

Inference and Regression

Conditions Necessary for Valid Inference in the Least Squares Regression Model Testing Individual Regression Parameters

Addressing Nonsignificant Independent Variables Multicollinearity



Inference and Regression

* **Statistical inference:**
  + Process of making estimates and drawing conclusions about one or more characteristics of a population (parameter) through the analysis of sample data drawn from the population.
* In regression, inference is commonly used to estimate and draw conclusions about:

The regression parameters

** , ** , ** ,, ** .

0 1 2

*q*

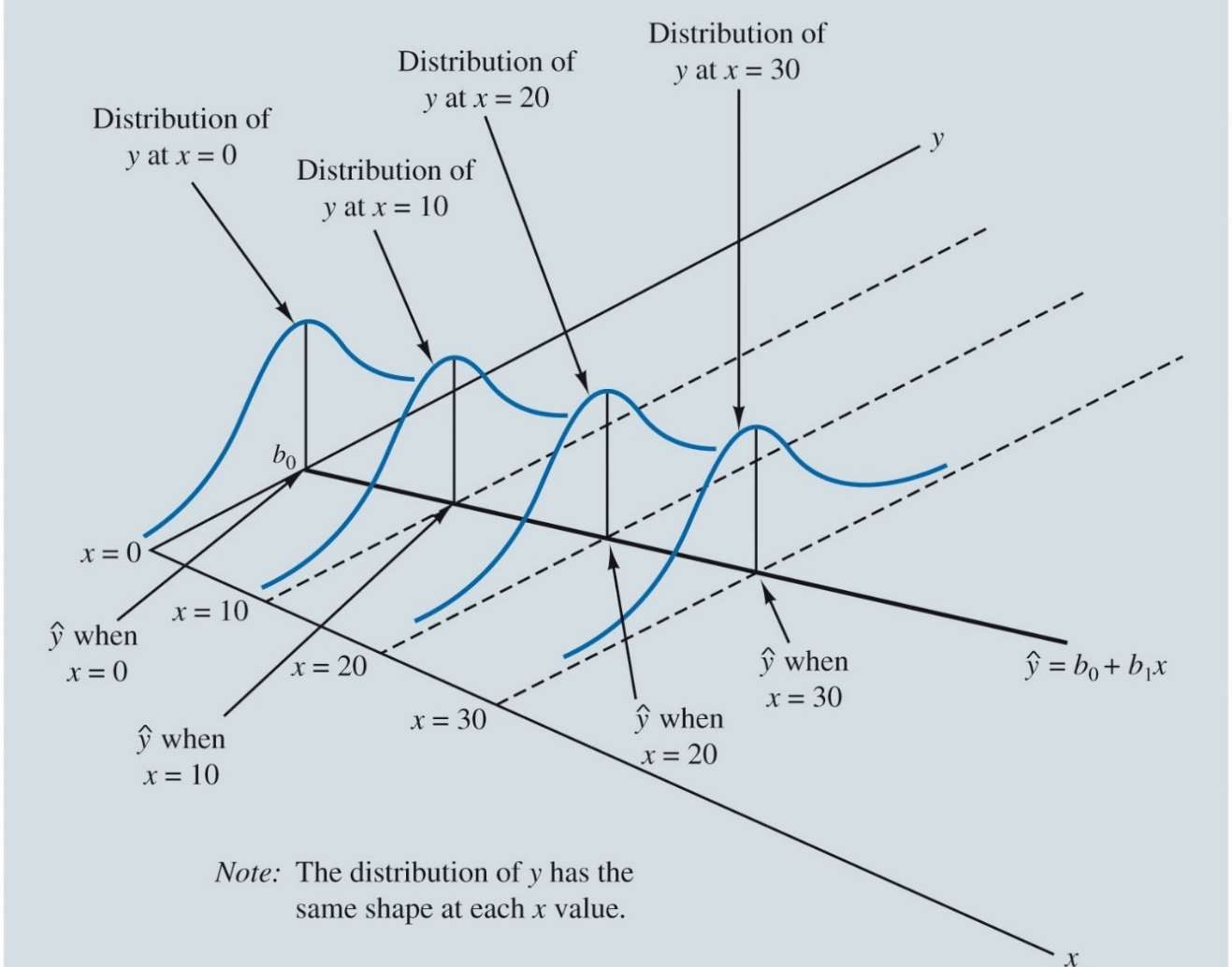
The mean value and/or the predicted value of the dependent variable *y* for specific values of the independent variables

*x*\* , *x*\* ,, *x*\*.

1 2

*q*

* Consider both **hypothesis testing** and **interval estimation**.

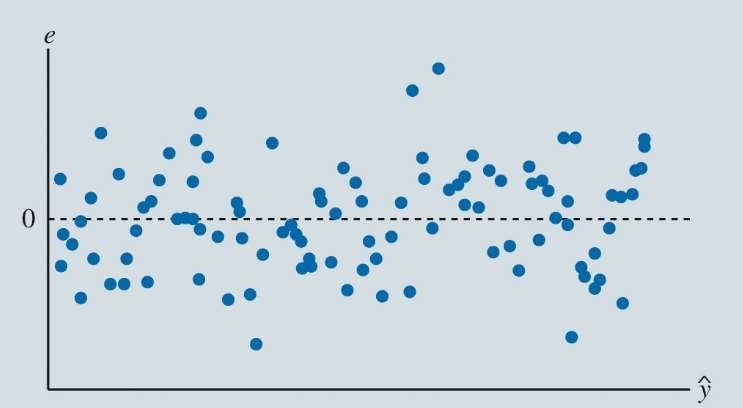


Inference and Regression

**Conditions Necessary for Valid Inference in** Illustration of the Conditions for Valid Inference in

**the Least Squares Regression Model:** Regression

* 1.For any given combination of values of the independent variables
  + 𝑥1, 𝑥2, … , 𝑥q, the population of potential error terms 𝜀 is normally distributed with a mean of 0 and a constant variance.
* 2. The values of 𝜀 are statistically independent



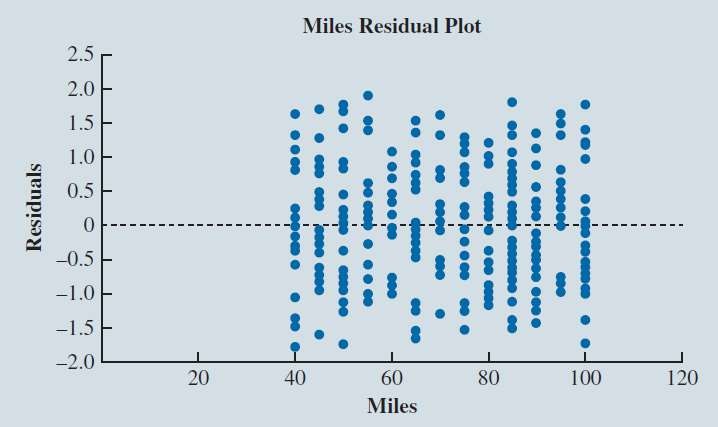
Inference and Regression

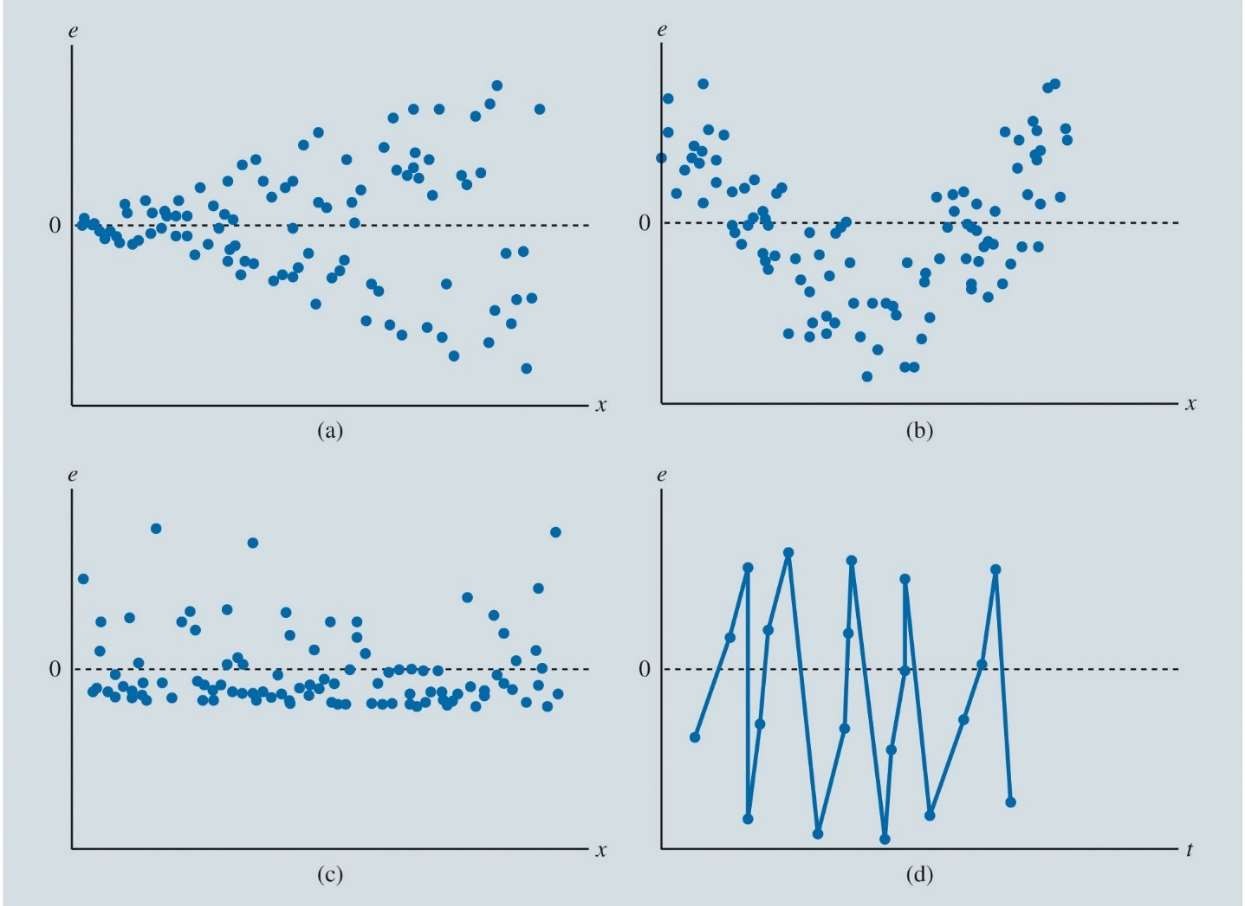
**Are the conditions violated?**

Example of a Random Error Pattern in a Scatter Chart of Residuals and Predicted Values of the Dependent Variable

* 1.Center of the residuals should be approximately 0.
  + Mean 0
* 2. The spread in data should be about the same through out
  + Constant variance
* 3. Errors should be symmetrically distributed with values near 0 occurring more frequently
  + Normally Distributed
* 4. Independent
  + Current data points do not depend on previous points

These residuals look good! – No violations





Inference and Regression

Examples of Diagnostic Scatter Charts of Residuals from Four Regressions

**Are the conditions violated?**

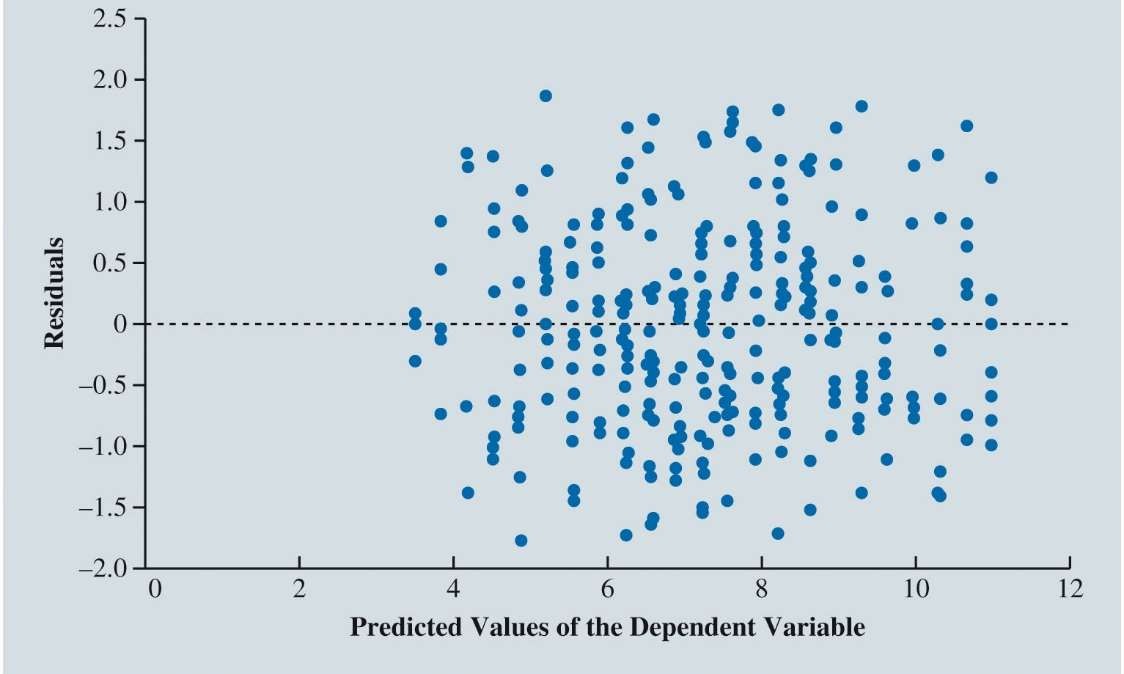
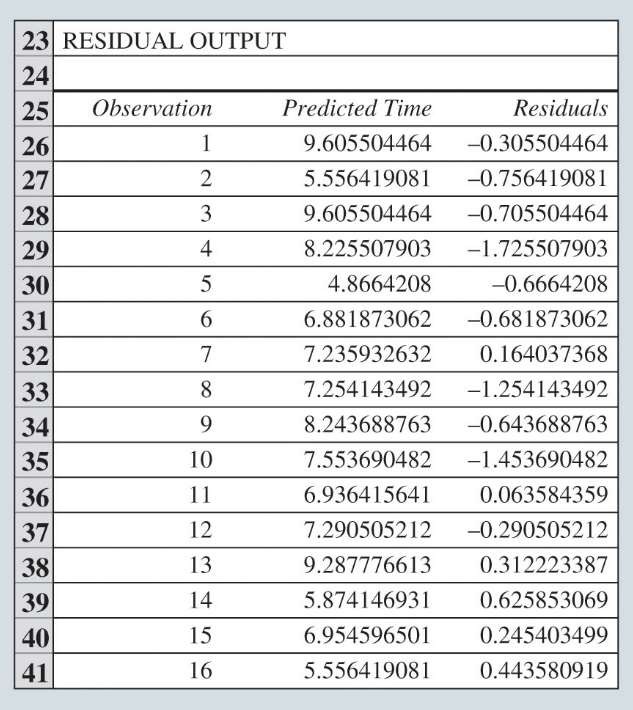
* 1.Center of the residuals should be approximately 0.
  + Mean 0
* 2. The spread in data should be about the same through out
  + Constant variance
* 3. Errors should be symmetrically distributed with values near 0 occurring more frequently
  + Normally Distributed
* 4. Independent
  + Current data points do not depend on previous points

These residuals do **NOT** look good!



Inference and Regression

Figure 7.18: Excel Residual Plots for the Butler Trucking Company Multiple Regression



Inference and Regression

Table of the First Several Predicted Values 𝑦^ and Residuals 𝑒 Generated by the Excel Regression Tool

Scatter chart of 𝑦^ vs Residuals 𝑒 –

- used to assess whether the regression model satisfies the conditions needed for inference



Inference and Regression

Scatter Chart of Predicted Values 𝑦^ and Residuals 𝑒

* Mean 0
* Similar Variance
* Concentrated around 0

No evidence for violation of the conditions

=> Trust the statistical inference!



Inference and Regression

**Testing Individual Regression Parameters:**

To determine whether statistically significant relationships exist between the dependent variable *y* and each of the independent variables 𝑥1, 𝑥2, … , 𝑥q , individually

If ** = 0, there is no linear relationship between the dependent variable *y*

*j*

and the independent variable *xj* .

If **  0, there is a linear relationship between *y* and *x* .

*j*

*j*

𝐻0: βj = 0

𝐻a: βj ≠ 0



Inference and Regression

**Testing Individual Regression Parameters (cont.):**

* Use a t test to test the Null Hypothesis
* The test statistic for this t test is,

𝑡 = 𝑠

𝑏j

bj

Where 𝑠bj is the estimated standard deviation of 𝑏j

* As the magnitude of *t* increases (as t deviates from zero in either direction),
  + we are more likely to reject the hypothesis that the regression parameter 𝛽j is 0.
  + Implies 𝛽j ≠ 0 and there is a relationship between y and 𝑥j



Inference and Regression

**Testing Individual Regression Parameters (cont.):**

* Typically, most software will provide a p-value to determine if 𝛽j is significant (not equal to 0)
* Confidence interval can be used to test whether each of the regression parameters

𝛽0, 𝛽1, 𝛽2, … , 𝛽q is equal to zero as well.

* **Confidence interval:**
  + An estimated interval believed to contain the value of the parameter at some level of confidence.
    - Example 95% confidence interval

𝑏j ± 𝑡a/2𝑆bj

* **Confidence level:** α - Alpha
  + Indicates how frequently interval estimates will contain the true value of the parameter we are estimating.
    - Example = 0.05

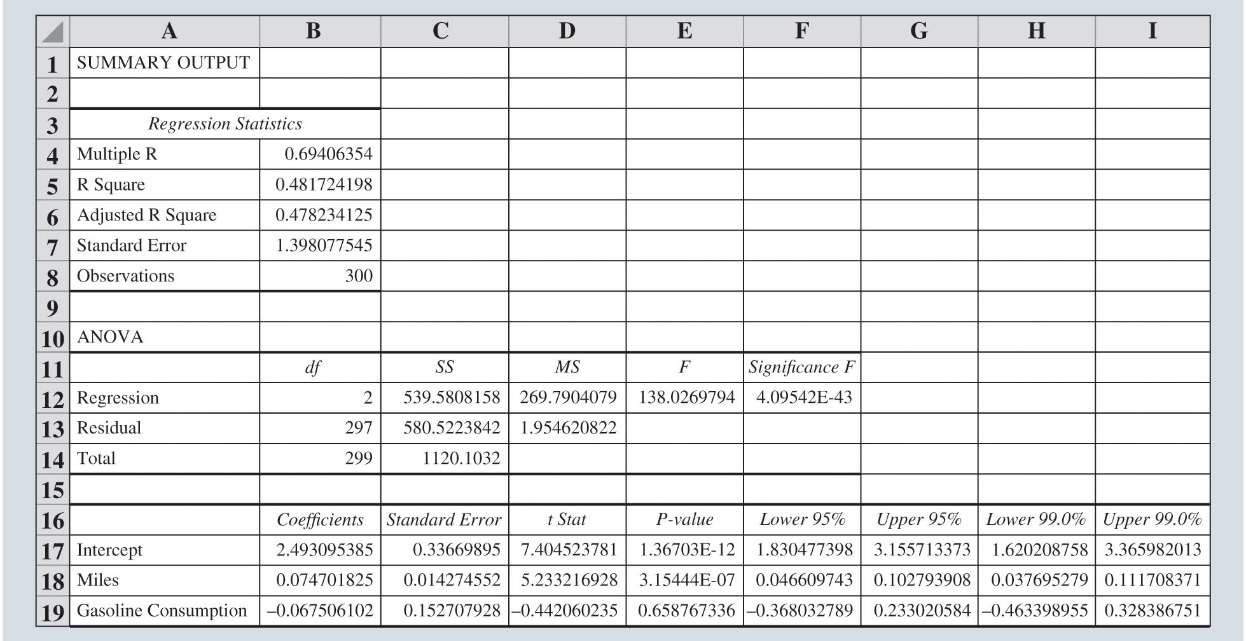


Inference and Regression

**Addressing Nonsignificant Independent Variables:**

* If practical experience dictates that the nonsignificant independent variable has a relationship with the dependent variable
  + the independent variable should be left in the model.
* If the model sufficiently explains the dependent variable without the nonsignificant independent variable
  + then consider rerunning the regression without the nonsignificant independent variable.
* The appropriate treatment of the inclusion or exclusion of the y-intercept

when *b*0 is not statistically significant may require special consideration.



Inference and Regression

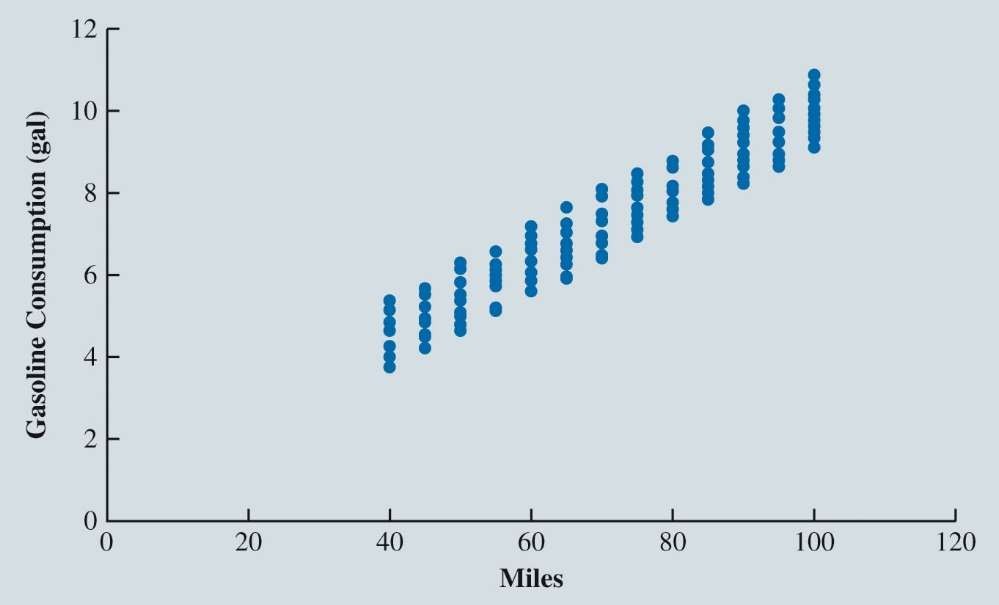
**Multicollinearity:**

* the correlation among the independent variables in multiple regression analysis.
* In *t* tests for the significance of individual parameters, multicollinearity may lead to:
  + concluding that a parameter associated with one of the multicollinear independent variables is not significantly different from zero when the independent variable actually has a strong relationship with the dependent variable.
* This problem is avoided when there is little correlation among the independent variables.



Inference and Regression

Figure 7.21: Excel Regression Output for the Butler Trucking Company with Miles and Gasoline Consumption as Independent Variables



Inference and Regression

Figure 7.22: Scatter Chart of Miles and Gasoline Consumed for Butler Trucking Company



Inference and Regression

**Multicollinearity (cont.):**

* Testing for an overall regression relationship:
  + Use an *F* test based on the *F* probability distribution.
  + If the *F* test leads us to reject the hypothesis that the values of

*b*1 , *b*2 ,, *bq*

are all zero:

* + - Conclude that there is an overall regression relationship.
    - Otherwise, conclude that there is no overall regression relationship.



Inference and Regression

Multicollinearity (cont.):

* Testing for an overall regression relationship (cont.):
  + The test statistic generated by the sample data for this test is:

*F* =

SSR *q*

SSE (*n*  *q*  1)

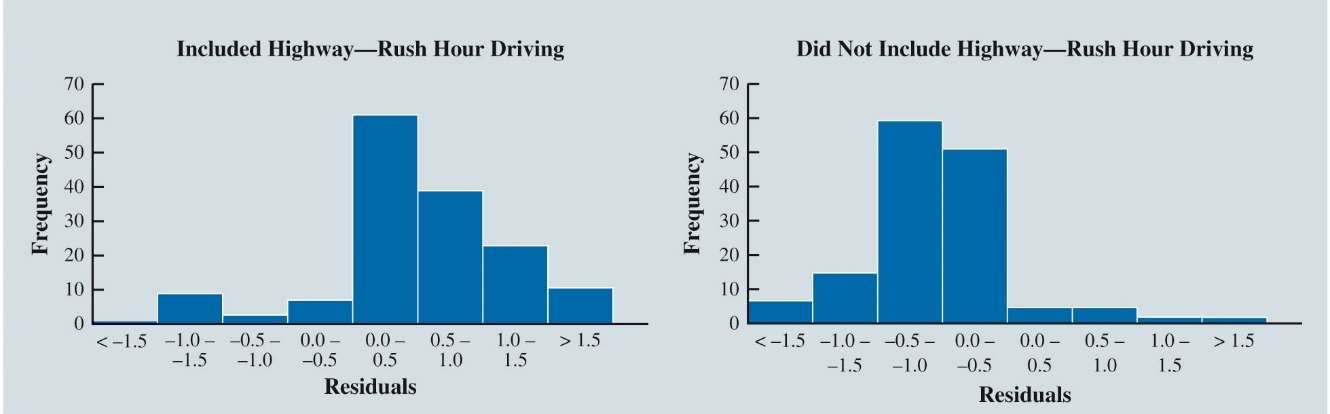
* SSR = Sum of squares due to regression.
* SSE = Sum of squares due to error.
* *q* = the number of independent variables in the regression model.
* *n* = the number of observations in the sample.
* Larger values of *F* provide stronger evidence of an overall regression relationship.
* For a small p-value => Reject null and conclude there is a regression relationship



Categorical Independent Variables

Butler Trucking Company and Rush Hour Interpreting the Parameters

More Complex Categorical Variables



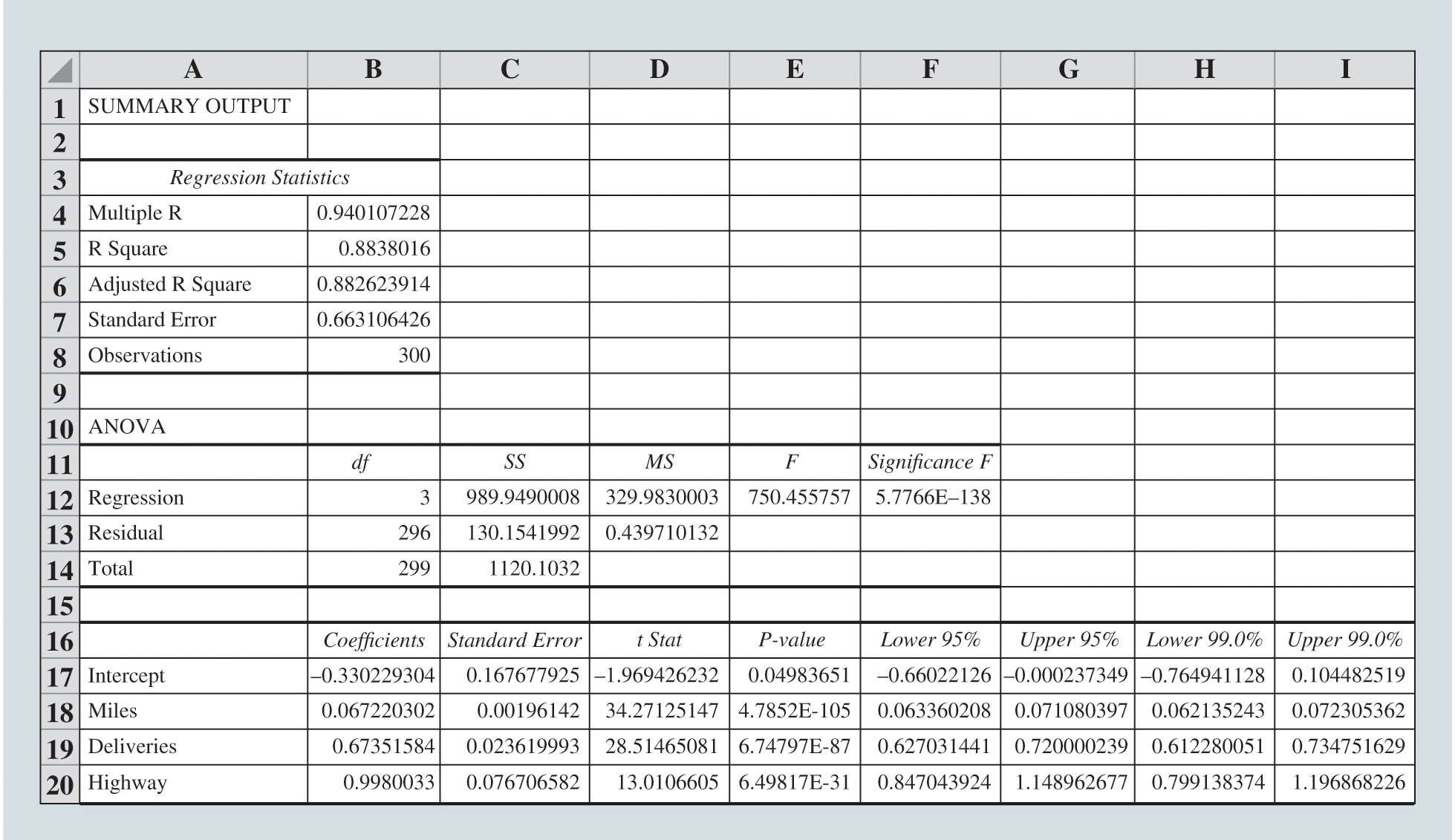
Categorical Independent Variables

**Butler Trucking Company and Rush Hour:**

* Dependent Variable, y : Travel Time
* Independent Variables
  + 𝑥1 - Miles Traveled
  + 𝑥2 - Number of Deliveries
  + 𝑥3 - Rush Hour
    - Categorical Variable
    - 𝑥3 = 0 if delivery trip took place during rush hour
    - 𝑥3 = 1 if delivery trip did not take place during rush hour



Categorical Independent Variables



Categorical Independent Variables

**Excel Data and Output for Butler Trucking with**

Miles Traveled (𝑥1),

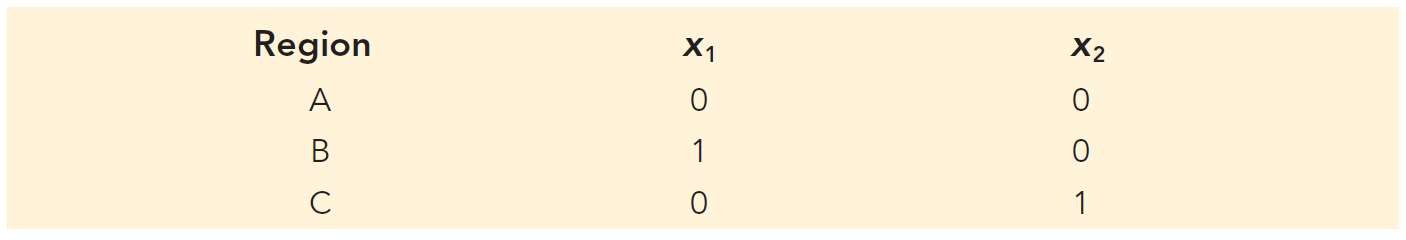
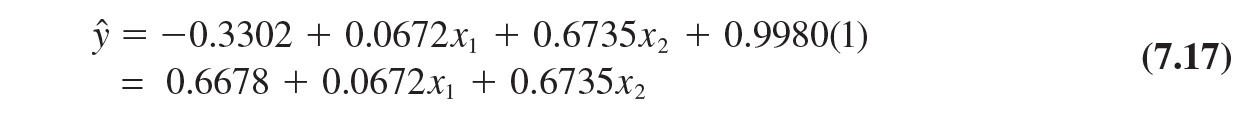
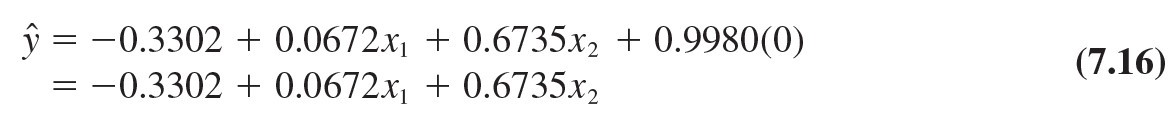
Number of Deliveries (𝑥2), and the Highway Rush Hour Dummy Variable (𝑥3), as the Independent Variables



Categorical Independent Variables

**Interpreting the Parameters:**

* The model estimates that travel time **increases** by:
  + **0.0672 hours (about 4 minutes)** for every increase of 1 mile traveled, holding all other variables constant
  + **0.6735 hours (about 40 minutes)** for every delivery, holding all other variables constant
  + **0.9980 hours (about 60 minutes)** if the driving route took place during the afternoon rush hour period, holding all other variables constant
  + 𝑅2 = 0.8838
    - indicates that the regression model explains approximately 88.4% of the variability in travel time for the driving assignments in the sample



Categorical Independent Variables

**Interpreting the Parameters (cont.):**

Compare the regression model for the case when 𝑥3 = 0 and when 𝑥3 = 1.

When *x*3 = 0:

When *x*3 = 1:



Categorical Independent Variables

**More Complex Categorical Variables:**

If a categorical variable has *k* levels, *k* minus 1 dummy variables are required, with each dummy variable corresponding to one of the levels of the categorical variable and coded as 0 or 1.

* Example:
  + Suppose a manufacturer of vending machines organized the sales territories for a particular state into three regions: A, B, and C.
  + Sales Region – Categorical variable with 3 levels (A, B, C)
  + Number of Dummy Variables = 3-1 = 2



Categorical Independent Variables

**More Complex Categorical Variables:**

* **Example Continued:**
  + The regression equation:

*y*ˆ = *b*0 + *b*1 *x*1 + *b*2 *x*2

* Observations corresponding to Region A -> 𝑥1 = 0, 𝑥2 = 0,
  + Estimated mean number of units sold in Region A

*y*ˆ = *b* + *b* (0) + *b* (0) = *b*

0 1

2

0



Categorical Independent Variables

**More Complex Categorical Variables:**

* **Example Continued:**
  + Observations corresponding to Region B -> 𝑥1 = 1, 𝑥2 = 0,
  + Estimated number of units sold in Region B:

*y*ˆ = *b* + *b* (1) + *b* (0) = *b* + *b*

0 1

2

0 1

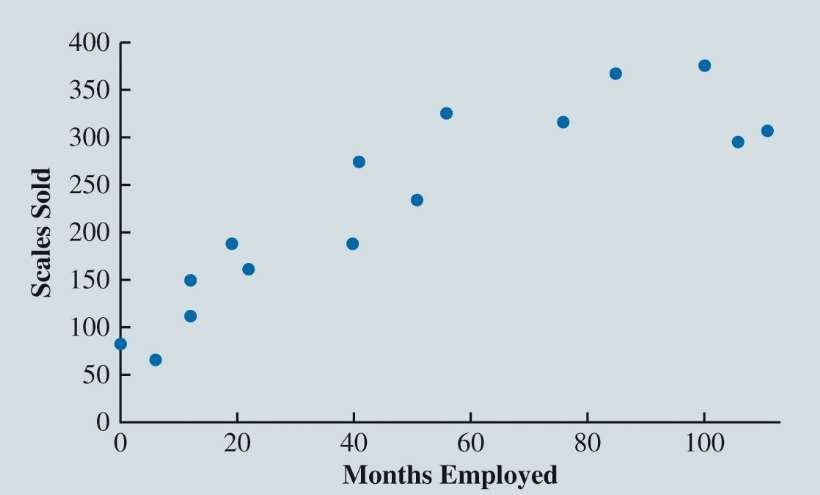
* Observations corresponding to Region C -> 𝑥1 = 0, 𝑥2 = 1,
* Estimated number of units sold in Region C:

*y*ˆ = *b* + *b* (0) + *b* (1) = *b* + *b*

0 1

2

0 2



Modeling Nonlinear Relationships

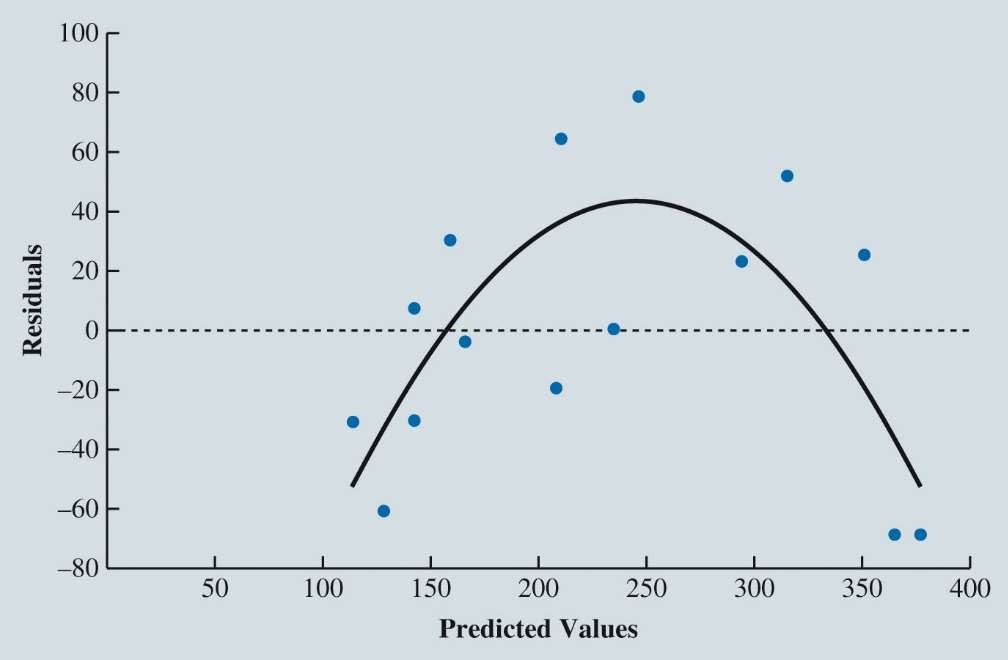
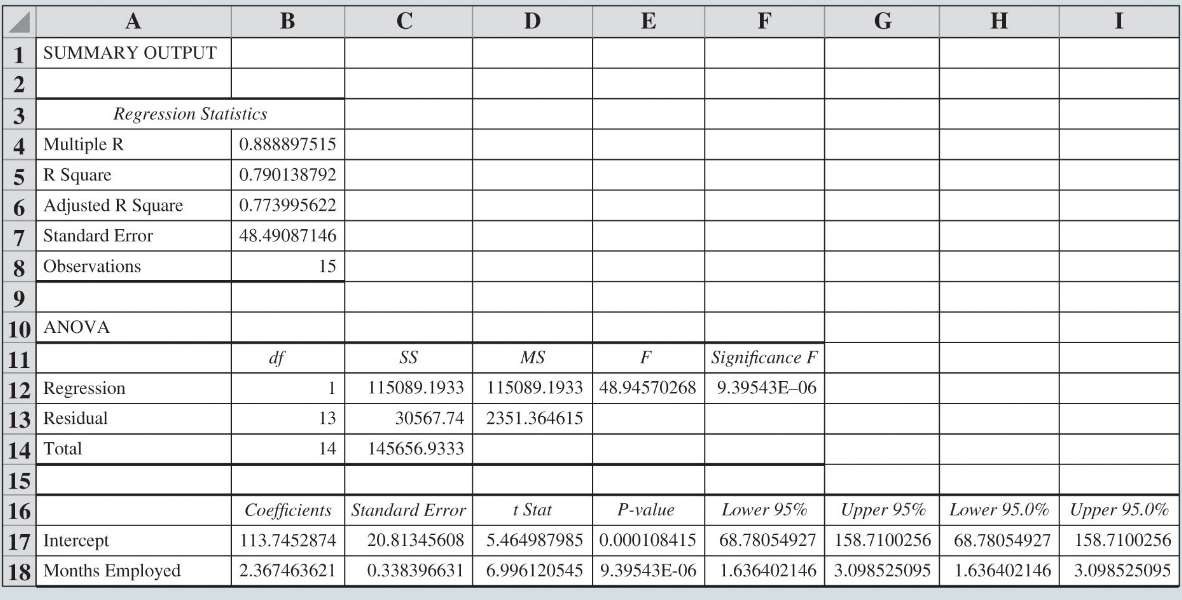
Quadratic Regression Models Piecewise Linear Regression Models

Interaction Between Independent Variables



Modeling Nonlinear Relationships

Figure 7.25: Scatter Chart for the Reynolds Example



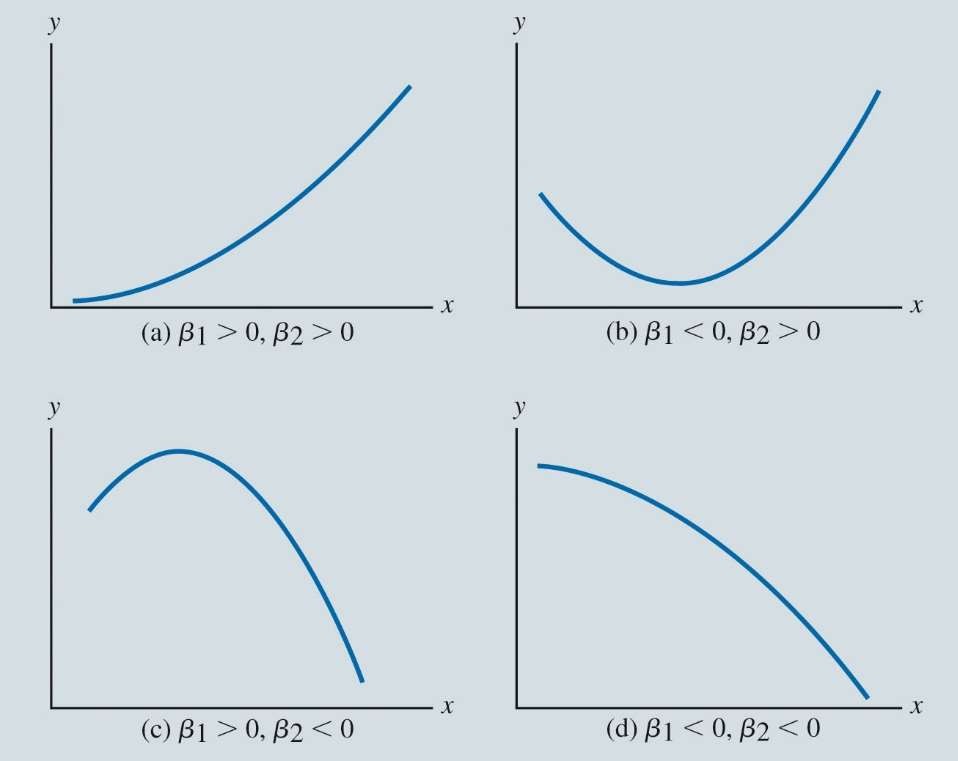
Modeling Nonlinear Relationships

Figure 7.26: Excel Regression Output for the Reynolds Example



Modeling Nonlinear Relationships

Figure 7.27: Scatter Chart of the Residuals and Predicted Values of the Dependent Variable for the Reynolds Simple Linear Regression



Modeling Nonlinear Relationships

* Equation (7.18) corresponds to a **quadratic regression model.**



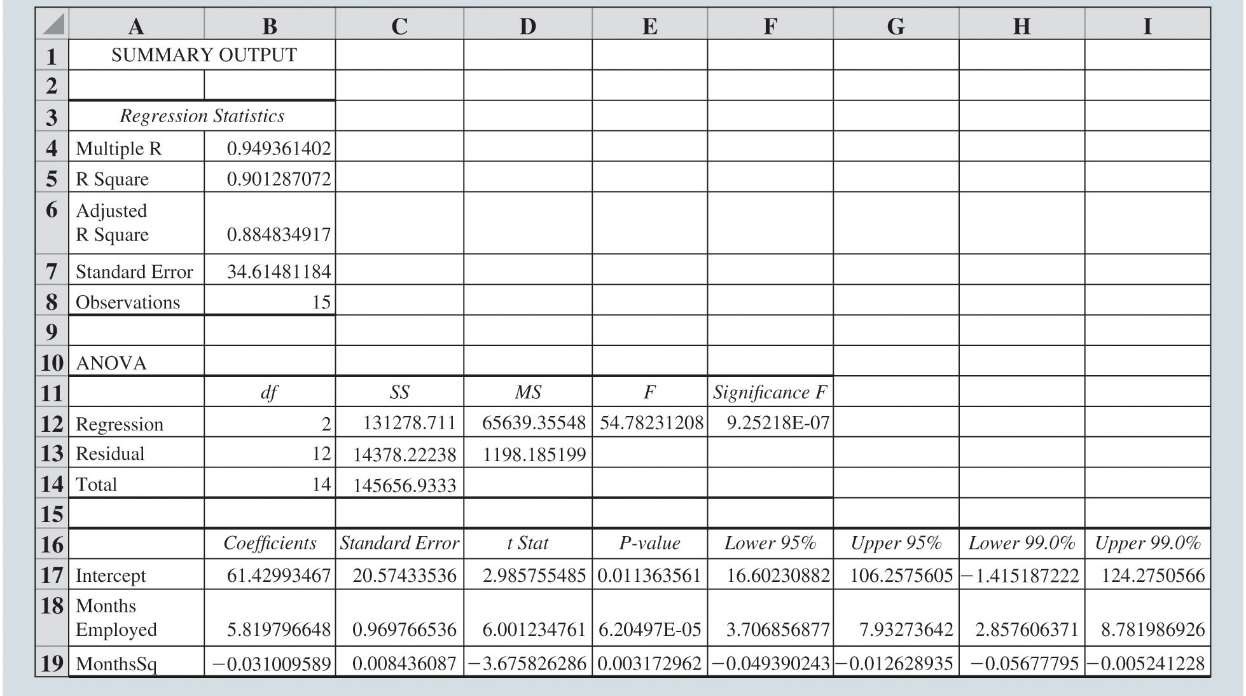
Quadratic Regression Models:

* In the Reynolds example,
  + To account for the curvilinear relationship between months employed and scales sold,
  + include the square of the number of months the salesperson has been employed



Modeling Nonlinear Relationships

Figure 7.28: Relationships That Can Be Fit with a Quadratic Regression Model



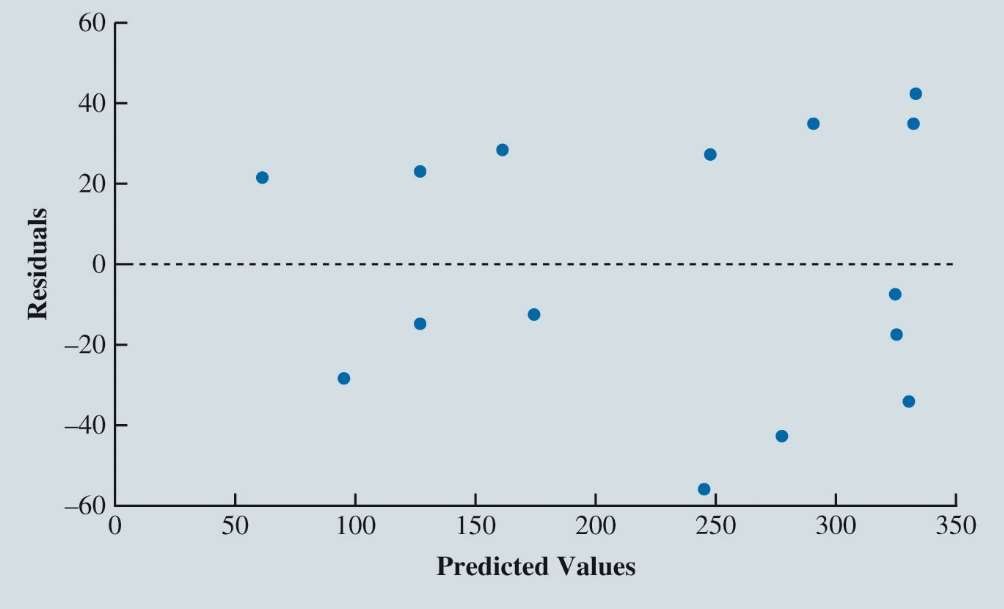
Modeling Nonlinear Relationships

Figure 7.29: Excel Data for the Reynolds Quadratic Regression Model



Modeling Nonlinear Relationships

Figure 7.30: Excel Output for the Reynolds Quadratic Regression Model



Modeling Nonlinear Relationships

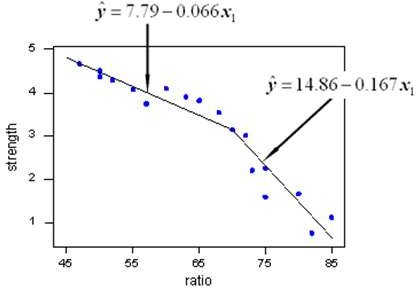
Figure 7.31: Scatter Chart of the Residuals and Predicted Values of the Dependent Variable for the Reynolds Quadratic Regression Model

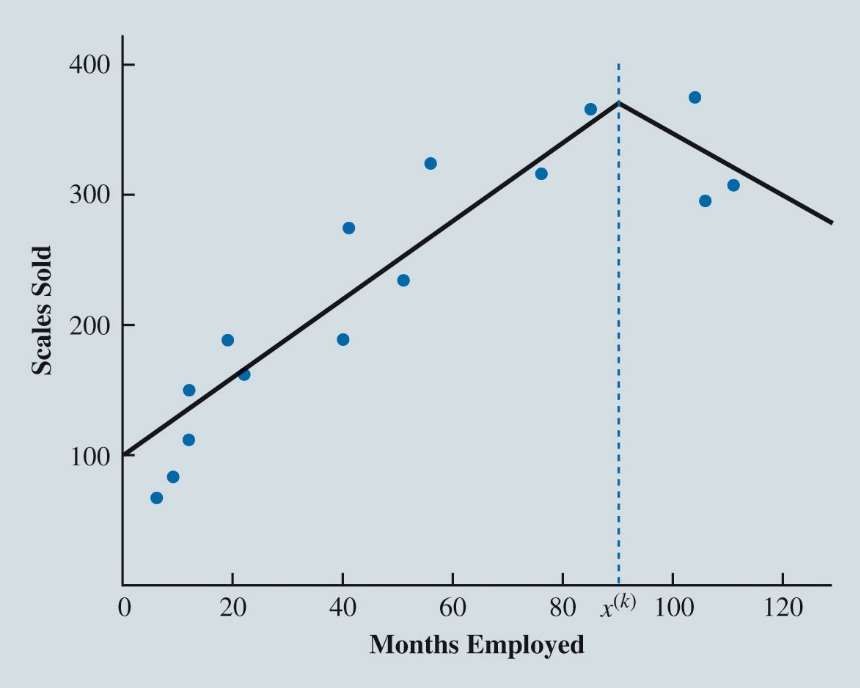
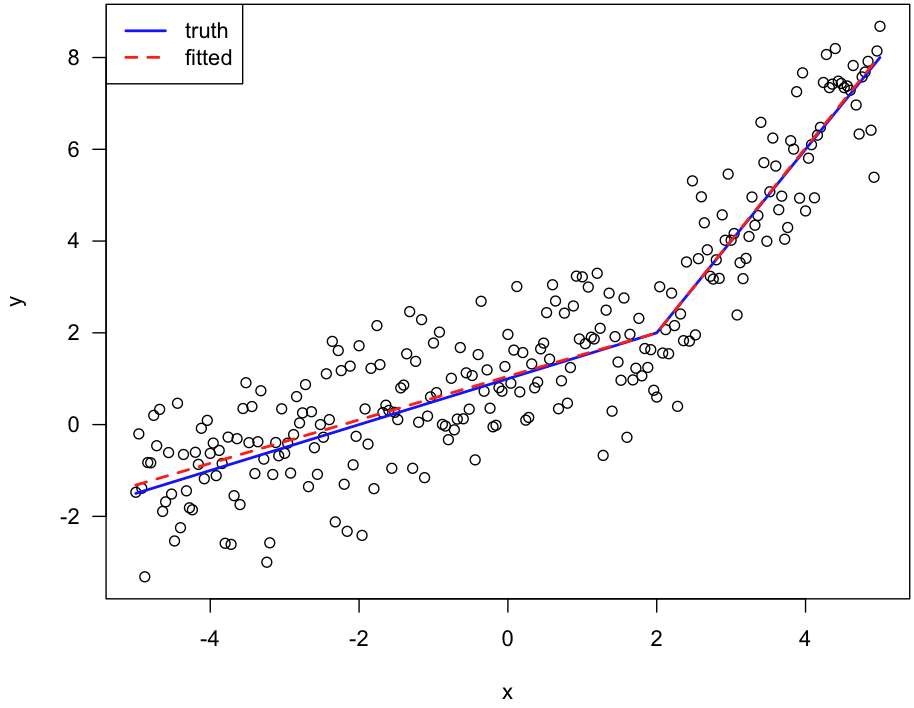


Modeling Nonlinear Relationships

**Piecewise Linear Regression Models:**

* For the Reynolds data, as an alternative to a quadratic regression model:
  + Recognize that up to a certain point of Months Employed
    - the relationship between Months Employed and Sales appears to be positive and linear.
  + After this point,
    - the relationship between Months Employed and Sales appears to be negative and linear
* **Piecewise linear regression model**:
  + This model will allow us to fit these relationships as two linear regressions
    - joined at the value of Months where the relationship between Months Employed and Sales changes.





Modeling Nonlinear Relationships

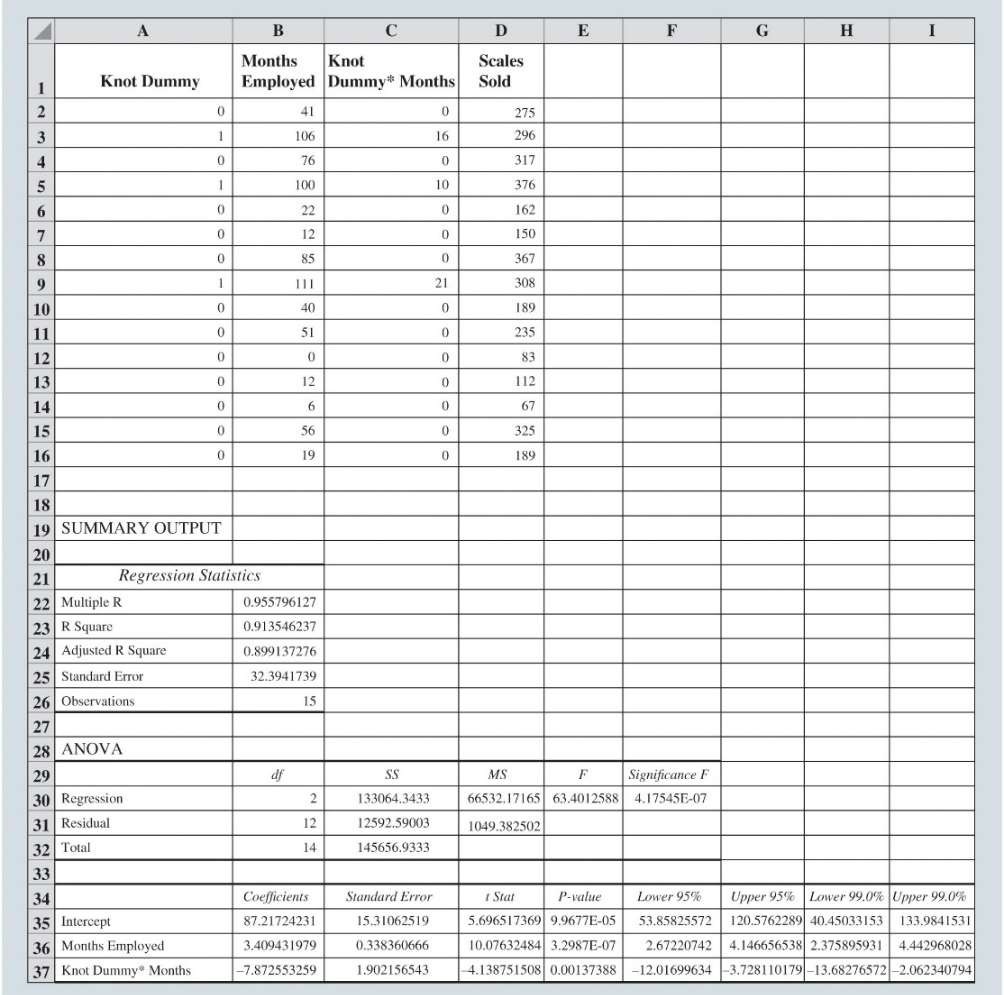
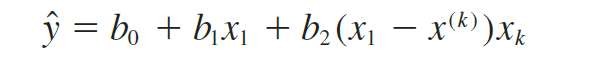
**Piecewise Linear Regression Models (cont.):**

* **Knot**:
  + The value of the independent variable at which the relationship between dependent variable and independent variable changes;
  + also called *breakpoint.*



Modeling Nonlinear Relationships

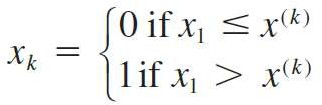
Figure 7.32: Possible Position of Knot *x*(*k*)



Modeling Nonlinear Relationships

**Piecewise Linear Regression Models (cont.):**

* Define a dummy variable:



*x*1 = Months.

*x*(*k*) = value of the knot (90 months for the Reynolds example).

*xk* = the knot dummy variable.

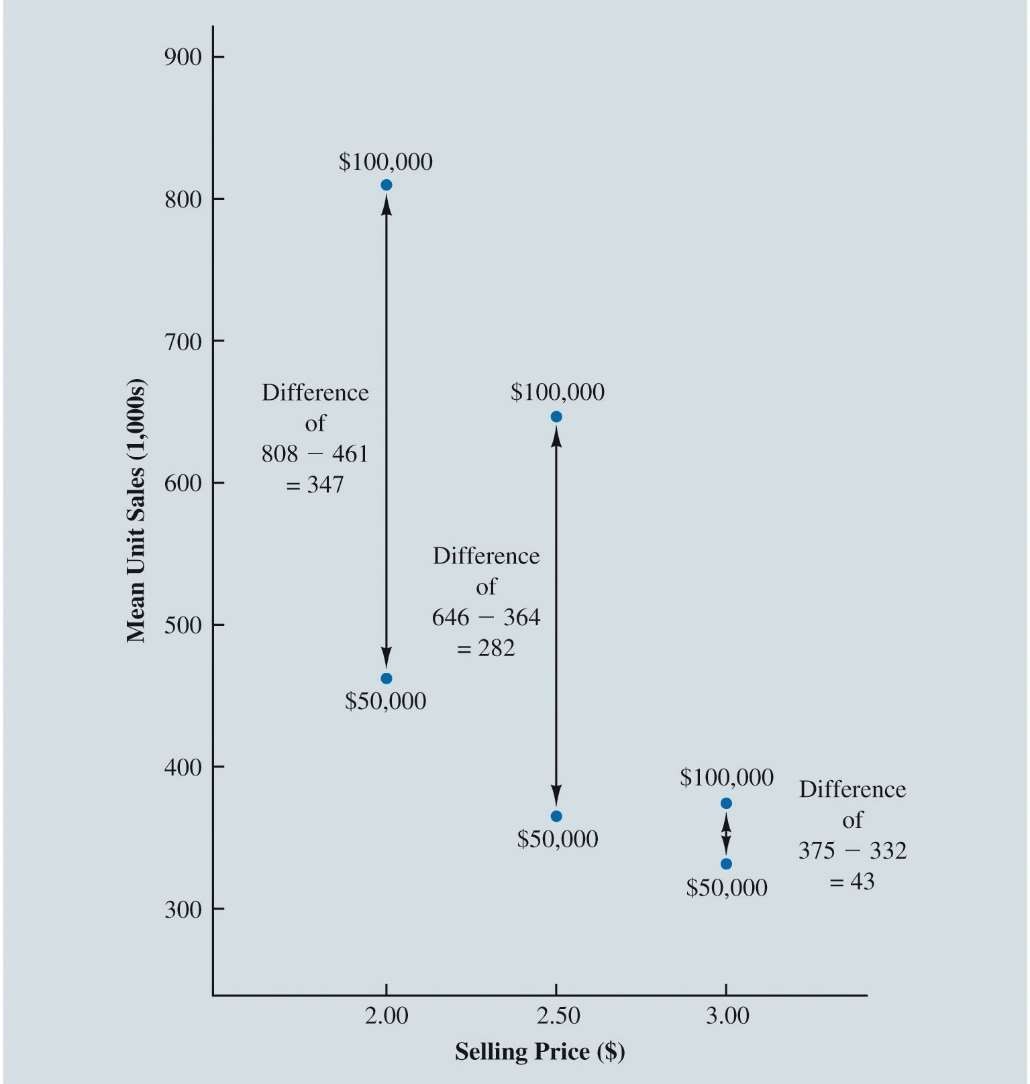
* Then fit the following estimated regression equation:



Modeling Nonlinear Relationships

Figure 7.33: Data and Excel Output for the Reynolds Piecewise Linear Regression Model





Modeling Nonlinear Relationships

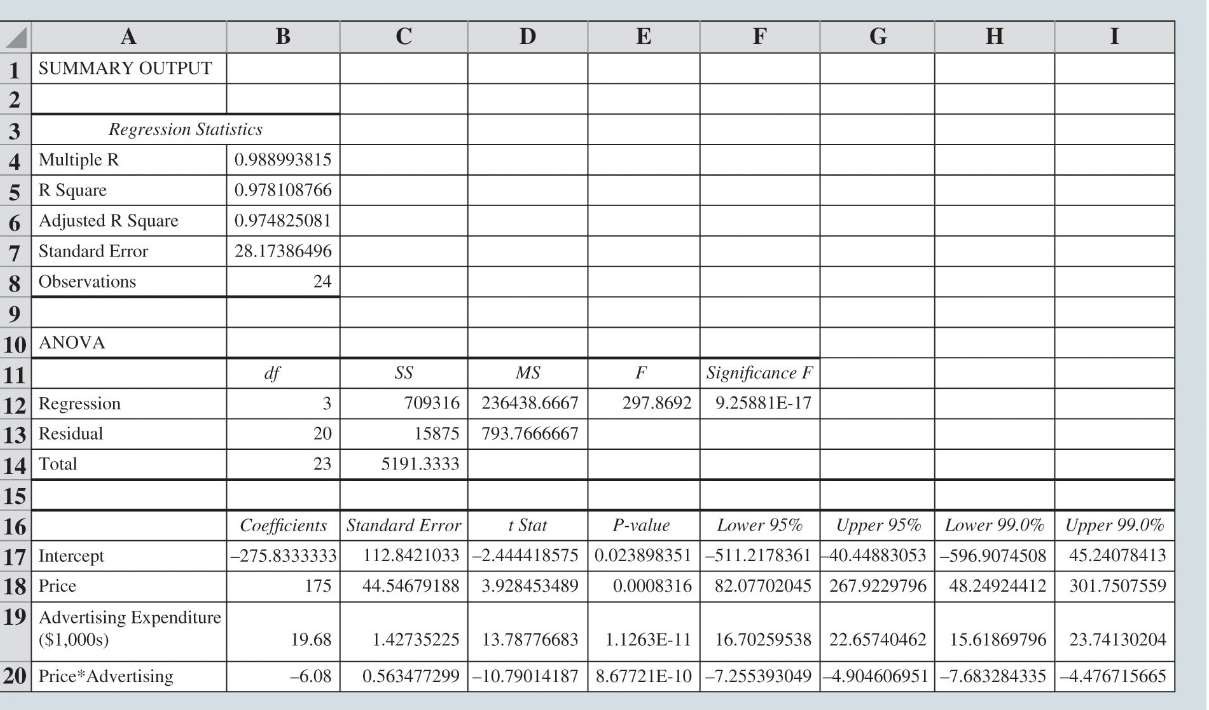
**Interaction Between Independent Variables:**

* **Interaction**:
  + This occurs when the relationship between the dependent variable and one independent variable is different at various values of a second independent variable.
* The estimated multiple linear regression equation is given as:



Modeling Nonlinear Relationships

Figure 7.34: Mean Unit Sales (1,000s) as a Function of Selling Price and Advertising Expenditures



Modeling Nonlinear Relationships

Figure 7.35: Excel Output for the Tyler Personal Care Linear Regression Model with Interaction