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

  $$\beta_0, \beta_1, \beta_2, \ldots, \beta_q.$$

  The mean value and/or the predicted value of the dependent variable *y* for specific values of the independent variables $x_1^*, x_2^*, \ldots, x_q^*$

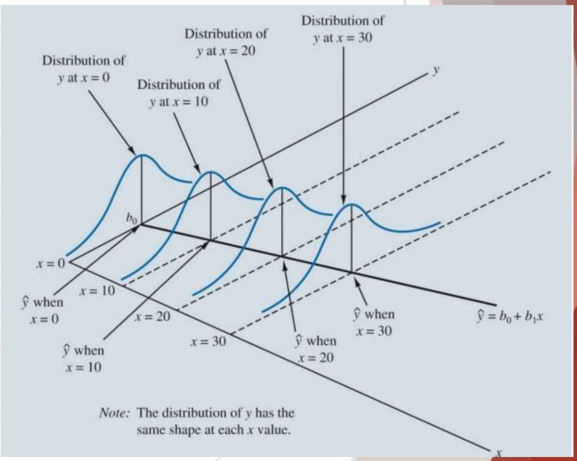
- Consider both **hypothesis testing** and **interval estimation**.

Inference and Regression

Conditions Necessary for Valid Inference in the Least Squares Regression Model:

- 1.For any given combination of values of the independent variables
  - $x_1, x_2, \dots, x_q$, 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

Illustration of the Conditions for Valid Inference in Regression



Note: The distribution of y has the same shape at each x value.

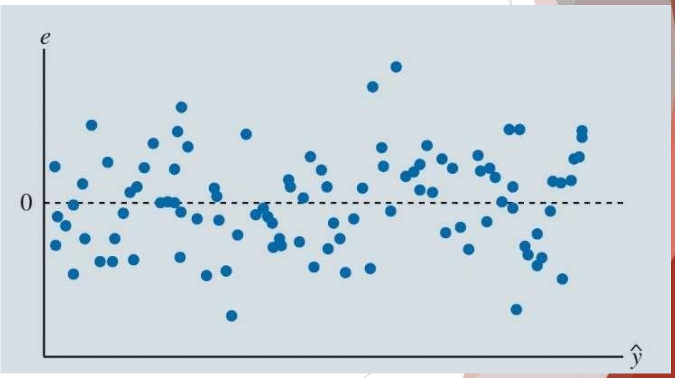
Inference and Regression

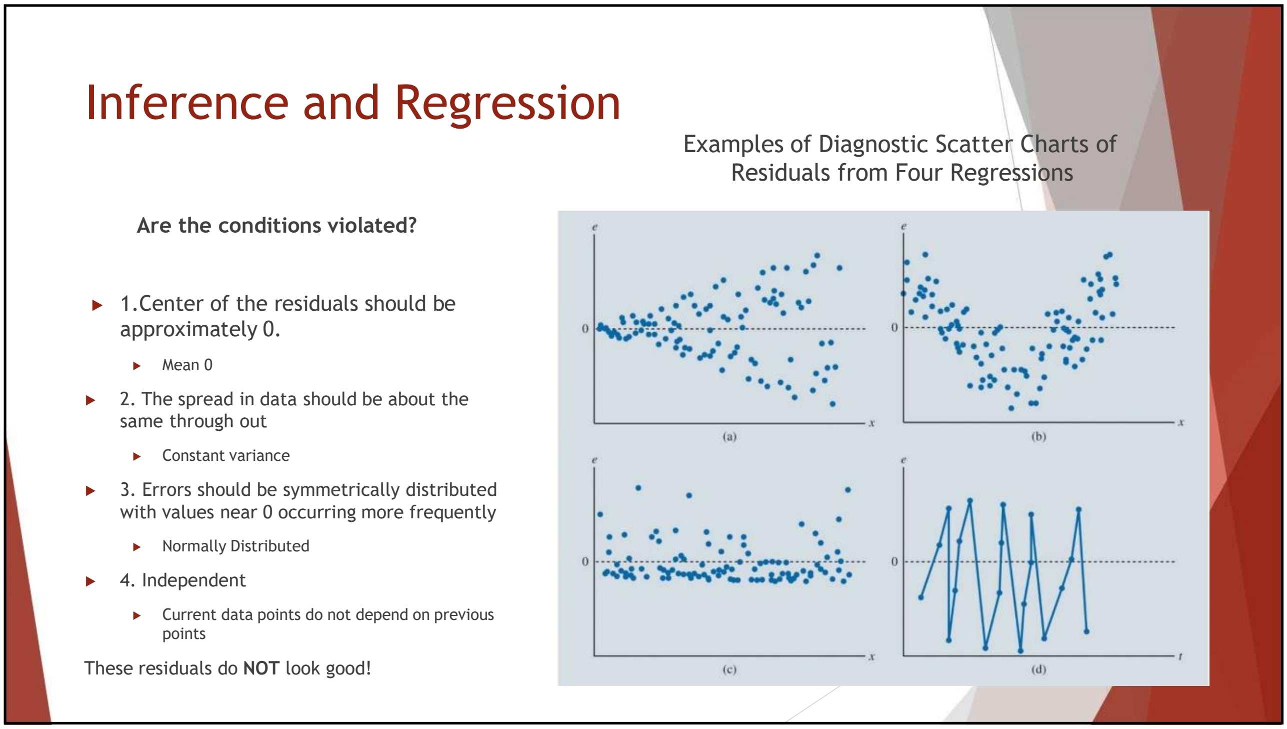
Are the conditions violated?

- 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 look good! – No violations

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





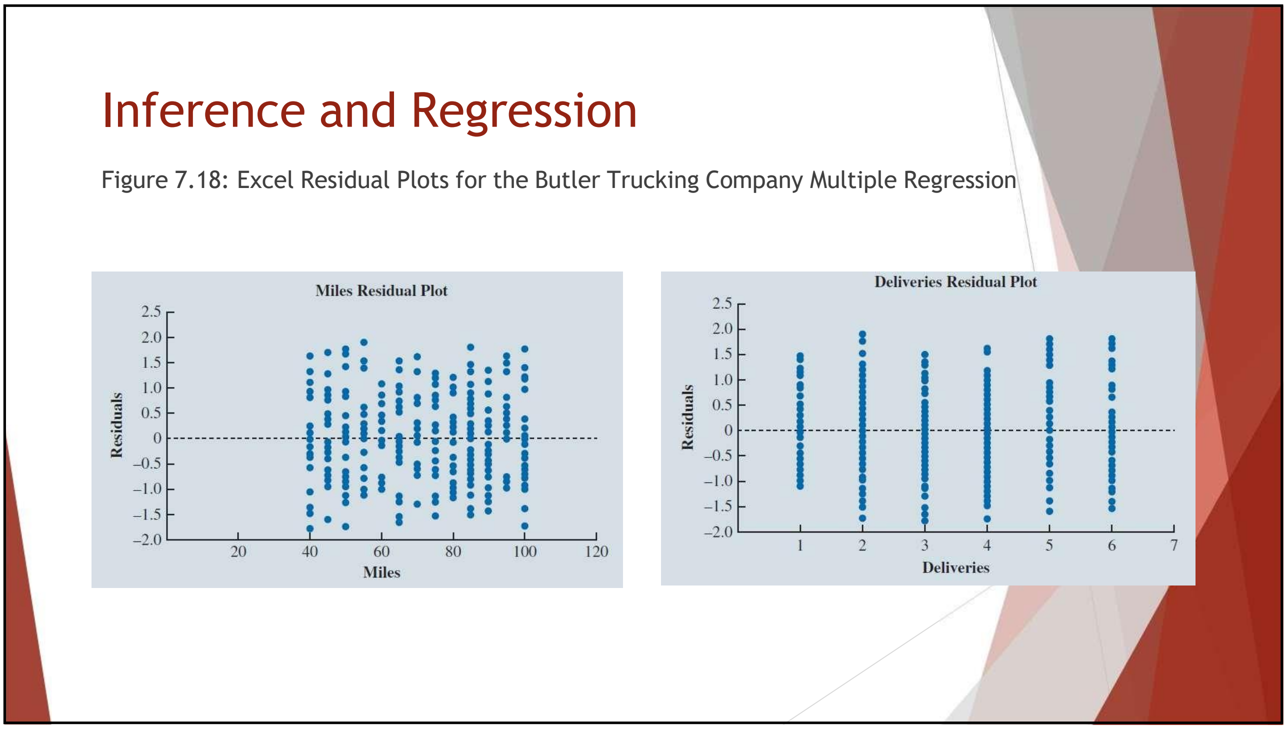
Inference and Regression

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

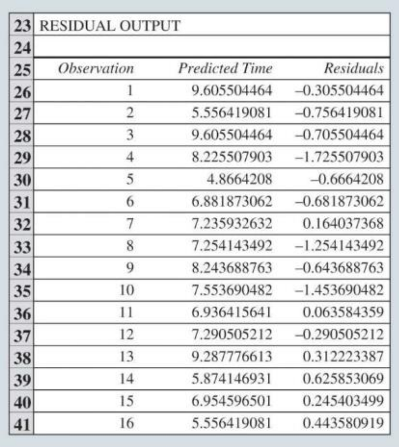
Figure 7.18: Excel Residual Plots for the Butler Trucking Company Multiple Regression

Inference and Regression

Table of the First Several Predicted Values $\hat{y}$ and Residuals e Generated by the Excel Regression Tool

Scatter chart of $\hat{y}$ vs Residuals e -

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



23	RESIDUAL OUTPUT		
24			
25	*Observation*	*Predicted Time*	*Residuals*
26	1	9.605504464	–0.305504464
27	2	5.556419081	–0.756419081
28	3	9.605504464	–0.705504464
29	4	8.225507903	–1.725507903
30	5	4.8664208	–0.6664208
31	6	6.881873062	–0.681873062
32	7	7.235932632	0.164037368
33	8	7.254143492	–1.254143492
34	9	8.243688763	–0.643688763
35	10	7.553690482	–1.453690482
36	11	6.936415641	0.063584359
37	12	7.290505212	–0.290505212
38	13	9.287776613	0.312223387
39	14	5.874146931	0.625853069
40	15	6.954596501	0.245403499
41	16	5.556419081	0.443580919

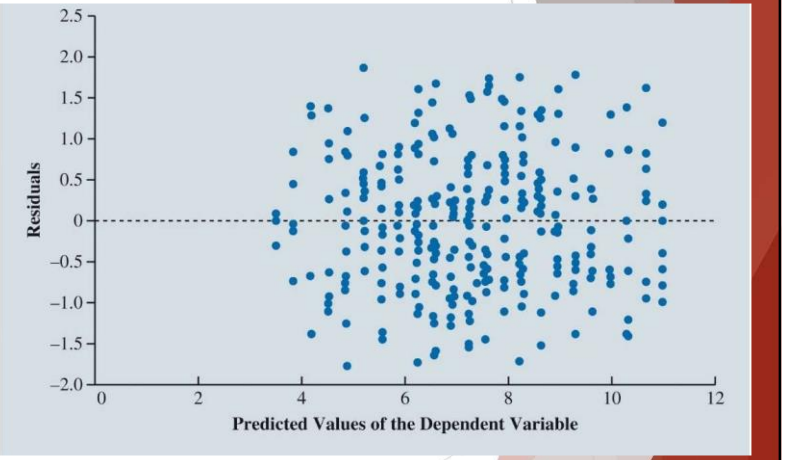
Inference and Regression

Scatter Chart of Predicted Values $\hat{y}$ 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 $x_1, x_2, \ldots, x_q$, individually

If β_j = 0, there is no linear relationship between the dependent variable *y* and the independent variable x_j.

If $\beta_j \neq 0$, there is a linear relationship between *y* and x_j.

$$H_0\colon \beta_j = 0$$
$$H_a\colon \beta_j \neq 0$$

Inference and Regression

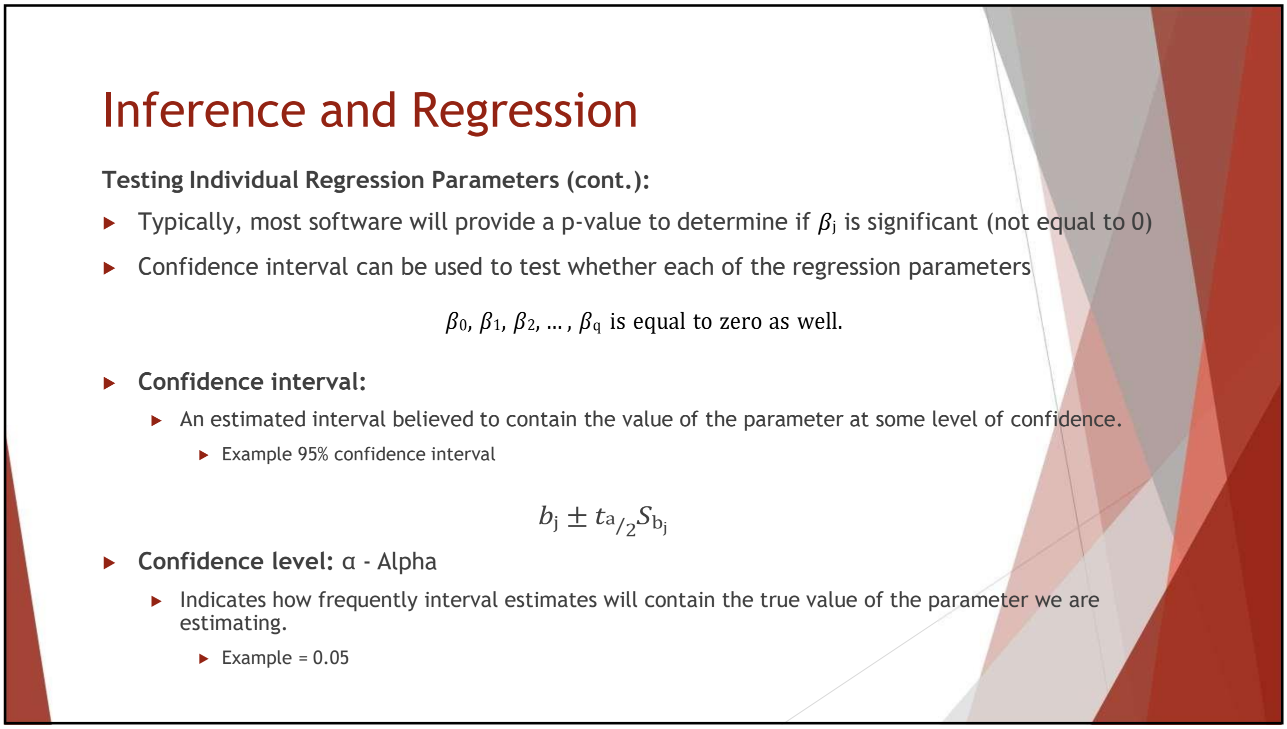
Testing Individual Regression Parameters (cont.):

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

$$t = \frac{b_j}{s_{b_j}}$$

Where s_{b_j} is the estimated standard deviation of b_j

- 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 β_j is 0.
  - Implies $\beta_j \neq 0$ and there is a relationship between y and x_j



Inference and Regression

Testing Individual Regression Parameters (cont.):

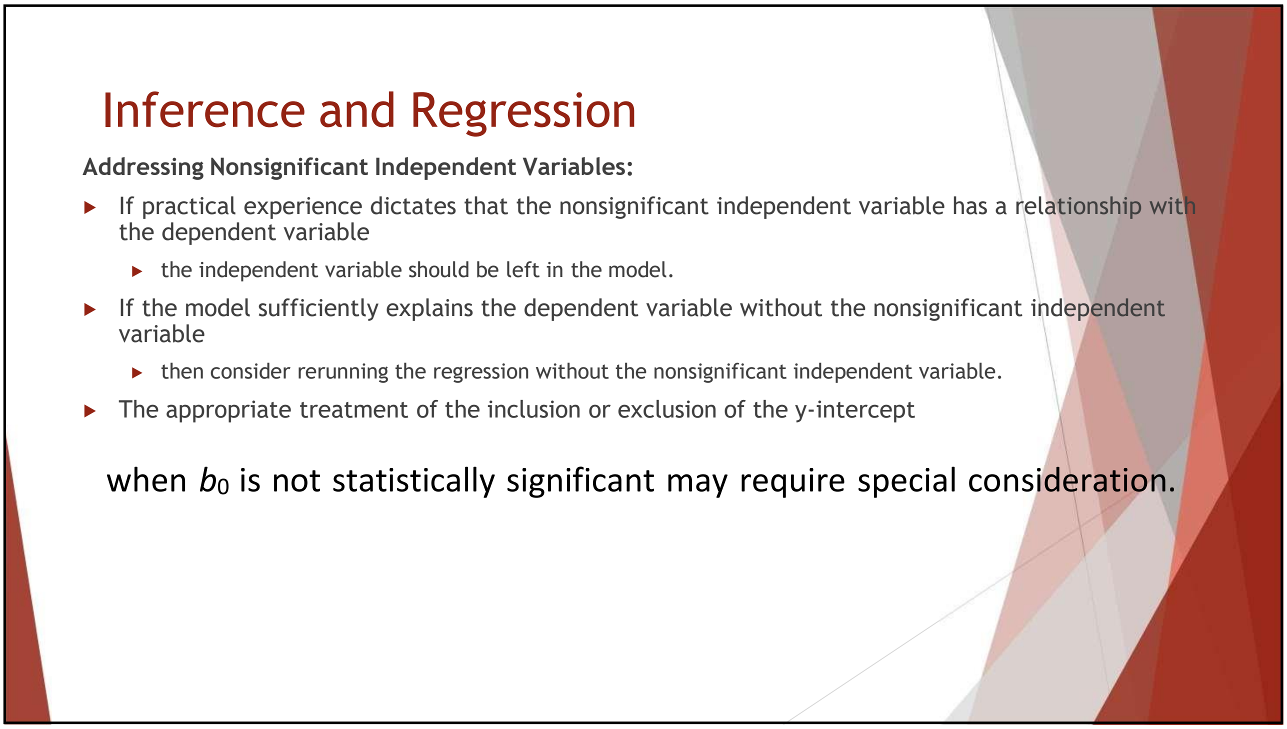
- Typically, most software will provide a p-value to determine if β_j is significant (not equal to 0)
- Confidence interval can be used to test whether each of the regression parameters

$\beta_0, \beta_1, \beta_2, \ldots, \beta_q$ 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

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

- **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

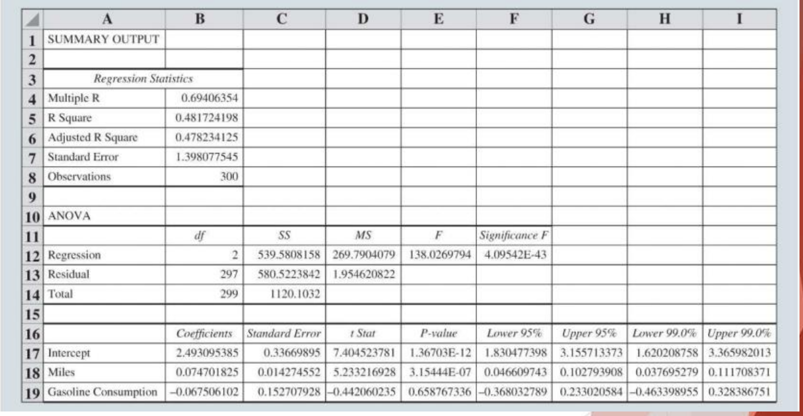
Multicollinearity:

- 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

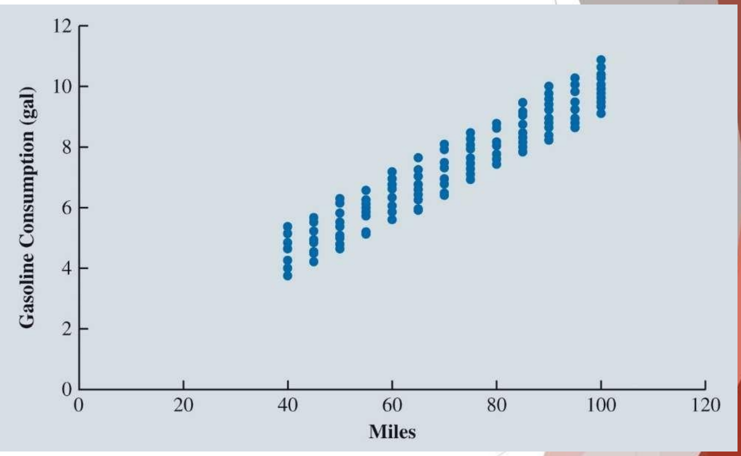
Figure 7.21: Excel Regression Output for the Butler Trucking Company with Miles and Gasoline Consumption as Independent Variables



	A	B	C	D	E	F	G	H	I
1	SUMMARY OUTPUT								
2									
3	*Regression Statistics*								
4	Multiple R	0.69406354							
5	R Square	0.481724198							
6	Adjusted R Square	0.478234125							
7	Standard Error	1.398077545							
8	Observations	300							
9									
10	ANOVA								
11		*df*	*SS*	*MS*	*F*	*Significance F*			
12	Regression	2	539.5808158	269.7904079	138.0269794	4.09542E-43			
13	Residual	297	580.5223842	1.954620822					
14	Total	299	1120.1032						
15									
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 99.0%*	*Upper 99.0%*
17	Intercept	2.493095385	0.33669895	7.404523781	1.36703E-12	1.830477398	3.155713373	1.620208758	3.365982013
18	Miles	0.074701825	0.014274552	5.233216928	3.15444E-07	0.046609743	0.102793908	0.037695279	0.111708371
19	Gasoline Consumption	–0.067506102	0.152707928	–0.442060235	0.658767336	–0.368032789	0.233020584	–0.463398955	0.328386751

Inference and Regression

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



Inference and Regression

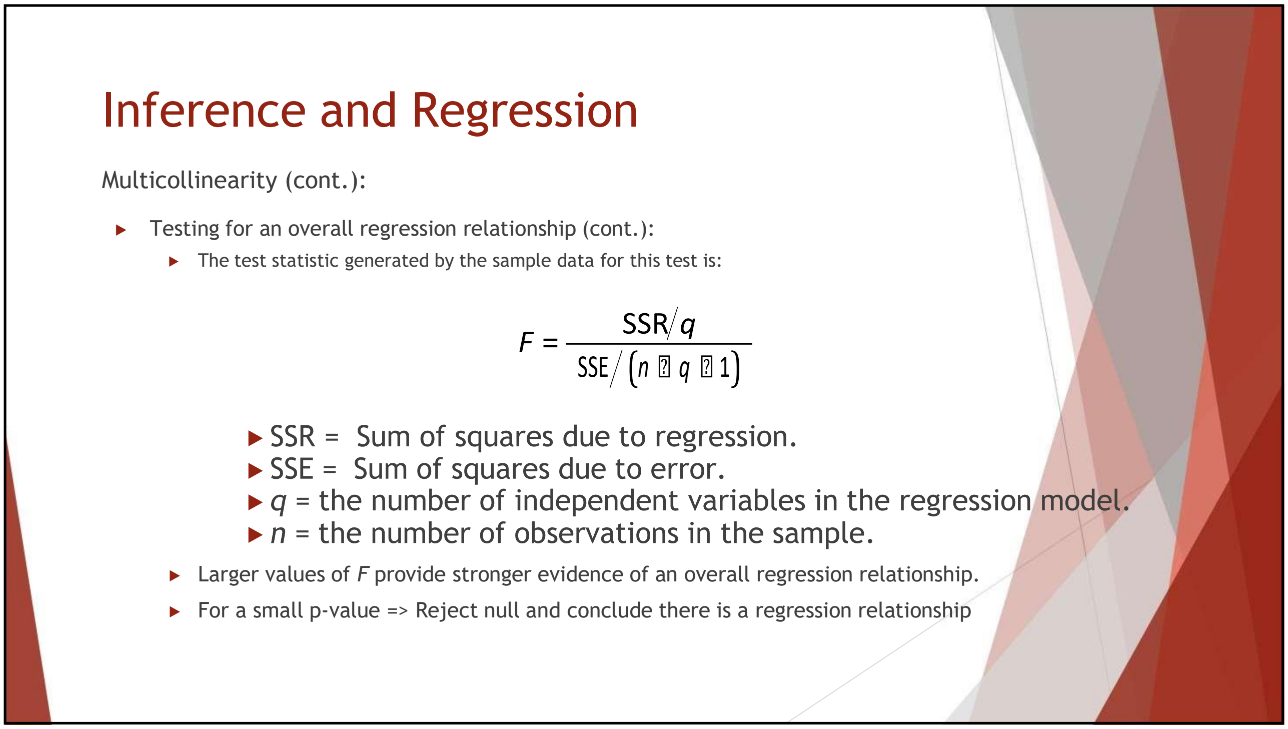
Multicollinearity (cont.):

- 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

$$b_1, b_2, \square, b_q$$

are all zero:

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



Inference and Regression

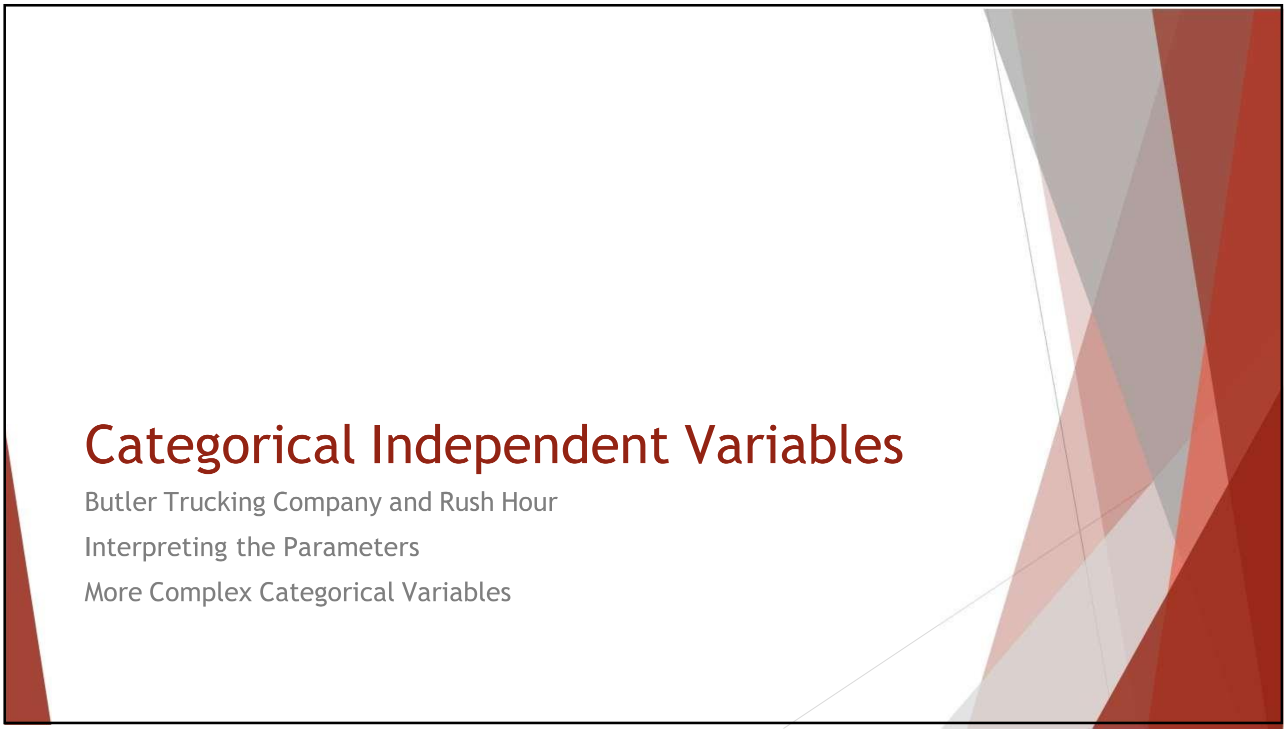
Multicollinearity (cont.):

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

$$F = \frac{\mathrm{SSR}/q}{\mathrm{SSE}/\left(n - q - 1\right)}$$

- 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

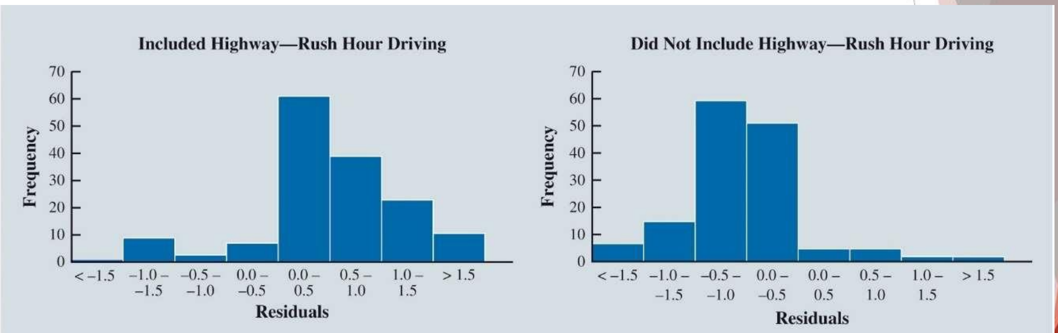
More Complex Categorical Variables

Categorical Independent Variables

Butler Trucking Company and Rush Hour:

- Dependent Variable, y : Travel Time
- Independent Variables
  - x_1 - Miles Traveled
  - x_2 - Number of Deliveries
  - x_3 - Rush Hour
    - Categorical Variable
    - $x_3 = 0$ if delivery trip took place during rush hour
    - $x_3 = 1$ if delivery trip did not take place during rush hour

Categorical Independent Variables

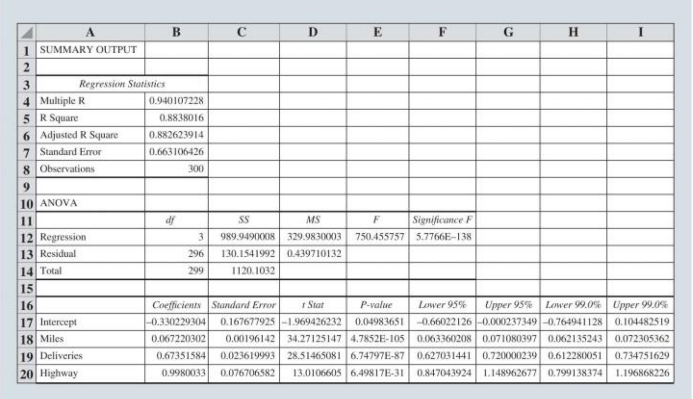


Included Highway—Rush Hour Driving

Did Not Include Highway—Rush Hour Driving

Categorical Independent Variables

Excel Data and Output for Butler Trucking with
Miles Traveled (x_1),
Number of Deliveries (x_2), and the Highway Rush Hour Dummy Variable (x_3), as the Independent Variables

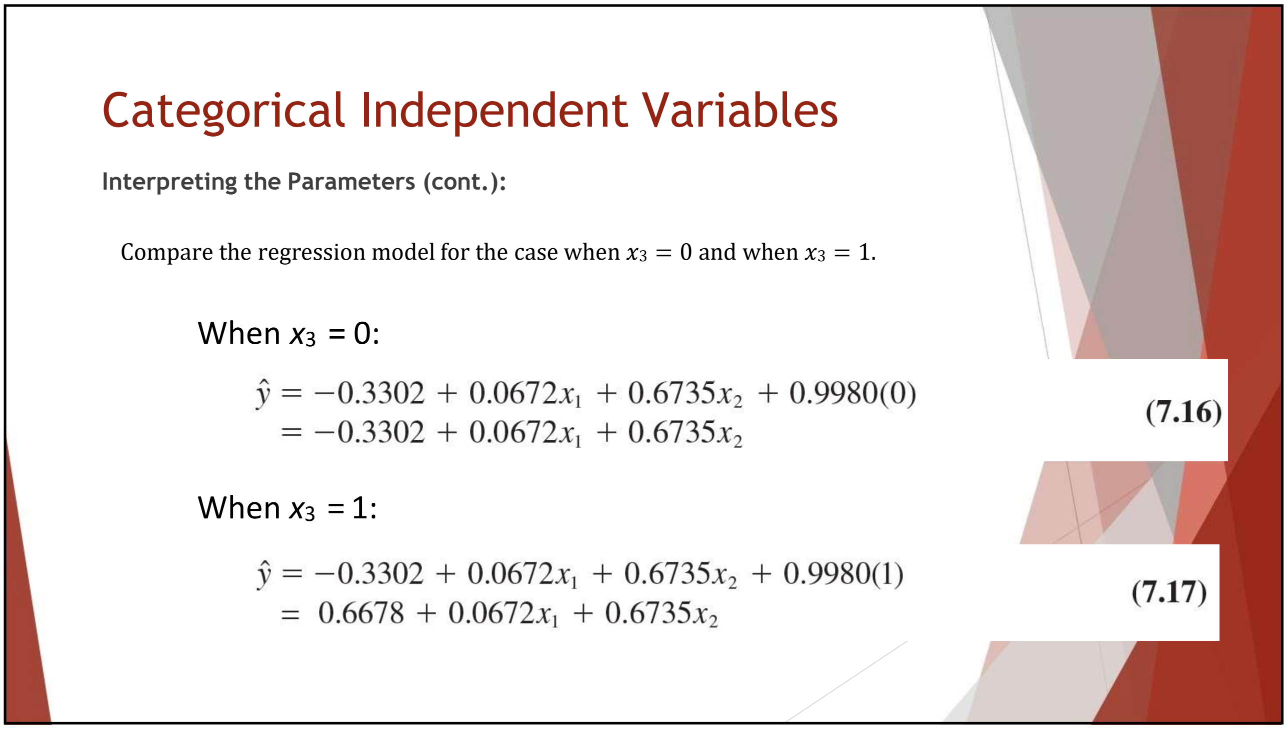


	A	B	C	D	E	F	G	H	I
1	SUMMARY OUTPUT								
2									
3	*Regression Statistics*								
4	Multiple R	0.940107228							
5	R Square	0.8838016							
6	Adjusted R Square	0.882623914							
7	Standard Error	0.663106426							
8	Observations	300							
9									
10	ANOVA								
11		*df*	*SS*	*MS*	*F*	*Significance F*			
12	Regression	3	989.9490008	329.9830003	750.455757	5.7766E–138			
13	Residual	296	130.1541992	0.439710132					
14	Total	299	1120.1032						
15									
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 99.0%*	*Upper 99.0%*
17	Intercept	–0.330229304	0.167677925	–1.969426232	0.04983651	–0.66022126	–0.000237349	–0.764941128	0.104482519
18	Miles	0.067220302	0.00196142	34.27125147	4.7852E-105	0.063360208	0.071080397	0.062135243	0.072305362
19	Deliveries	0.67351584	0.023619993	28.51465081	6.74797E-87	0.627031441	0.720000239	0.612280051	0.734751629
20	Highway	0.9980033	0.076706582	13.0106605	6.49817E-31	0.847043924	1.148962677	0.799138374	1.196868226

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
  - $R^2 = 0.8838$
    - 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.):

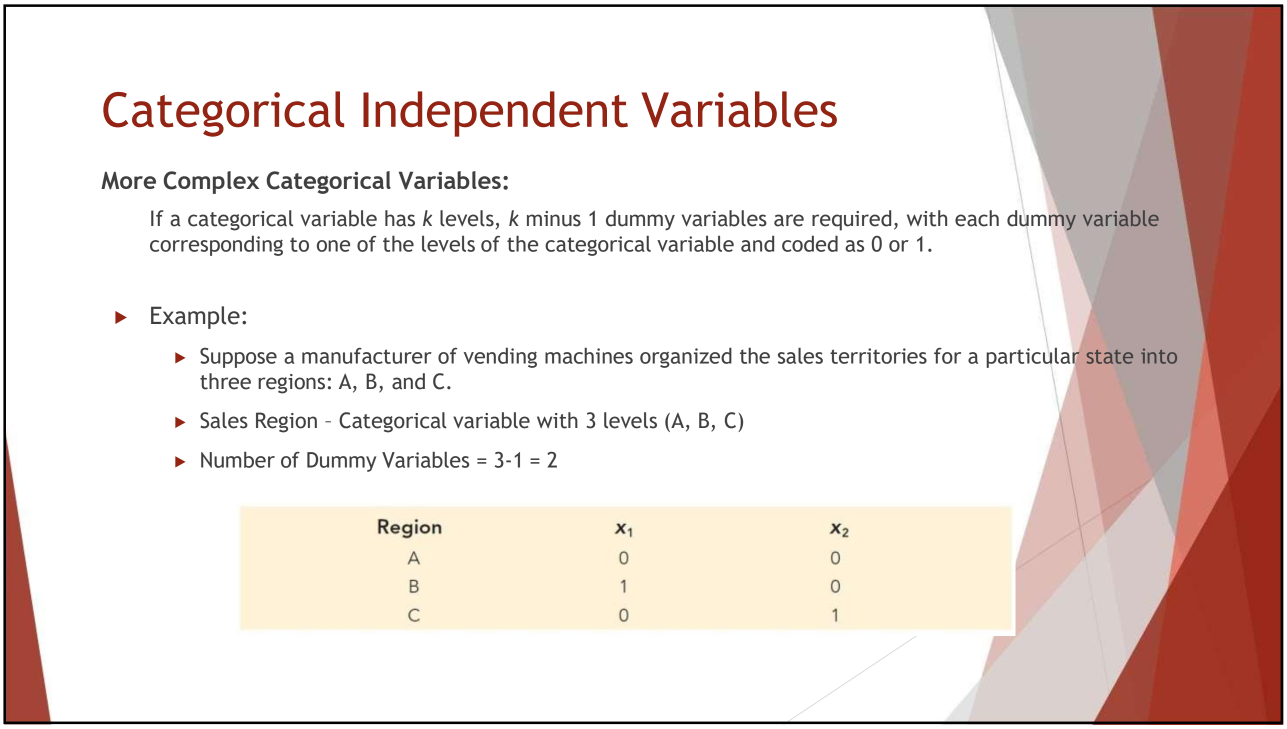
Compare the regression model for the case when $x_3 = 0$ and when $x_3 = 1$.

When $x_3 = 0$:

$$\begin{aligned}\hat{y} &= -0.3302 + 0.0672x_1 + 0.6735x_2 + 0.9980(0) \\ &= -0.3302 + 0.0672x_1 + 0.6735x_2\end{aligned} \qquad \textbf{(7.16)}$$

When $x_3 = 1$:

$$\begin{aligned}\hat{y} &= -0.3302 + 0.0672x_1 + 0.6735x_2 + 0.9980(1) \\ &= 0.6678 + 0.0672x_1 + 0.6735x_2\end{aligned} \qquad \textbf{(7.17)}$$



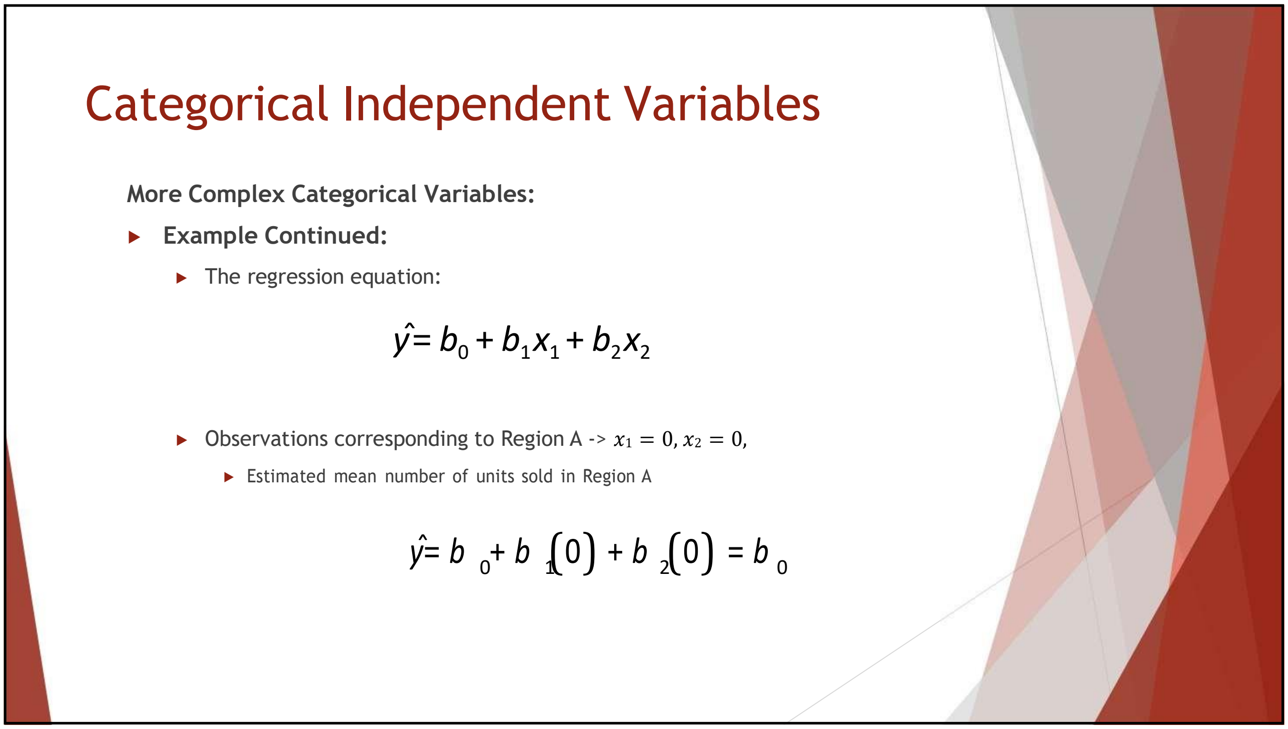
Categorical Independent Variables

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

Region	x_1	x_2
A	0	0
B	1	0
C	0	1



Categorical Independent Variables

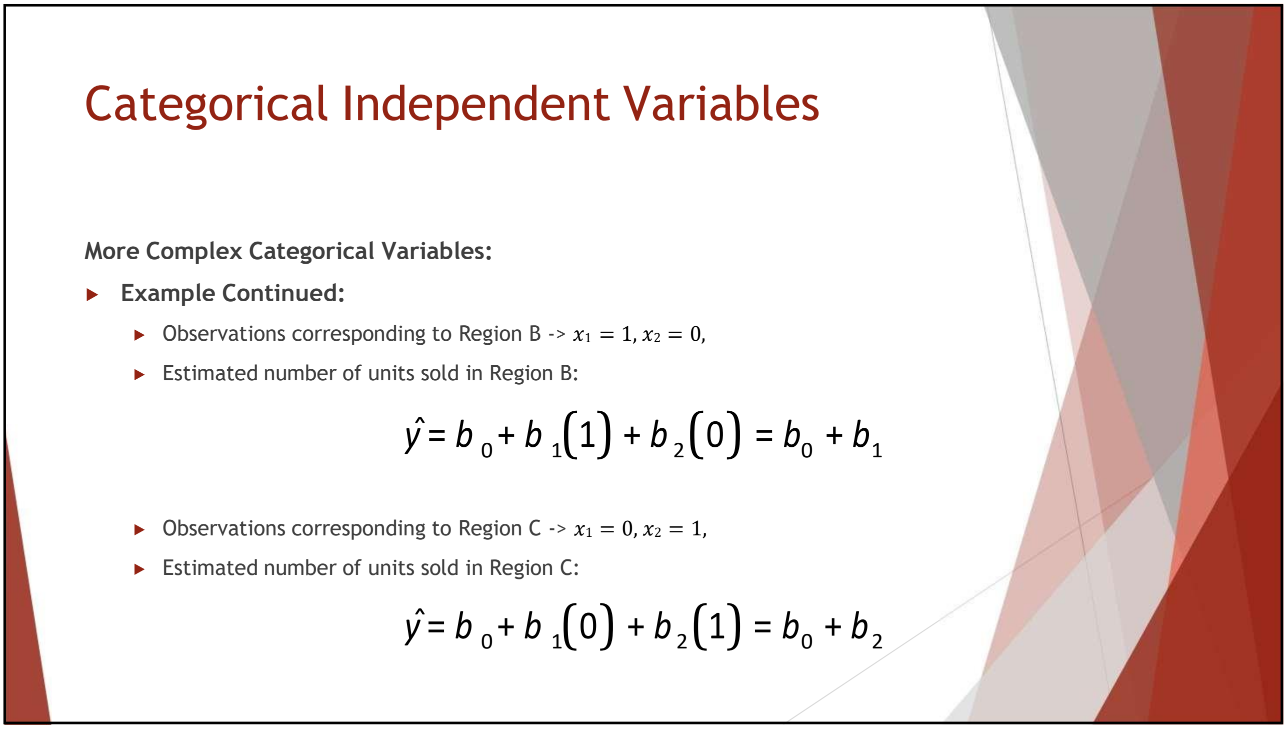
More Complex Categorical Variables:

- **Example Continued:**
  - The regression equation:

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

  - Observations corresponding to Region A -> $x_1 = 0, x_2 = 0,$
    - Estimated mean number of units sold in Region A

$$\hat{y} = b_0 + b_1(0) + b_2(0) = b_0$$



Categorical Independent Variables

More Complex Categorical Variables:

- **Example Continued:**
  - Observations corresponding to Region B -> $x_1 = 1, x_2 = 0,$
  - Estimated number of units sold in Region B:

$$\hat{y} = b_0 + b_1(1) + b_2(0) = b_0 + b_1$$

  - Observations corresponding to Region C -> $x_1 = 0, x_2 = 1,$
  - Estimated number of units sold in Region C:

$$\hat{y} = b_0 + b_1(0) + b_2(1) = b_0 + b_2$$

Modeling Nonlinear Relationships

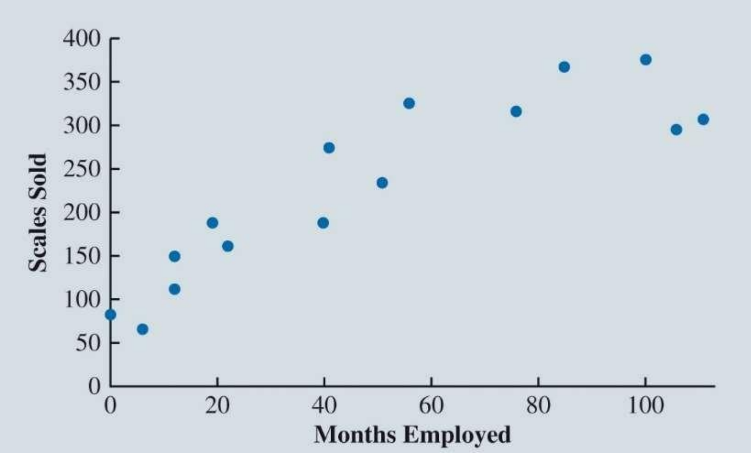
Quadratic Regression Models

Piecewise Linear Regression Models

Interaction Between Independent Variables

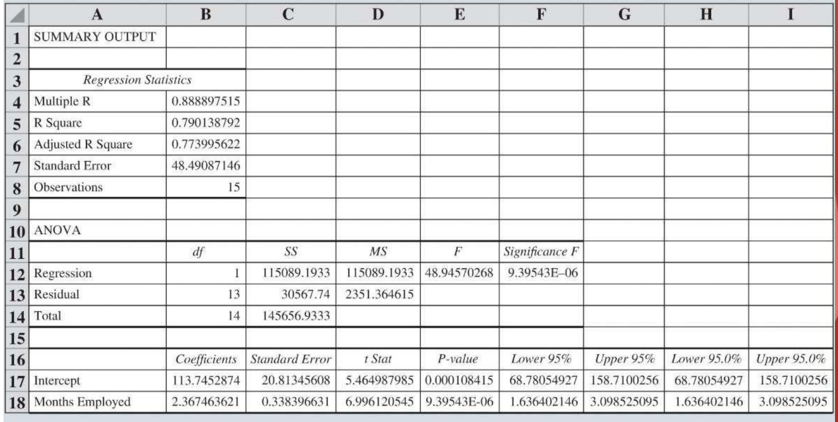
Modeling Nonlinear Relationships

Figure 7.25: Scatter Chart for the Reynolds Example



Modeling Nonlinear Relationships

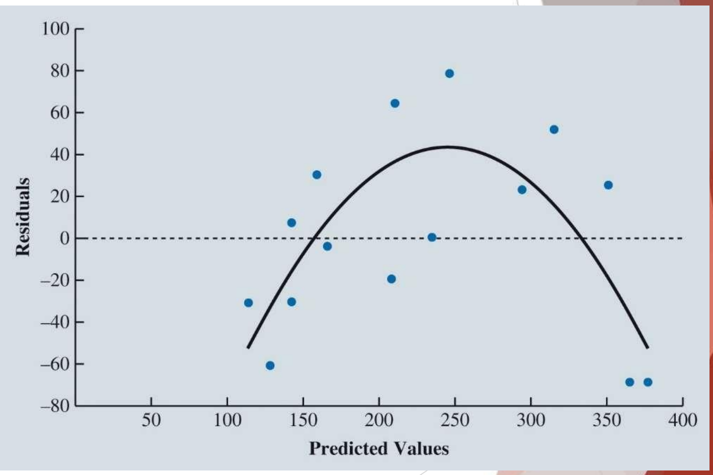
Figure 7.26: Excel Regression Output for the Reynolds Example



	A	B	C	D	E	F	G	H	I
1	SUMMARY OUTPUT								
2									
3	*Regression Statistics*								
4	Multiple R	0.888897515							
5	R Square	0.790138792							
6	Adjusted R Square	0.773995622							
7	Standard Error	48.49087146							
8	Observations	15							
9									
10	ANOVA								
11		*df*	*SS*	*MS*	*F*	*Significance F*			
12	Regression	1	115089.1933	115089.1933	48.94570268	9.39543E–06			
13	Residual	13	30567.74	2351.364615					
14	Total	14	145656.9333						
15									
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
17	Intercept	113.7452874	20.81345608	5.464987985	0.000108415	68.78054927	158.7100256	68.78054927	158.7100256
18	Months Employed	2.367463621	0.338396631	6.996120545	9.39543E-06	1.636402146	3.098525095	1.636402146	3.098525095

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.**

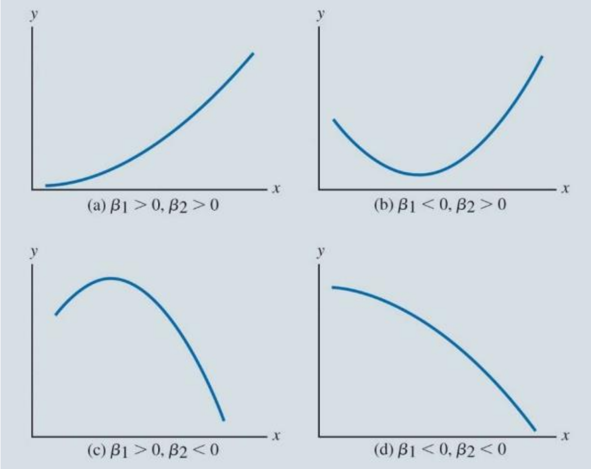
$$\hat{y} = b_0 + b_1 x_1 + b_2 x_1^2$$

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

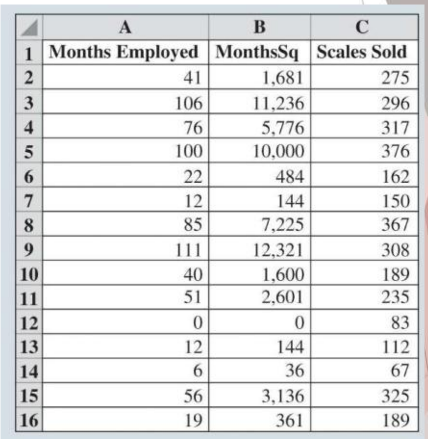
Modeling Nonlinear Relationships

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



Modeling Nonlinear Relationships

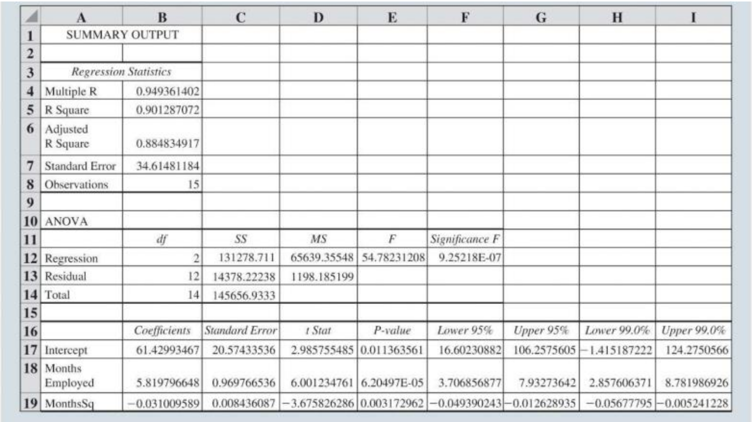
Figure 7.29: Excel Data for the Reynolds Quadratic Regression Model



	A	B	C
1	Months Employed	MonthsSq	Scales Sold
2	41	1,681	275
3	106	11,236	296
4	76	5,776	317
5	100	10,000	376
6	22	484	162
7	12	144	150
8	85	7,225	367
9	111	12,321	308
10	40	1,600	189
11	51	2,601	235
12	0	0	83
13	12	144	112
14	6	36	67
15	56	3,136	325
16	19	361	189

Modeling Nonlinear Relationships

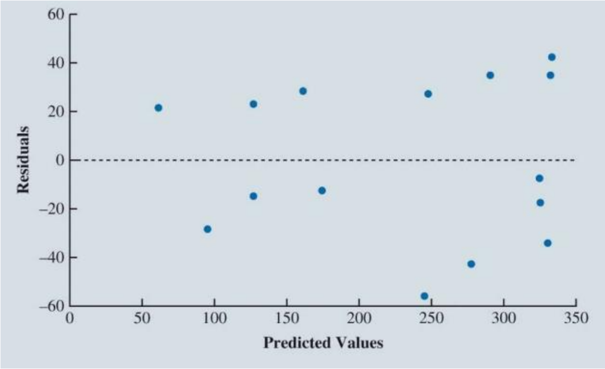
Figure 7.30: Excel Output for the Reynolds Quadratic Regression Model



	A	B	C	D	E	F	G	H	I
1	SUMMARY OUTPUT								
2									
3	*Regression Statistics*								
4	Multiple R	0.949361402							
5	R Square	0.901287072							
6	Adjusted R Square	0.884834917							
7	Standard Error	34.61481184							
8	Observations	15							
9									
10	ANOVA								
11		*df*	*SS*	*MS*	*F*	*Significance F*			
12	Regression	2	131278.711	65639.35548	54.78231208	9.25218E-07			
13	Residual	12	14378.22238	1198.185199					
14	Total	14	145656.9333						
15									
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 99.0%*	*Upper 99.0%*
17	Intercept	61.42993467	20.57433536	2.985755485	0.011363561	16.60230882	106.2575605	−1.415187222	124.2750566
18	Months Employed	5.819796648	0.969766536	6.001234761	6.20497E-05	3.706856877	7.93273642	2.857606371	8.781986926
19	MonthsSq	−0.031009589	0.008436087	−3.675826286	0.003172962	−0.049390243	−0.012628935	−0.05677795	−0.005241228

Modeling Nonlinear Relationships

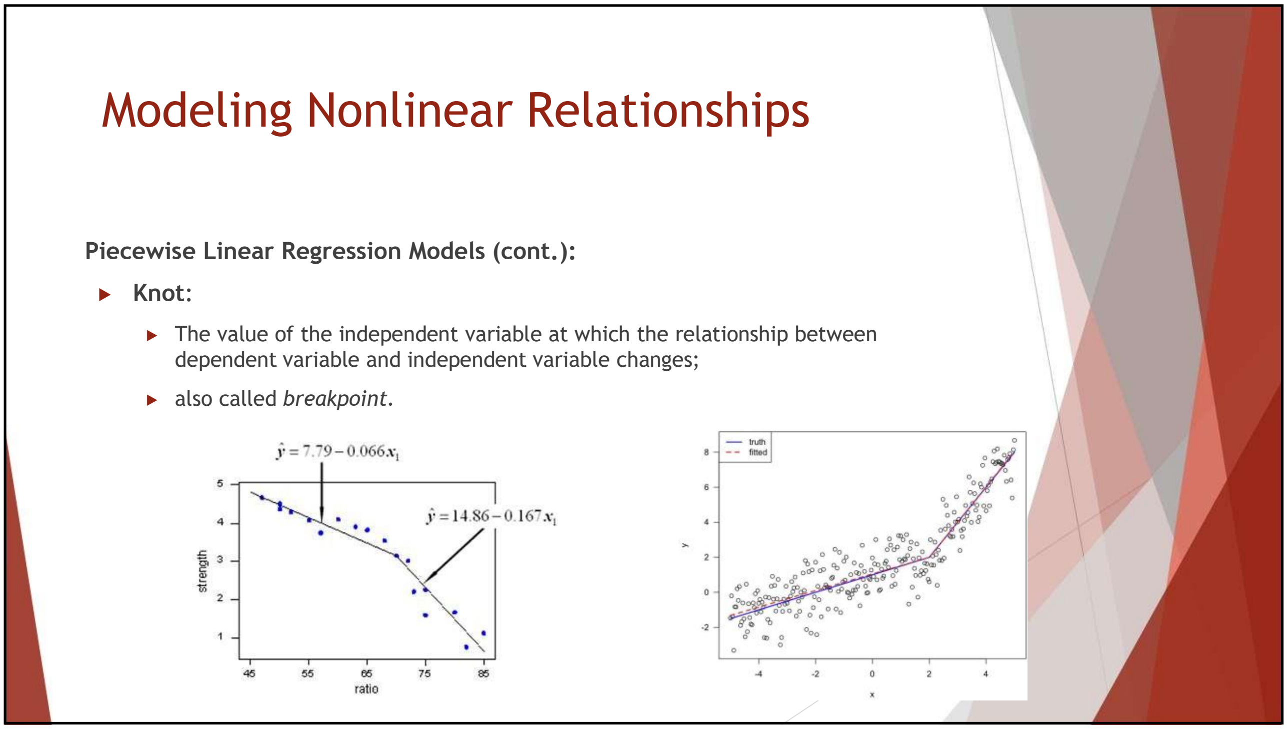
Figure 7.31: Scatter Chart of the Residuals and Predicted Values of the Dependent Variable for the Reynolds Quadratic Regression Model



Modeling Nonlinear Relationships

Piecewise Linear Regression Models:

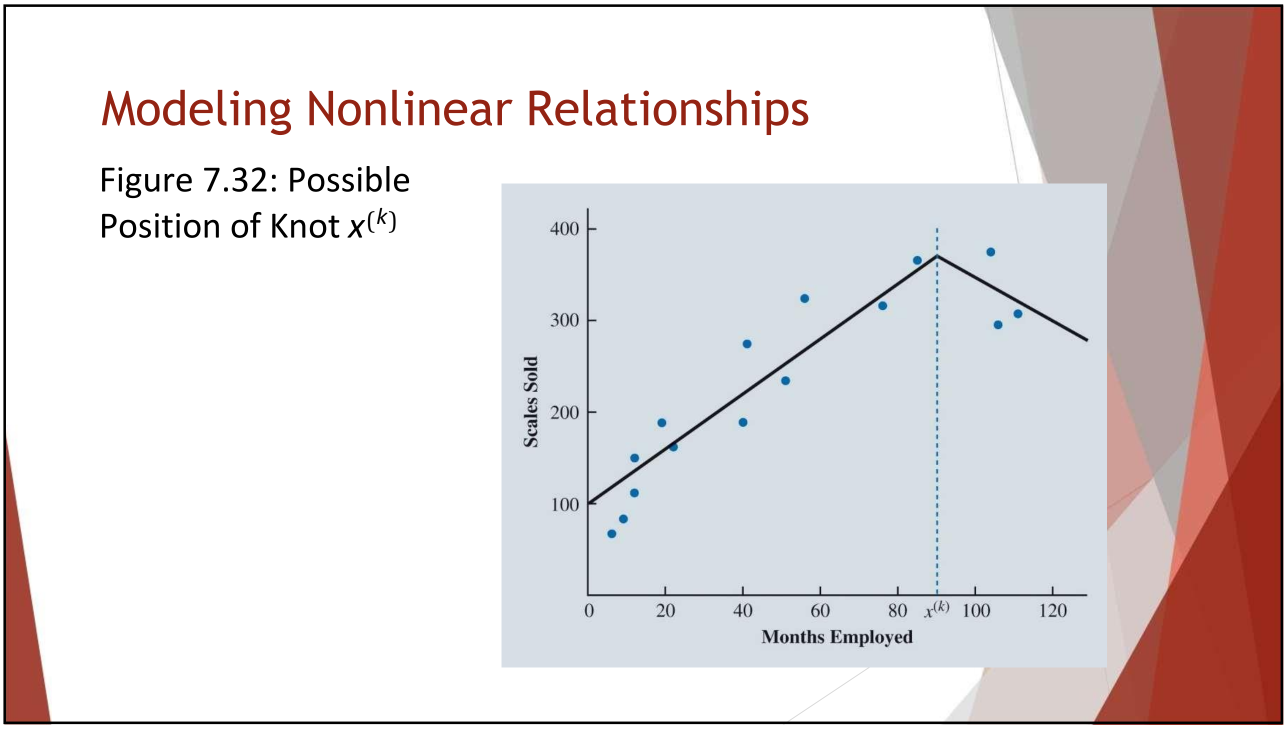
- For the Reynolds data, as an alternative to a quadratic regression model:
  - Recognize that up to a certain point of Months Employed
    - the relationship between Months Employed and Sales appears to be positive and linear.
  - After this point,
    - the relationship between Months Employed and 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:

$$x_k = \begin{cases} 0 \text{ if } x_1 \leq x^{(k)} \\ 1 \text{ if } x_1 > x^{(k)} \end{cases}$$

x_1 = Months.

$x^{(k)}$ = value of the knot (90 months for the Reynolds example).

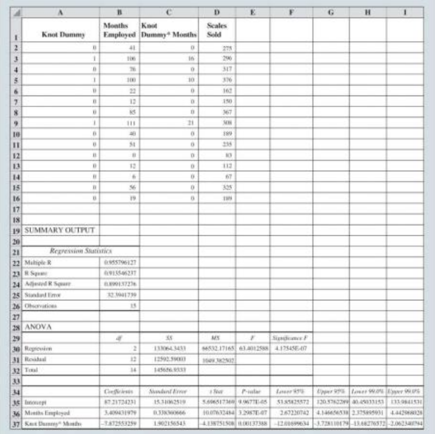
x_k = the knot dummy variable.

- Then fit the following estimated regression equation:

$$\hat{y} = b_0 + b_1 x_1 + b_2 (x_1 - x^{(k)}) x_k$$

Modeling Nonlinear Relationships

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



	A	B	C	D	E	F	G	H	I
1	Knot Dummy	Months Employed	Knot Dummy* Months	Scales Sold					
2	0	41	0	275					
3	1	106	16	296					
4	0	76	0	317					
5	1	100	10	376					
6	0	22	0	162					
7	0	12	0	150					
8	0	85	0	367					
9	1	111	21	308					
10	0	40	0	189					
11	0	51	0	235					
12	0	0	0	83					
13	0	12	0	112					
14	0	6	0	67					
15	0	56	0	325					
16	0	19	0	189					
17									
18									
19	SUMMARY OUTPUT								
20									
21	*Regression Statistics*								
22	Multiple R	0.955796127							
23	R Square	0.913546237							
24	Adjusted R Square	0.899137276							
25	Standard Error	32.3941739							
26	Observations	15							
27									
28	ANOVA								
29		*df*	*SS*	*MS*	*F*	*Significance F*			
30	Regression	2	133064.3433	66532.17165	63.4012588	4.17545E-07			
31	Residual	12	12592.59003	1049.382502					
32	Total	14	145656.9333						
33									
34		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 99.0%*	*Upper 99.0%*
35	Intercept	87.21724231	15.31062519	5.696517369	9.9677E-05	53.85825572	120.5762289	40.45033153	133.9841531
36	Months Employed	3.409431979	0.338360666	10.07632484	3.2987E-07	2.67220742	4.146656538	2.375895931	4.442968028
37	Knot Dummy* Months	-7.872553259	1.902156543	-4.138751508	0.00137388	-12.01699634	-3.728110179	-13.68276572	-2.062340794

Modeling Nonlinear Relationships

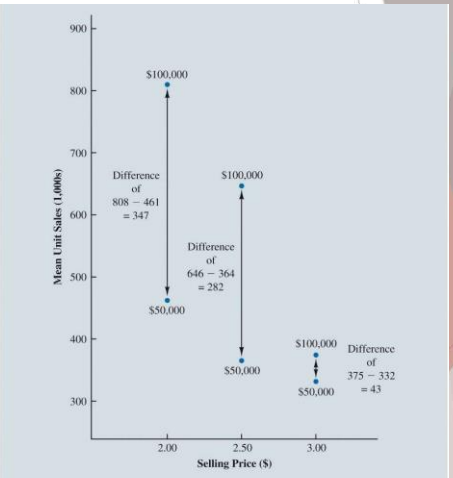
Interaction Between Independent Variables:

- **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:

$$\hat{y} = b_0 + b_1x_1 + b_2x_2 + b_3x_1x_2$$

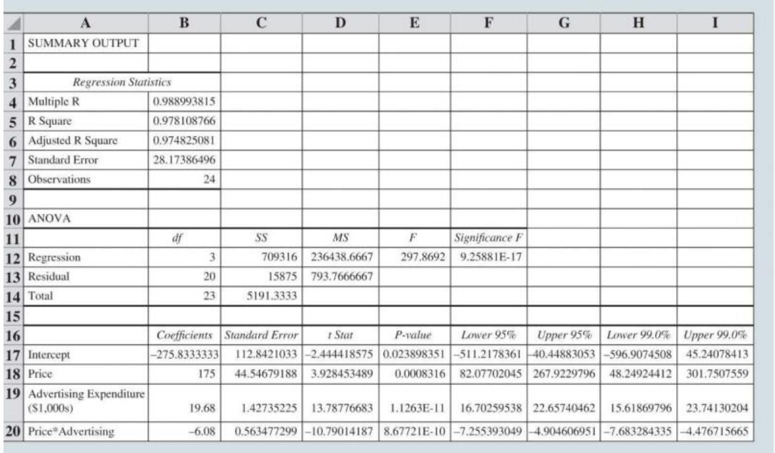
Modeling Nonlinear Relationships

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



	A	B	C	D	E	F	G	H	I
1	SUMMARY OUTPUT								
2									
3	*Regression Statistics*								
4	Multiple R	0.988993815							
5	R Square	0.978108766							
6	Adjusted R Square	0.974825081							
7	Standard Error	28.17386496							
8	Observations	24							
9									
10	ANOVA								
11		*df*	*SS*	*MS*	*F*	*Significance F*			
12	Regression	3	709316	236438.6667	297.8692	9.25881E-17			
13	Residual	20	15875	793.7666667					
14	Total	23	5191.3333						
15									
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 99.0%*	*Upper 99.0%*
17	Intercept	–275.8333333	112.8421033	–2.444418575	0.023898351	–511.2178361	–40.44883053	–596.9074508	45.24078413
18	Price	175	44.54679188	3.928453489	0.0008316	82.07702045	267.9229796	48.24924412	301.7507559
19	Advertising Expenditure ($1,000s)	19.68	1.42735225	13.78776683	1.1263E-11	16.70259538	22.65740462	15.61869796	23.74130204
20	Price*Advertising	–6.08	0.563477299	–10.79014187	8.67721E-10	–7.255393049	–4.904606951	–7.683284335	–4.476715665