# Business Statistics: A Decision-Making Approach

6th Edition

## Chapter 14

Multiple Regression Analysis and Model Building

 Chap 14-1

## Chapter Goals

**After completing this chapter, you should be able to:**

- understand model building using multiple regression analysis
- apply multiple regression analysis to business decision-making situations
- analyze and interpret the computer output for a multiple regression model
- test the significance of the independent variables in a multiple regression model

 Chap 14-2

## Chapter Goals

*(continued)*

**After completing this chapter, you should be able to:**

- use variable transformations to model nonlinear relationships
- recognize potential problems in multiple regression analysis and take the steps to correct the problems.
- incorporate qualitative variables into the regression model by using dummy variables.

 Chap 14-3

## The Multiple Regression Model

Idea: Examine the linear relationship between 1 dependent (y) & 2 or more independent variables ($x_i$)

**Population model:**

Y-intercept / Population slopes / Random Error

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

**Estimated multiple regression model:**

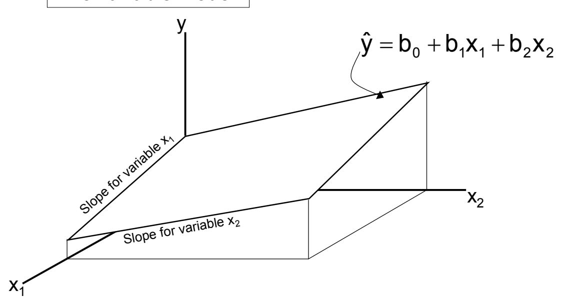
Estimated (or predicted) value of y / Estimated intercept / Estimated slope coefficients

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

Business Statistics: A Decision-Making Approach, 6e © 2005 Prentice-Hall, Inc. Chap 14-4

## Multiple Regression Model

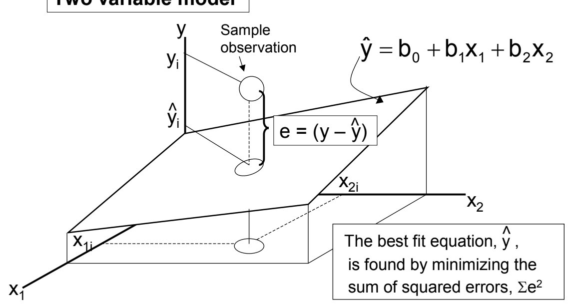
**Two variable model**




Business Statistics: A Decision-Making Approach, 6e © 2005 Prentice-Hall, Inc. Chap 14-5

## Multiple Regression Model

**Two variable model**




The best fit equation, $\hat{y}$, is found by minimizing the sum of squared errors, $\Sigma e^2$

Business Statistics: A Decision-Making Approach, 6e © 2005 Prentice-Hall, Inc. Chap 14-6

## Multiple Regression Assumptions

**Errors (residuals) from the regression model:**

$$e = (y - \hat{y})$$

- The errors are normally distributed
- The mean of the errors is zero
- Errors have a constant variance
- The model errors are independent

## Model Specification

- Decide what you want to do and select the dependent variable
- Determine the potential independent variables for your model
- Gather sample data (observations) for all variables

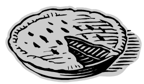
## The Correlation Matrix

- Correlation between the dependent variable and selected independent variables can be found using Excel:
  - Tools / Data Analysis… / Correlation
- Can check for statistical significance of correlation with a t test

## Example

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





## Pie Sales Model

| •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 model:

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

Correlation matrix:

| • | •Pie Sales | •Price | •Advertising |
|---|---|---|---|
| •Pie Sales | •1 | | |
| •Price | •-0.44327 | •1 | |
| •Advertising | •0.55632 | •0.03044 | •1 |



## Interpretation of Estimated Coefficients

- Slope ($b_i$)
  - Estimates that the average value of y changes by $b_i$ units for each 1 unit increase in $X_i$ holding all other variables constant
  - Example: if $b_1 = -20$, then sales (y) is expected to decrease by an estimated 20 pies per week for each $1 increase in selling price ($x_1$), net of the effects of changes due to advertising ($x_2$)
- y-intercept ($b_0$)
  - The estimated average value of y when all $x_i = 0$ (assuming all $x_i = 0$ is within the range of observed values)

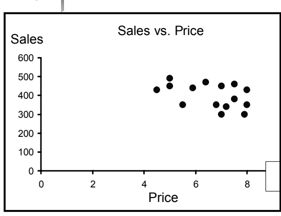
## Pie Sales Correlation Matrix

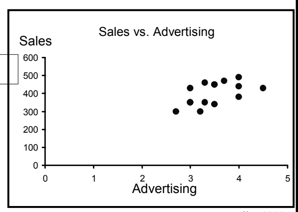
| • | •Pie Sales | •Price | •Advertising |
|---|---|---|---|
| •Pie Sales | •1 | | |
| •Price | $\boxed{\text{•-0.44327}}$ | •1 | |
| •Advertising | $\boxed{\text{•0.55632}}$ | •0.03044 | •1 |

- Price vs. Sales : r = -0.44327
  - There is a negative association between price and sales
- Advertising vs. Sales : r = 0.55632
  - There is a positive association between advertising and sales




## Scatter Diagrams







## Estimating a Multiple Linear Regression Equation

- Computer software is generally used to generate the coefficients and measures of goodness of fit for multiple regression
- Excel:
  - Tools / Data Analysis... / Regression
- PHStat:
  - PHStat / Regression / Multiple Regression…



## Multiple Regression Output

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

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

| 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 |



## 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



## Using The Model to Make Predictions

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

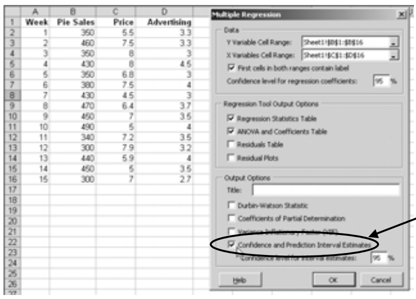
$$\begin{aligned}\widehat{\text{Sales}} &= 306.526 - 24.975(\text{Price}) + 74.131(\text{Advertising})\\ &= 306.526 - 24.975(5.50) + 74.131(3.5)\\ &= 428.62\end{aligned}$$

Predicted sales is 428.62 pies

Note that Advertising is in $100's, so $350 means that $x_2$ = 3.5

## Predictions in PHStat

- PHStat | regression | multiple regression …



Check the "confidence and prediction interval estimates" box



## 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 $\hat{y}$ value, given these x's

Prediction interval for an individual $\hat{y}$ value, given these x's




## Multiple Coefficient of Determination

- Reports the proportion of total variation in y explained by all x variables taken together

$$R^2 = \frac{SSR}{SST} = \frac{\text{Sum of squares regression}}{\text{Total sum of squares}}$$



## Multiple Coefficient of Determination

*(continued)*

| •Regression Statistics | |
|---|---|
| •Multiple R | •0.72213 |
| •R Square | •0.52148 |
| 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 |

Business Statistics: A Decision-Making Approach, 6e © 2005 Prentice-Hall, Inc. Chap 14-22

## 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?

Business Statistics: A Decision-Making Approach, 6e © 2005 Prentice-Hall, Inc. Chap 14-23

## Adjusted $R^2$

*(continued)*

- Shows the proportion of variation in y explained by all x variables adjusted for the number of x variables used

$$R_A^2 = 1 - (1 - R^2)\left(\frac{n-1}{n-k-1}\right)$$

(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

Business Statistics: A Decision-Making Approach, 6e © 2005 Prentice-Hall, Inc. Chap 14-24

## Multiple Coefficient of Determination

*(continued)*

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

$$R_A^2 = .44172$$

**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 |



## 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:
  - $H_0$: $\beta_1 = \beta_2 = \ldots = \beta_k = 0$ (no linear relationship)
  - $H_A$: at least one $\beta_i \neq 0$ (at least one independent variable affects y)



## F-Test for Overall Significance

*(continued)*

- Test statistic:

$$\boxed{F = \frac{\frac{SSR}{k}}{\frac{SSE}{n-k-1}} = \frac{MSR}{MSE}}$$

where F has (numerator) $D_1 = k$ and (denominator) $D_2 = (n - k - 1)$ degrees of freedom



## F-Test for Overall Significance

*(continued)*

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

$$F = \frac{MSR}{MSE} = \frac{14730.0}{2252.8} = 6.5386$$

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 |



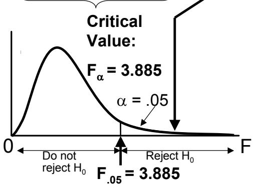
## F-Test for Overall Significance

*(continued)*

$H_0$: $\beta_1 = \beta_2 = 0$
$H_A$: $\beta_1$ and $\beta_2$ not both zero

$\alpha = .05$

$\underbrace{df_1 = 2 \quad df_2 = 12}$




**Test Statistic:**

$$F = \frac{MSR}{MSE} = 6.5386$$

**Decision:**

Reject $H_0$ at $\alpha = 0.05$

**Conclusion:**

**The regression model does explain a significant portion of the variation in pie sales**

**(There is evidence that at least one independent variable affects y)**



## Are Individual Variables Significant?

- Use t-tests of individual variable slopes
- Shows if there is a linear relationship between the variable $x_i$ and y
- Hypotheses:
  - $H_0$: $\beta_i = 0$ (no linear relationship)
  - $H_A$: $\beta_i \neq 0$ (linear relationship does exist between $x_i$ and y)

## Are Individual Variables Significant?

*(continued)*

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

$H_A$: $\beta_i \neq 0$ (linear relationship does exist between $x_i$ and y)

Test Statistic:

$$t = \frac{b_i - 0}{s_{b_i}}$$

(df = n – k – 1)



## Are Individual Variables Significant?

*(continued)*

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

**t-value for Price is t = -2.306, with p-value .0398**

**t-value for Advertising is t = 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 |



## Inferences about the Slope: t Test Example

$H_0$: $\beta_i = 0$

$H_A$: $\beta_i \neq 0$

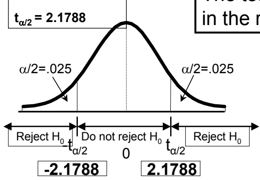
**From Excel output:**

| • | •Coefficients | •Standard Error | •t Stat | •P-value |
|---|---|---|---|---|
| •Price | •-24.97509 | •10.83213 | •-2.30565 | •0.03979 |
| •Advertising | •74.13096 | •25.96732 | •2.85478 | •0.01449 |

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

$\alpha$ **= .05**

$t_{\alpha/2}$ **= 2.1788**




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 $\alpha$ = .05

## Confidence Interval Estimate for the Slope

Confidence interval for the population slope $\beta_1$ (the effect of changes in price on pie sales):

$$b_i \pm t_{\alpha/2} s_{b_i}$$

where t has (n – k – 1) d.f.

| • | •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: Weekly sales are estimated to be reduced by between 1.37 to 48.58 pies for each increase of $1 in the selling price

 Chap 14-34

## Standard Deviation of the Regression Model

- The estimate of the standard deviation of the regression model is:

$$s_\varepsilon = \sqrt{\frac{SSE}{n-k-1}} = \sqrt{MSE}$$

- Is this value large or small? Must compare to the mean size of y for comparison

 Chap 14-35

## Standard Deviation of the Regression Model

*(continued)*

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

**The standard deviation of the regression model is 47.46**

| •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 |

 Chap 14-36

## Standard Deviation of the Regression Model

*(continued)*

- The standard deviation of the regression model is 47.46
- A rough prediction range for pie sales in a given week is $\pm 2(47.46) = 94.2$
- Pie sales in the sample were in the 300 to 500 per week range, so this range is probably too large to be acceptable. The analyst may want to look for additional variables that can explain more of the variation in weekly sales



## Multicollinearity

- Multicollinearity: High correlation exists between two independent variables
- This means the two variables contribute redundant information to the multiple regression model



## Multicollinearity

*(continued)*

- Including two highly correlated independent variables can adversely affect the regression results
  - No new information provided
  - Can lead to unstable coefficients (large standard error and low t-values)
  - Coefficient signs may not match prior expectations

## Some Indications of Severe Multicollinearity

- Incorrect signs on the coefficients
- Large change in the value of a previous coefficient when a new variable is added to the model
- A previously significant variable becomes insignificant when a new independent variable is added
- The estimate of the standard deviation of the model increases when a variable is added to the model

 Chap 14-40

## Detect Collinearity (Variance Inflationary Factor)

$VIF_j$ is used to measure collinearity:

$$VIF_j = \frac{1}{1 - R_j^2}$$

$R^2_j$ is the coefficient of determination when the $j^{th}$ independent variable is regressed against the remaining $k - 1$ independent variables

If $VIF_j > 5$, $x_j$ is highly correlated with the other explanatory variables

 Chap 14-41

## Detect Collinearity in PHStat

PHStat / regression / multiple regression …
Check the "variance inflationary factor (VIF)" box

| •Regression Analysis | |
|---|---|
| •Price and all other X | |
| *•Regression Statistics* | |
| •Multiple R | •0.030437581 |
| •R Square | •0.000926446 |
| •Adjusted R Square | •-0.075925366 |
| •Standard Error | •1.21527235 |
| •Observations | •15 |
| **•VIF** | **•1.000927305** |

Output for the pie sales example:
- Since there are only two explanatory variables, only one VIF is reported
  - VIF is < 5
  - There is no evidence of collinearity between Price and Advertising

 Chap 14-42

## Qualitative (Dummy) Variables

- Categorical explanatory variable (dummy variable) with two or more levels:
  - yes or no, on or off, male or female
  - coded as 0 or 1
- Regression intercepts are different if the variable is significant
- Assumes equal slopes for other variables
- The number of dummy variables needed is (number of levels - 1)

 Chap 14-43

## Dummy-Variable Model Example (with 2 Levels)

Let:

y = pie sales

$$\hat{y} = b_0 + b_1x_1 + b_2x_2$$

$x_1$ = price

$x_2$ = holiday ($X_2$ = 1 if a holiday occurred during the week)
($X_2$ = 0 if there was no holiday that week)

 Chap 14-44

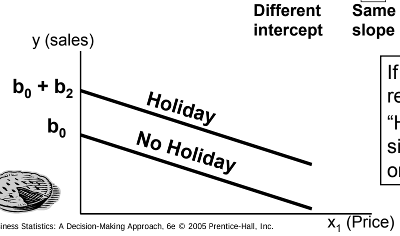
## Dummy-Variable Model Example (with 2 Levels)

*(continued)*

$$\hat{y} = b_0 + b_1x_1 + b_2(1) = \boxed{(b_0 + b_2)} + \boxed{b_1}x_1 \quad \text{Holiday}$$

$$\hat{y} = b_0 + b_1x_1 + b_2(0) = \boxed{b_0} + \boxed{b_1}x_1 \quad \text{No Holiday}$$

**Different intercept** **Same slope**




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

 Chap 14-45

## Interpretation of the Dummy Variable Coefficient (with 2 Levels)

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

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)

- The number of dummy variables is **one less than the number of levels**
- Example:
  y = house price ; $x_1$ = square feet
- The style of the house is also thought to matter:
  Style = ranch, split level, condo

Three levels, so two dummy variables are needed



## Dummy-Variable Models (more than 2 Levels)

*(continued)*

**Let the default category be "condo"**

$$x_2 = \begin{cases} 1 & \text{if ranch} \\ 0 & \text{if not} \end{cases} \qquad x_3 = \begin{cases} 1 & \text{if split level} \\ 0 & \text{if not} \end{cases}$$

$$\hat{y} = b_0 + b_1x_1 + b_2x_2 + b_3x_3$$

$b_2$ shows the impact on price if the house is a ranch style, compared to a condo

$b_3$ shows the impact on price if the house is a split level style, compared to a condo



## Interpreting the Dummy Variable Coefficients (with 3 Levels)

Suppose the estimated equation is

$$\hat{y} = 20.43 + 0.045x_1 + 23.53x_2 + 18.84x_3$$

For a condo: $x_2 = x_3 = 0$

$$\hat{y} = 20.43 + 0.045x_1$$

For a ranch: $x_3 = 0$

$$\hat{y} = 20.43 + 0.045x_1 + 23.53$$

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

For a split level: $x_2 = 0$

$$\hat{y} = 20.43 + 0.045x_1 + 18.84$$

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



## Nonlinear Relationships

- The relationship between the dependent variable and an independent variable may not be linear
- Useful when scatter diagram indicates non-linear relationship
- Example: Quadratic model
  - $y = \beta_0 + \beta_1 x_j + \beta_2 x_j^2 + \varepsilon$
  - The second independent variable is the square of the first variable



## Polynomial Regression Model

General form:

$$y = \beta_0 + \beta_1 x_j + \beta_2 x_j^2 + \ldots + \beta_p x_j^p + \varepsilon$$

- where:

  $\beta_0$ = Population regression constant
  $\beta_i$ = Population regression coefficient for variable $x_j$ : j = 1, 2, …*k*
  p = Order of the polynomial
  $\varepsilon_i$ = Model error

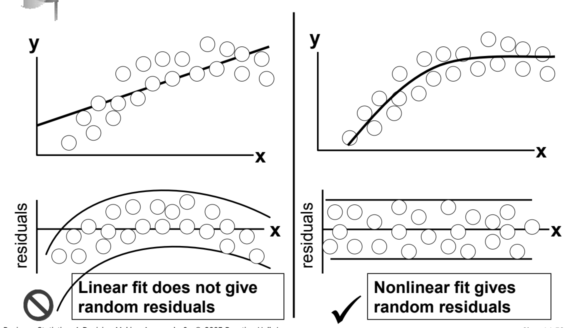
If p = 2 the model is a quadratic model:

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



## Linear vs. Nonlinear Fit

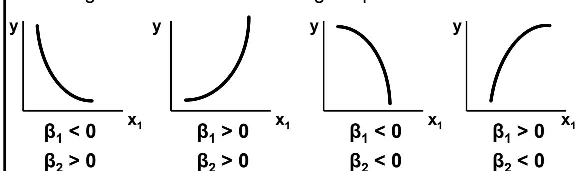



Business Statistics: A Decision-Making Approach, 6e © 2005 Prentice-Hall, Inc. Chap 14-52

## Quadratic Regression Model

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

Quadratic models may be considered when scatter diagram takes on the following shapes:




$\beta_1$ = the coefficient of the linear term
$\beta_2$ = the coefficient of the squared term

Business Statistics: A Decision-Making Approach, 6e © 2005 Prentice-Hall, Inc. Chap 14-53

## Testing for Significance: Quadratic Model

- Test for Overall Relationship
  - F test statistic = $\frac{MSR}{MSE}$
- Testing the Quadratic Effect
  - Compare quadratic model

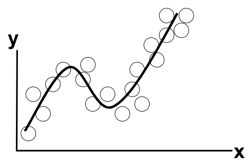
    $$y = \beta_0 + \beta_1 x_j + \beta_2 x_j^2 + \varepsilon$$

    with the linear model

    $$y = \beta_0 + \beta_1 x_j + \varepsilon$$
  - Hypotheses
    - $H_0$: $\beta_2 = 0$ (No 2nd order polynomial term)
    - $H_A$: $\beta_2 \neq 0$ (2nd order polynomial term is needed)

Business Statistics: A Decision-Making Approach, 6e © 2005 Prentice-Hall, Inc. Chap 14-54

## Higher Order Models




If p = 3 the model is a cubic form:

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

 Chap 14-55

## Interaction Effects

- Hypothesizes interaction between pairs of x variables
  - Response to one x variable varies at different levels of another x variable
- Contains two-way cross product terms

$$y = \beta_0 + \underbrace{\beta_1 x_1 + \beta_2 x_1^2 + \beta_3 x_3}_{\text{Basic Terms}} + \underbrace{\beta_4 x_1 x_2 + \beta_5 x_1^2 x_2}_{\text{Interactive Terms}}$$

 Chap 14-56

## Effect of Interaction

- Given:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \varepsilon$$

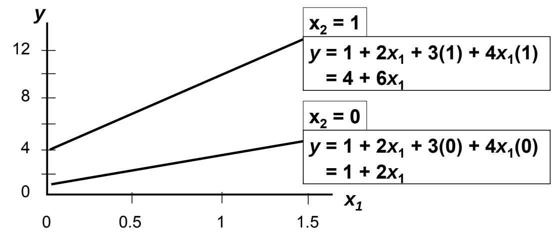
- Without interaction term, effect of $x_1$ on y is measured by $\beta_1$
- With interaction term, effect of $x_1$ on y is measured by $\beta_1 + \beta_3 x_2$
- Effect changes as $x_2$ increases

 Chap 14-57

## Interaction Example

$y = 1 + 2x_1 + 3x_2 + 4x_1x_2$ where $x_2$ = 0 or 1 (dummy variable)




**Effect (slope) of $x_1$ on y does depend on $x_2$ value**



## Interaction Regression Model Worksheet

| Case, i | $y_i$ | $x_{1i}$ | $x_{2i}$ | $x_{1i}\,x_{2i}$ |
|---|---|---|---|---|
| 1 | 1 | 1 | 3 | 3 |
| 2 | 4 | 8 | 5 | 40 |
| 3 | 1 | 3 | 2 | 6 |
| 4 | 3 | 5 | 6 | 30 |
| : | : | : | : | : |

multiply $x_1$ by $x_2$ to get $x_1x_2$, then run regression with y, $x_1$, $x_2$ , $x_1x_2$



## Evaluating Presence of Interaction

- Hypothesize interaction between pairs of independent variables

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \varepsilon$$

- Hypotheses:
  - $H_0$: $\beta_3 = 0$ (no interaction between $x_1$ and $x_2$)
  - $H_A$: $\beta_3 \neq 0$ ($x_1$ interacts with $x_2$)

## Model Building

- Goal is to develop a model with the best set of independent variables
  - Easier to interpret if unimportant variables are removed
  - Lower probability of collinearity
- Stepwise regression procedure
  - Provide evaluation of alternative models as variables are added
- Best-subset approach
  - Try all combinations and select the best using the highest adjusted $R^2$ and lowest $s_\varepsilon$

## Stepwise Regression

- Idea: develop the least squares regression equation in steps, either through forward selection, backward elimination, or through standard stepwise regression
- The coefficient of partial determination is the measure of the marginal contribution of each independent variable, given that other independent variables are in the model

## Best Subsets Regression

- Idea: estimate all possible regression equations using all possible combinations of independent variables
- Choose the best fit by looking for the highest adjusted $R^2$ and lowest standard error $s_\varepsilon$

Stepwise regression and best subsets regression can be performed using PHStat, Minitab, or other statistical software packages

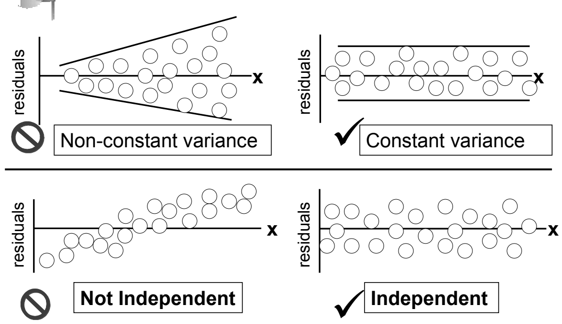
## Aptness of the Model

- Diagnostic checks on the model include verifying the assumptions of multiple regression:
  - Each $x_i$ is linearly related to y
  - Errors have constant variance
  - Errors are independent
  - Error are normally distributed

Errors (or Residuals) are given by $\boxed{e_i = (y - \hat{y})}$



## Residual Analysis






## The Normality Assumption

- Errors are assumed to be normally distributed
- Standardized residuals can be calculated by computer
- Examine a histogram or a normal probability plot of the standardized residuals to check for normality

## Chapter Summary

- Developed the multiple regression model
- Tested the significance of the multiple regression model
- Developed adjusted $R^2$
- Tested individual regression coefficients
- Used dummy variables
- Examined interaction in a multiple regression model

 Chap 14-67

## Chapter Summary

*(continued)*

- Described nonlinear regression models
- Described multicollinearity
- Discussed model building
  - Stepwise regression
  - Best subsets regression
- Examined residual plots to check model assumptions

 Chap 14-68