0 + b_1X_1 + b_2X_2 + b_3X_3$$

Interpreting the Dummy Variable Coefficients (with 3 Levels)

Consider the regression equation:

$$\boxed{\hat{Y} = 20.43 + 0.045X_1 + 23.53X_2 + 18.84X_3}$$

For a colonial: $X_2 = X_3 = 0$

$$\boxed{\hat{Y} = 20.43 + 0.045X_1}$$

For a ranch: $X_2 = 1$; $X_3 = 0$

$$\boxed{\hat{Y} = 20.43 + 0.045X_1 + 23.53}$$

With the same square feet, a ranch will have an estimated average price of 23.53 thousand dollars more than a colonial.

For a split level: $X_2 = 0$; $X_3 = 1$

$$\boxed{\hat{Y} = 20.43 + 0.045X_1 + 18.84}$$

With the same square feet, a split-level will have an estimated average price of 18.84 thousand dollars more than a colonial.

 © 2006 Prentice Hall, Inc.

Interaction Between Independent Variables

- Hypothesizes interaction between pairs of X variables
  - Response to one X variable may vary at different levels of another X variable
- Contains two-way cross product terms
  - $$\hat{Y} = b_0 + b_1X_1 + b_2X_2 + b_3X_3$$
    $$= b_0 + b_1X_1 + b_2X_2 + b_3(X_1X_2)$$

Basic Business Statistics, 11e © 2009 Prentice-Hall, Inc.. Chap 14-53

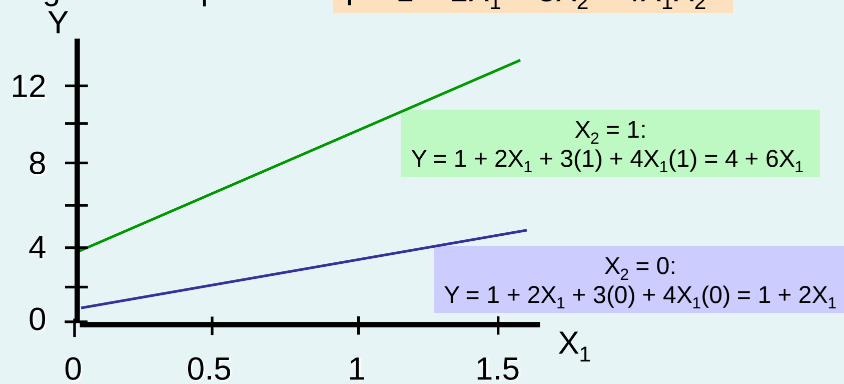
Effect of Interaction

- Given: $Y = \beta_0 + \beta_1X_1 + \beta_2X_2 + \beta_3X_1X_2 + \varepsilon$
- Without interaction term, effect of X_1 on Y is measured by β_1
- With interaction term, effect of X_1 on Y is measured by $\beta_1 + \beta_3 X_2$
- Effect changes as X_2 changes

Basic Business Statistics, 11e © 2009 Prentice-Hall, Inc.. Chap 14-54

Interaction Example

Suppose X_2 is a dummy variable and the estimated regression equation is $\hat{Y} = 1 + 2X_1 + 3X_2 + 4X_1X_2$



Slopes are different if the effect of X_1 on Y depends on X_2 value

Significance of Interaction Term

- Can perform a partial F test for the contribution of a variable to see if the addition of an interaction term improves the model
- Multiple interaction terms can be included
  - Use a partial F test for the simultaneous contribution of multiple variables to the model

Simultaneous Contribution of Independent Variables

- Use partial F test for the simultaneous contribution of multiple variables to the model
  - Let m variables be an additional set of variables added simultaneously
  - To test the hypothesis that the set of m variables improves the model:

$$F_{STAT} = \frac{[\mathrm{SSR(all)} - \mathrm{SSR\ (all\ except\ new\ set\ of\ m\ variables)}]/m}{\mathrm{MSE(all)}}$$

(where F_{STAT} has m and n-k-1 d.f.)

Basic Business Statistics, 11e © 2009 Prentice-Hall, Inc.. Chap 14-57

Logistic Regression

- Used when the dependent variable Y is binary (i.e., Y takes on only two values)
- Examples
  - Customer prefers Brand A or Brand B
  - Employee chooses to work full-time or part-time
  - Loan is delinquent or is not delinquent
  - Person voted in last election or did not
- Logistic regression allows you to predict the probability of a particular categorical response

Basic Business Statistics, 11e © 2009 Prentice-Hall, Inc.. Chap 14-58

Logistic Regression

(continued)

- Logistic regression is based on the **odds ratio**, which represents the probability of a success compared with the probability of failure

$$\text{Odds ratio} = \frac{\text{probability of success}}{1 - \text{probability of success}}$$

- The logistic regression model is based on the natural log of this odds ratio

Logistic Regression

(continued)

Logistic Regression Model:

$$\ln(\text{odds ratio}) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + \varepsilon_i$$

Where k = number of independent variables in the model

ε_i = random error in observation i

Logistic Regression Equation:

$$\ln(\text{estimated odds ratio}) = b_0 + b_1 X_{1i} + b_2 X_{2i} + \cdots + b_k X_{ki}$$

Estimated Odds Ratio and Probability of Success

- Once you have the logistic regression equation, compute the estimated odds ratio:

$$\text{Estimated odds ratio} = e^{\ln(\text{estimated odds ratio})}$$

- The estimated probability of success is

$$\text{Estimated probability of success} = \frac{\text{estimated odds ratio}}{1 + \text{estimated odds ratio}}$$

Chapter Summary

- Developed the multiple regression model
- Tested the significance of the multiple regression model
- Discussed adjusted r^2
- Discussed using residual plots to check model assumptions
- Tested individual regression coefficients
- Tested portions of the regression model
- Used dummy variables
- Evaluated interaction effects
- Discussed logistic regression

© 2006 Prentice Hall, Inc.