Chapter 14

Introduction to Multiple Regression

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 most important in predicting a dependent variable
- How to use categorical independent variables in a regression model
- How to predict a categorical dependent variable using logistic regression
- How to identify individual observations that may be unduly influencing the multiple regression model

The Multiple Regression Model

DCOVA

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

DCOVA

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

Multiple regression equation with k independent variables:

$$\underbrace{\hat{Y}_i}_{\text{Estimated (or predicted) value of Y}} = \underbrace{b_0}_{\text{Estimated intercept}} + b_1 X_{1i} + b_2 X_{2i} + \cdots + b_k X_{ki}$$

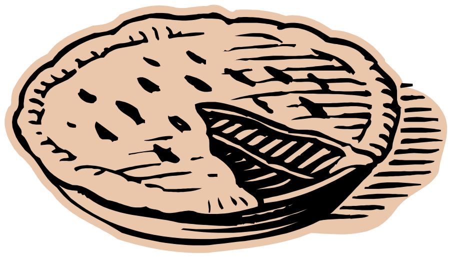
Estimated slope coefficients

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

Example:
2 Independent Variables

DCOVA

- 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



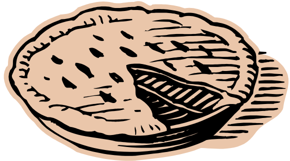
Pie Sales Example

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

DCOVA

Multiple regression equation:

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



Minitab Multiple Regression Output

DCOVA

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

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

DCOVA

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

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

Using The Equation to Make Predictions

DCOVA

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 $100s, so $350 means that $X_2 = 3.5$

Predictions in Minitab

DCOVA

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)

Predicted $\hat{Y}$ value

Values of Predictors for New Observations

New Obs	Price	Advertising
1	5.50	3.50

Input values

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

The Coefficient of Multiple Determination, r^2

DCOVA

- 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 Minitab

DCOVA

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

DCOVA

- 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?

Adjusted r^2

(continued)

DCOVA

- 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)

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

Adjusted r^2 in Minitab

DCOVA

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

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			

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?

DCOVA

- 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

DCOVA

- 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 Minitab

DCOVA

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

F Test for Overall Significance

(continued)

DCOVA

H_0: $\beta_1 = \beta_2 = 0$

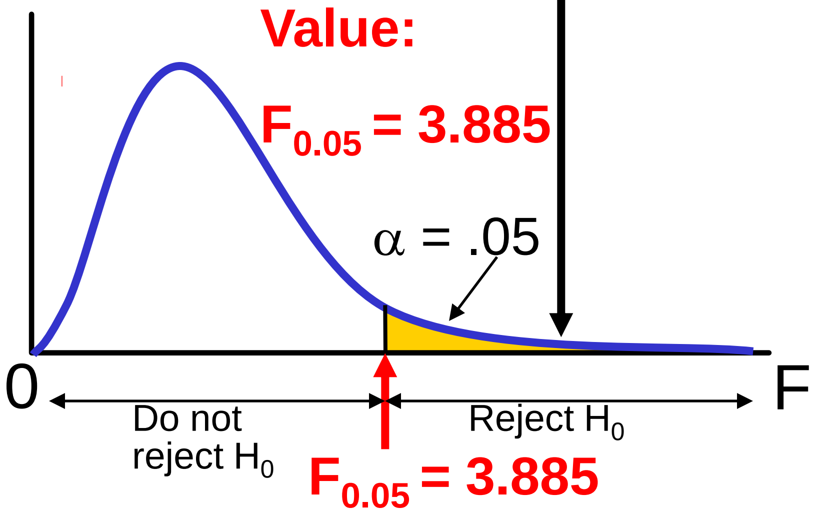
H_1: β_1 and β_2 not both zero

$\alpha = .05$

$df_1 = 2 \quad df_2 = 12$

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

Residuals in Multiple Regression

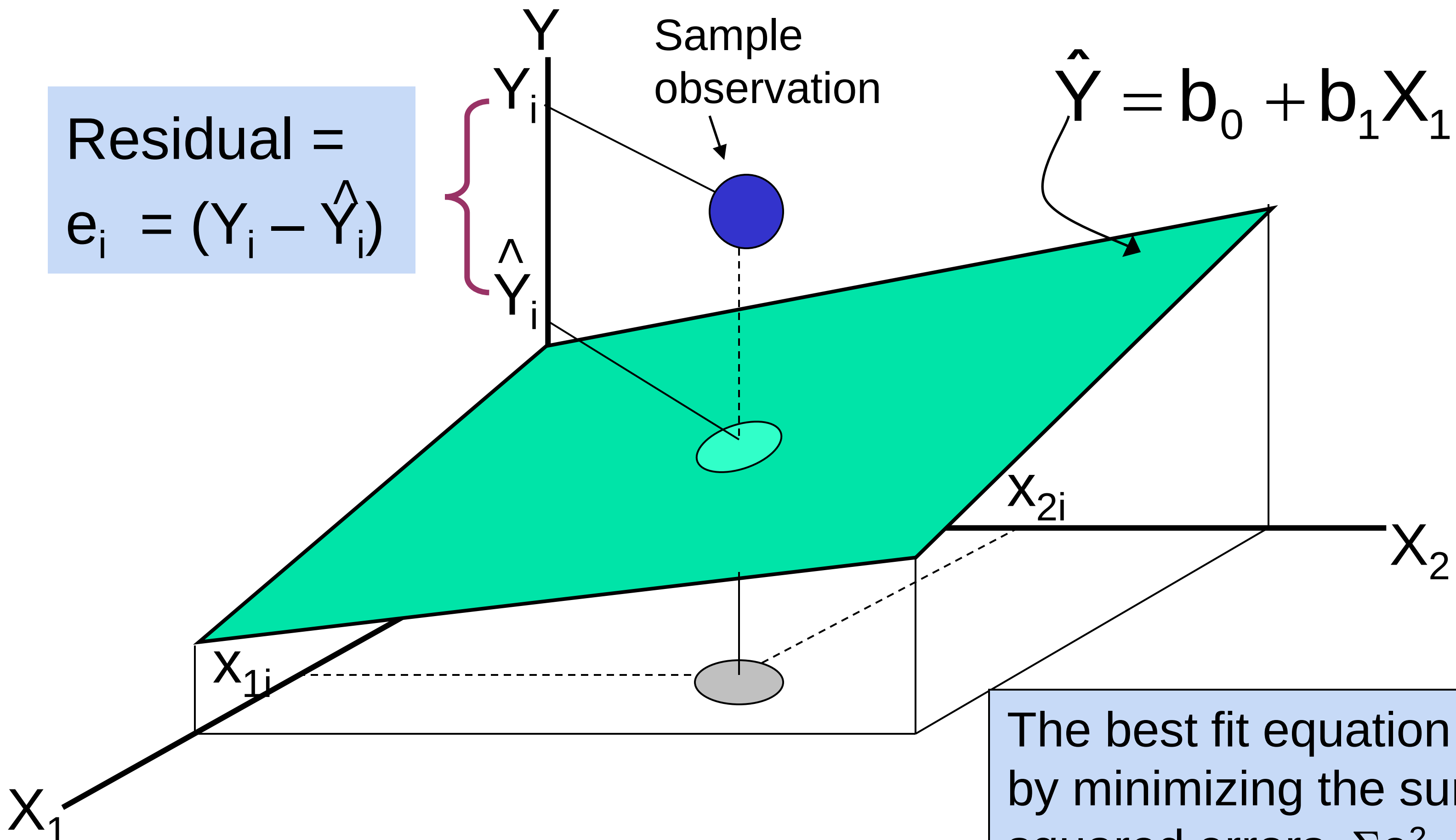
DCOVA

Two variable model

Residual =

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

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



The best fit equation is found by minimizing the sum of squared errors, Σe^2

Multiple Regression Assumptions

DCOVA

Errors (residuals) from the regression model:

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

Assumptions:

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

Residual Plots Used in Multiple Regression

DCOVA

- 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?

DCOVA

- 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)

Are Individual Variables Significant?

(continued)

DCOVA

H_0: $\beta_j = 0$ (no linear relationship between X_j and Y)

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

Test Statistic:

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

(df = n – k – 1)

Are Individual Variables Significant? Minitab Output

DCOVA

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} = -2.31, with p-value .040

t Stat for Advertising is t_{STAT} = 2.85, with p-value .014

Inferences about the Slope: t Test Example

DCOVA

$H_0: \beta_j = 0$

$H_1: \beta_j \neq 0$

d.f. = 15-2-1 = 12

α = .05

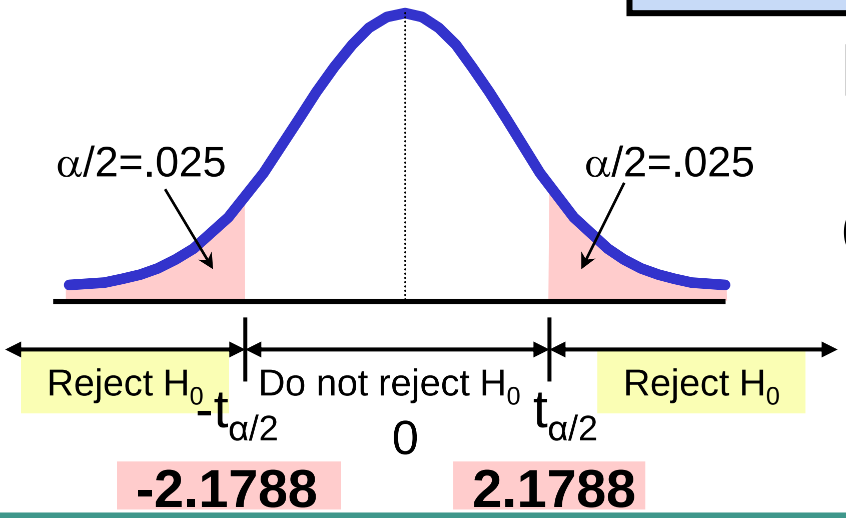
$t_{\alpha/2}$ = 2.1788

From the Excel 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

Confidence Interval Estimate for the Slope

DCOVA

Confidence interval for the population slope β_j

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

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:

$$-24.975 \pm (2.1788)(10.832)$$

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

DCOVA

(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 advertising constant

Testing Portions of the Multiple Regression Model

DCOVA

- Contribution of a Single Independent Variable X_j

$$SSR(X_j \mid \text{all variables except } X_j)$$
$$= SSR\ (\text{all variables}) - SSR(\text{all variables except } X_j)$$

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

Testing Portions of the Multiple Regression Model

(continued)

DCOVA

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

$$SSR(X_1 \mid X_2) = SSR\text{ (all variables)} - SSR(X_2)$$

From ANOVA section of regression for

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

From ANOVA section of regression for

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

Measures the contribution of X_1 in explaining SST

The Partial F-Test Statistic

DCOVA

- 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.)

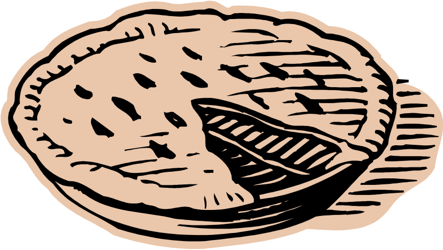
$$F_{STAT} = \frac{\text{SSR }(X_j \mid \text{all variables except j})}{\text{MSE}}$$

Testing Portions of Model: Example

DCOVA

Example: Frozen dessert pies

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



Testing Portions of Model: Example

(continued)

DCOVA

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

Testing Portions of Model: Example

(continued)

DCOVA

(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

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

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

Testing Portions of Model: Example

(continued)

DCOVA

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

H_1: X_2 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_1 only)

ANOVA		
	df	*SS*
Regression	1	**11100.43803**
Residual	13	45392.8953
Total	14	56493.33333

Testing Portions of Model: Example

(continued)

DCOVA

(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_1 only)

ANOVA		
	df	*SS*
Regression	1	**11100.43803**
Residual	13	45392.8953
Total	14	56493.33333

$$F_{STAT} = \frac{\text{SSR}(X_2 \mid X_1)}{\text{MSE(all)}} = \frac{29{,}460.03 - 11{,}100.44}{2252.78} = 8.150$$

Conclusion: Since $F_{STAT} = 8.150 > F_{0.05} = 4.75$ Reject H_0;
Adding X_2 does improve model

Coefficient of Partial Determination for k variable model

DCOVA

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

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

Coefficient of Partial Determination in Excel

DCOVA

- 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

DCOVA

- 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)

DCOVA

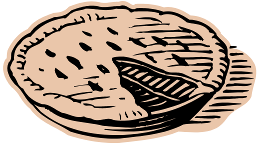
$$\hat{Y} = b_0 + b_1 X_1 + b_2 X_2$$

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)

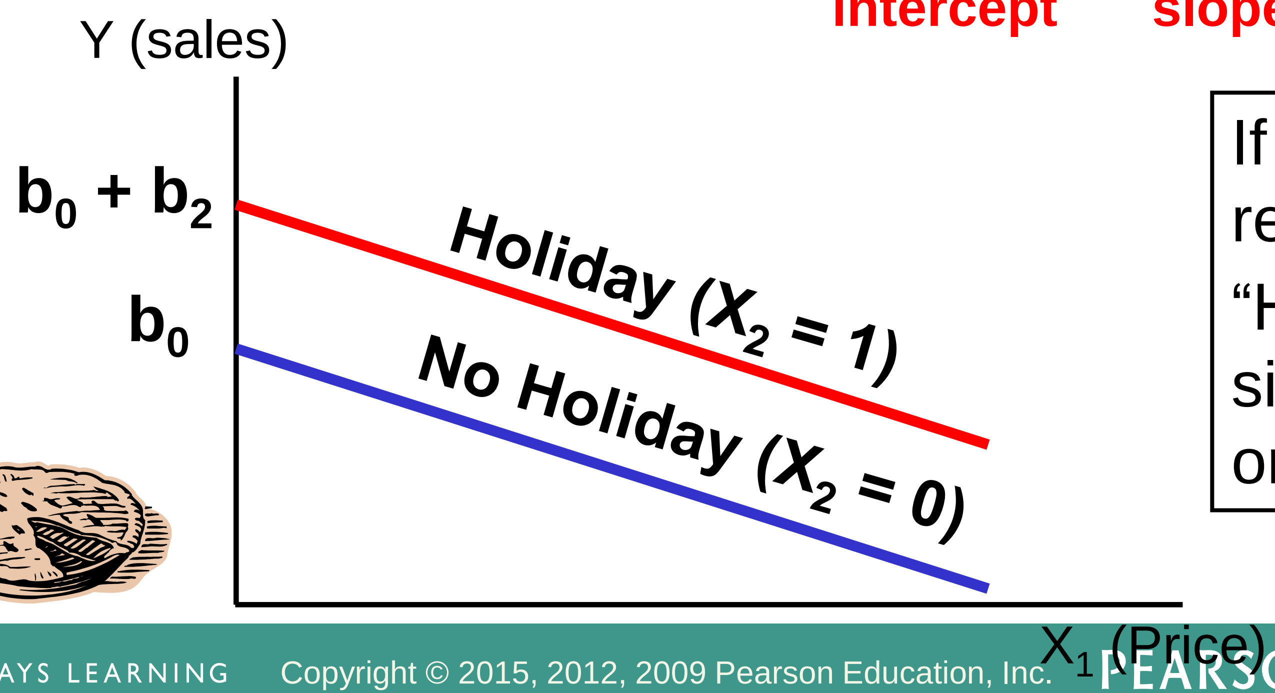
(continued)

DCOVA

$$\hat{Y} = b_0 + b_1X_1 + b_2(1) = (b_0 + b_2) + b_1X_1 \quad \textbf{Holiday}$$

$$\hat{Y} = b_0 + b_1X_1 + b_2(0) = \quad b_0 \quad + b_1X_1 \quad \textbf{No Holiday}$$

Different intercept **Same slope**



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

Interpreting the Dummy Variable Coefficient (with 2 Levels)

DCOVA

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

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

Dummy-Variable Models (more than 2 Levels)

DCOVA

- 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

Dummy-Variable Models (more than 2 Levels)

(continued)

DCOVA

- 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:

$$\hat{Y} = b_0 + b_1 X_1 + b_2 X_2 + b_3 X_3$$

Interpreting the Dummy Variable Coefficients (with 3 Levels)

DCOVA

Consider the regression equation:

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

For a colonial: $X_2 = X_3 = 0$

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

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

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

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: $X_2 = 0$; $X_3 = 1$

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

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

Interaction Between Independent Variables

DCOVA

- 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
  - $$\hat{Y} = b_0 + b_1X_1 + b_2X_2 + b_3X_3$$
    $$= b_0 + b_1X_1 + b_2X_2 + b_3(X_1X_2)$$

Effect of Interaction

DCOVA

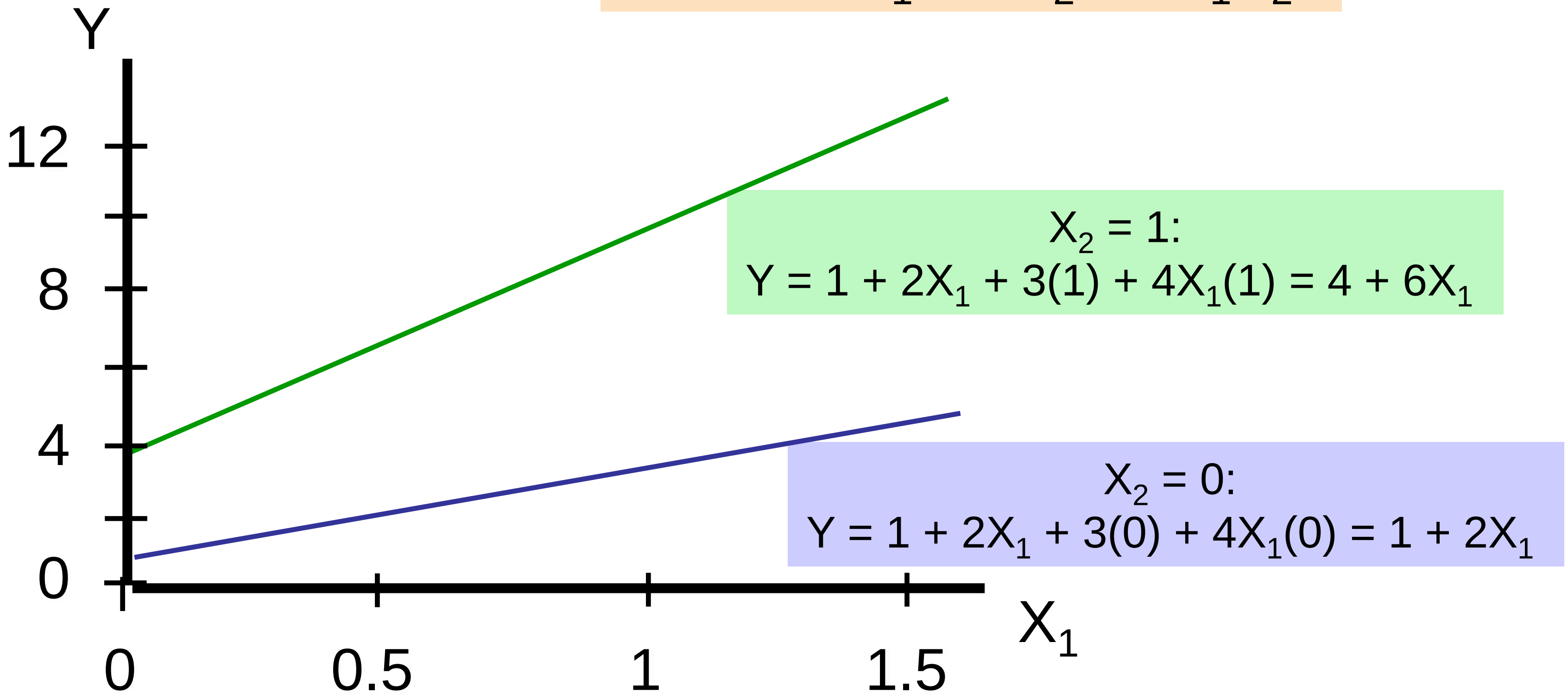
- Given: $$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \varepsilon$$

- 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

Interaction Example

DCOVA

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

DCOVA

- 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

DCOVA

- 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:

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

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

Chapter Summary

In this chapter we discussed

- The multiple regression model
- Testing the significance of the multiple regression model
- Adjusted r^2
- Using residual plots to check model assumptions
- Testing individual regression coefficients
- Testing portions of the regression model
- Using dummy variables
- Evaluating interaction effects