Chapter 15
Multiple Regression

- Multiple Regression Model
- Least Squares Method
- Multiple Coefficient of Determination
- Model Assumptions
- Testing for Significance
- Using the Estimated Regression Equation for Estimation and Prediction
- Categorical Independent Variables
- Residual Analysis

CENGAGE

Multiple Regression

- In this chapter, we continue our study of regression analysis by considering situations involving two or more independent variables.
- This subject area, called multiple regression analysis, enables us to consider more factors and thus obtain better estimates than are possible with simple linear regression.

Multiple Regression Model

The equation that describes how the dependent variable y is related to the independent variables $x_1, x_2, \ldots x_p$ and an error term is:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p + \varepsilon$$

where:

$\beta_0, \beta_1, \beta_2, \ldots, \beta_p$ are the parameters, and ε is a random variable called the error term.

Multiple Regression Equation

The equation that describes how the mean value of y is related to $x_1, x_2, \ldots x_p$ is:

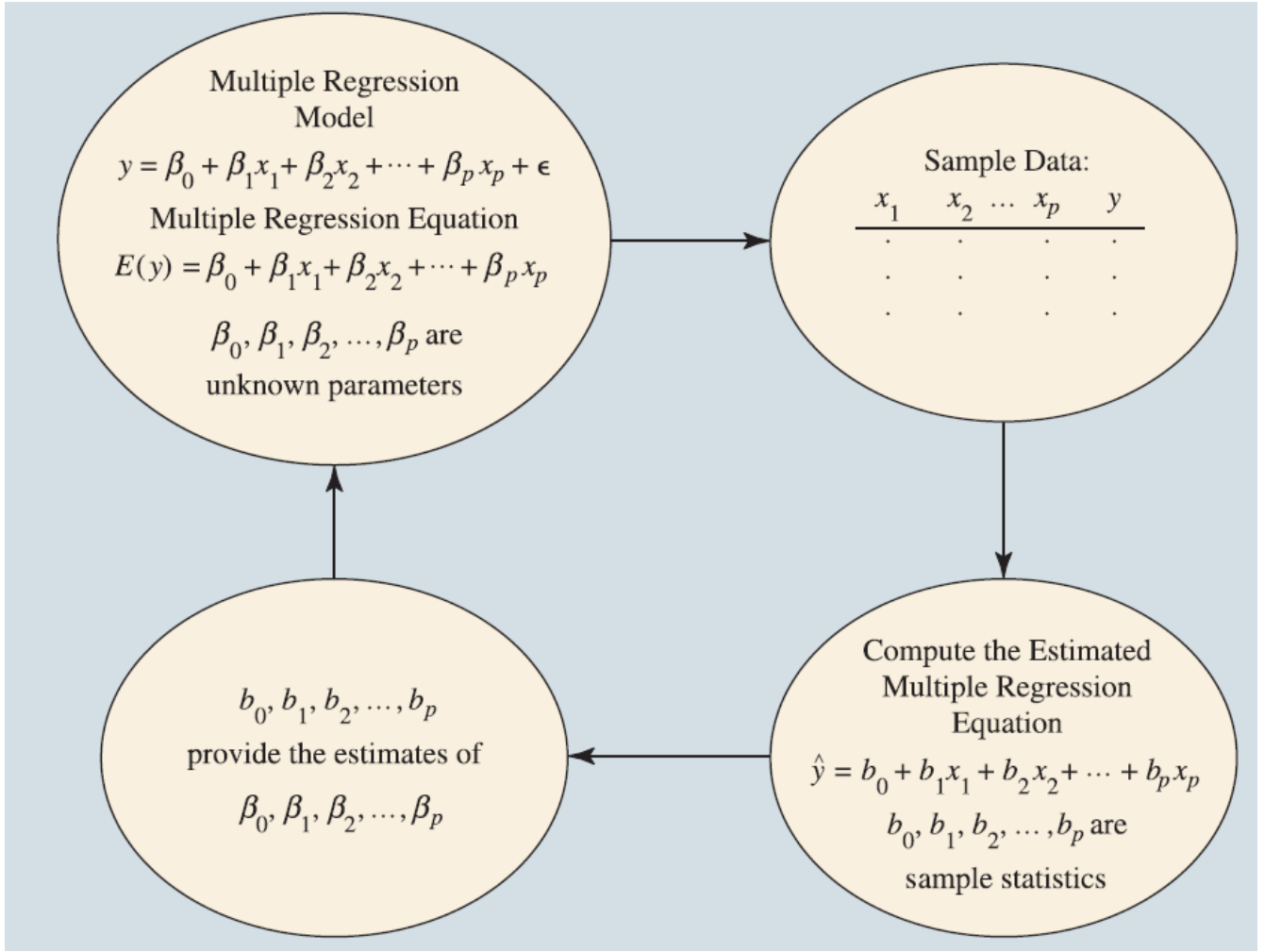
$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$$

Estimated Multiple Regression Equation

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

A simple random sample is used to compute sample statistics $b_0, b_1, b_2, \ldots, b_p$ that are used as the point estimators of the parameters $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$.

Estimation Process



Least Squares Method

Least Squares Criterion

$$\min \sum (y_i - \hat{y}_i)^2$$

Computation of Coefficient Values

The formulas for the regression coefficients b_0, b_1, b_2,…b_p involve the use of matrix algebra. We will rely on computer software packages to perform the calculations.

The emphasis will be on how to interpret the computer output rather than on how to make the multiple regression computations.

Multiple Regression Model (1 of 3)

Example: Butler Trucking Company

Managers at Butler Trucking Company want to develop better work schedules for their drivers. They believed that the total daily travel time would be closely related to the number of miles traveled in making the daily deliveries and also to the number of deliveries.

A simple random sample of 10 driving assignments was taken.

Multiple Regression Model (2 of 3)

Example: Butler Trucking Company

Driving Assignment	Miles traveled x_1	Deliveries x_2	y = Travel Time (hours)
1	100	4	9.3
2	50	3	4.8
3	100	4	8.9
4	100	2	6.5
5	50	2	4.2
6	80	2	6.2
7	75	3	7.4
8	65	4	6.0
9	90	3	7.6
10	90	2	6.1

Multiple Regression Model (3 of 3)

Example: Butler Trucking Company

Suppose we believe that total daily travel time (y) is related to the miles traveled (x_1) and the number of deliveries made (x_2) by the following regression model:

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

where

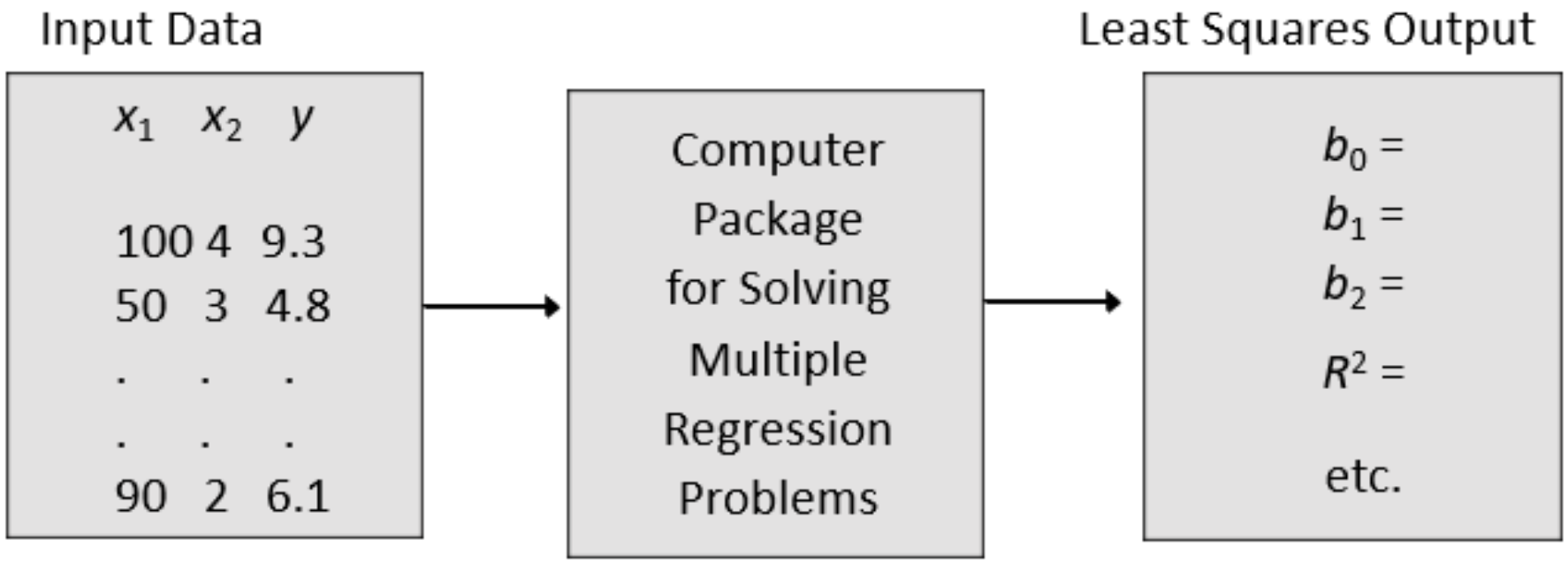
$y =$ Total travel time

$x_1 =$ Miles traveled

$x_2 =$ Deliveries made

Solving for the Estimates of β_0, β_1, β_2 (1 of 2)

Example: Butler Trucking Company



Solving for the Estimates of β_0, β_1, β_2 (2 of 2)

Example:
Butler Trucking Company

	A	B	C	D	E	F	G	H	I	J
1	**Assignment**	**Miles**	**Deliveries**	**Time**						
2	1	100	4	9.3						
3	2	50	3	4.8						
4	3	100	4	8.9						
5	4	100	2	6.5						
6	5	50	2	4.2						
7	6	80	2	6.2						
8	7	75	3	7.4						
9	8	65	4	6						
10	9	90	3	7.6						
11	10	90	2	6.1						
12										
13	SUMMARY OUTPUT									
14										
15	*Regression Statistics*									
16	Multiple R	0.9507								
17	R Square	0.9038								
18	Adjusted R Square	0.8763								
19	Standard Error	0.5731								
20	Observations	10								
21										
22	ANOVA									
23		*df*	*SS*	*MS*	*F*	*Significance F*				
24	Regression	2	21.6006	10.8003	32.8784	0.0003				
25	Residual	7	2.2994	0.3285						
26	Total	9	23.9							
27										
28		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 99.0%*	*Upper 99.0%*	
29	Intercept	-0.8687	0.9515	-0.9129	0.3916	-3.1188	1.3813	-4.1986	2.4612	
30	Miles	0.0611	0.0099	6.1824	0.0005	0.0378	0.0845	0.0265	0.0957	
31	Deliveries	0.9234	0.2211	4.1763	0.0042	0.4006	1.4463	0.1496	1.6972	
32										

CENGAGE

Estimated Regression Equation

Example: Butler Trucking Company

$$\hat{y} = -.8687 + .0611x_1 + .9234x_2$$

Interpreting the Coefficients (1 of 3)

In multiple regression analysis, we interpret each regression coefficient as follows:

b_i represents an estimate of the change in y corresponding to a 1-unit change in x_i when all other independent variables are held constant.

Interpreting the Coefficients (2 of 3)

Example: Butler Trucking Company

$$b_1 = .0611$$

.0611 is an estimate of the expected increase in travel time corresponding to an increase of one mile in the distance traveled when the number of deliveries is held constant.

Interpreting the Coefficients (3 of 3)

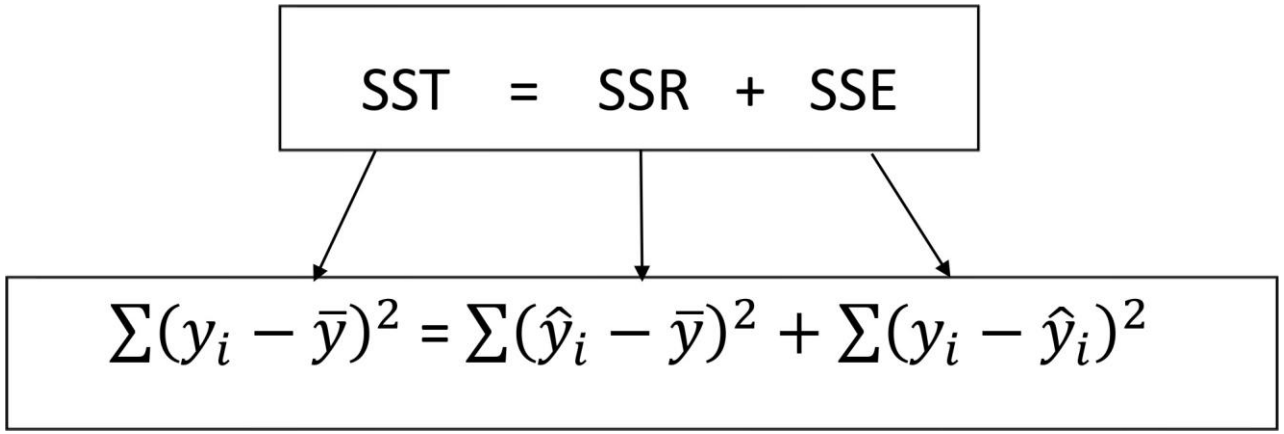
Example: Butler Trucking Company

$$b_2 = .9234$$

.9234 is an estimate of the expected increase in travel time corresponding to an increase of one delivery when the number of miles traveled is held constant.

Multiple Coefficient of Determination (1 of 3)

- Relationship Among SST, SSR, SSE



where:

SST = Total sum of squares

SSR = Sum of squares due to regression

SSE = Sum of squares due to error

Multiple Coefficient of Determination (2 of 3)

Example: Butler Trucking Company

ANOVA Output

22	ANOVA					
23		*df*	*SS*	*MS*	*F*	*Significance F*
24	Regression	2	21.6006	10.8003	32.8784	0.0003
25	Residual	7	2.2994	0.3285		
26	Total	9	23.9			

Multiple Coefficient of Determination (3 of 3)

Example: Butler Trucking Company

$$R^2 = \frac{SSR}{SST}$$

$$R^2 = \frac{21.6006}{23.9} = .9038$$

Adjusted Multiple Coefficient of Determination

Adding independent variables, even ones that are not statistically significant, causes the prediction errors to become smaller, thus reducing the sum of squares due to error, SSE.

Because $SSR = SST - SSE$, when SSE becomes smaller, SSR becomes larger, causing $R^2 =$ SSR/SST to increase.

The adjusted multiple coefficient of determination compensates for the number of independent variables in the model.

Adjusted Multiple Coefficient of Determination (2 of 2)

Example: Butler Trucking Company

$$R_a^2 = 1 - (1 - R^2)\frac{n-1}{n-p-1}$$

$$R_a^2 = 1 - (1 - .9038)\frac{10-1}{10-2-1} = .8763$$

Assumptions About the Error Term ε in the Multiple Regression Model

- The error term ε is a random variable with mean of zero.
- The variance of ε, denoted by σ^2, is the same for all values of the independent variables.
- The values of ε are independent.
- The error term ε is a normally distributed random variable reflecting the deviation between the *y* value and the expected value of *y* given by $\beta_0 + \beta_1 x_1 + \beta_2 x_2 + .. + \beta_p x_p$.

Testing for Significance

- In simple linear regression, the *F* and *t* tests provide the same conclusion.
- In multiple regression, the *F* and *t* tests have different purposes.

Testing for Significance: *F* Test (1 of 3)

- The *F* test is used to determine whether a significant relationship exists between the dependent variable and the set of all the independent variables.
- The *F* test is referred to as the test for overall significance.

Testing for Significance: *t* Test (1 of 3)

- If the *F* test shows an overall significance, the *t* test is used to determine whether each of the individual independent variables is significant.
- A separate *t* test is conducted for each of the independent variables in the model.
- We refer to each of these *t* tests as a test for individual significance.

Testing for Significance: *F* Test

Hypotheses

$H_0 : \beta_1 = \beta_2 = \ldots = \beta_p = 0$

H_a : One or more of the parameters is not equal to zero

Test Statistics

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

Rejection Rule

Reject H_0 if p-value $\leq \alpha$ or if $F \geq F_\alpha$,

Where F_α is based on an F distribution with p d.f. in the numerator and $n - p - 1$ d.f. in the denominator.

Testing for Significance: *F* Test (3 of 3)

Example: Butler Trucking Company

Hypotheses

$H_0 : \beta_1 = \beta_2 = \ldots = \beta_p = 0$

H_a: One or more of the parameters is not equal to zero

Rejection Rule

For $\alpha = .01$ and d.f. $= 2, 7$; $F_{.01} = 9.55$

Reject H_0 if p-value $\leq .01$ or $F \geq 9.55$

F Test for Overall Significance (1 of 2)

Example: Butler Trucking Company

ANOVA Output

ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	21.6006	10.8003	32.8784	0.0003
Residual	7	2.2994	0.3285		
Total	9	23.9			

p-value used to test for overall significance

F Test for Overall Significance (2 of 2)

Example: Butler Trucking Company

Test Statistics

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

$$\frac{10.8003}{.3285} = 32.9$$

Conclusion

p-value $< .01$, so we can reject H_0.

(Also, $F = 32.9 > 9.55$).

Testing for Significance: *t* Test (2 of 3)

Hypotheses $H_0 : \beta_i = 0$

$H_a : \beta_i \neq 0$

Test statistics $t = \dfrac{b_i}{s_{b_i}}$

Rejection Rule Reject H_0 if p-value $\leq \alpha$ or if $t \leq -t_{\alpha/2}$ or $t \geq t_{\alpha/2}$

Where $t_{\alpha/2}$ is based on a t distribution with $n - p - 1$ degrees of freedom.

Testing for Significance: *t* Test (3 of 3)

Example: Butler Trucking Company

Hypotheses

$$H_0 : \beta_i = 0$$
$$H_a : \beta_i \neq 0$$

Rejection Rule

For $\alpha = .01$ and d.f. $= 7$, $t_{.005} = 3.499$

Reject H_0 if p-value $\leq .01$, or if $t \leq -3.499$ or $t \geq 3.499$.

t Test for Significance of Individual Parameters (1 of 2)

Example: Butler Trucking Company

Regression Equation Output

27					
28		*Coefficients*	*Standard Error*	*t Stat*	*P-value*
29	Intercept	-0.8687	0.9515	-0.9129	0.3916
30	Miles	0.0611	0.0099	6.1824	0.0005
31	Deliveries	0.9234	0.2211	4.1763	0.0042
32					
33					
34					

The *p*-value in cell E30 is used to test for the individual significance of Miles.

The *p*-value in cell E31 is used to test for the individual significance of Deliveries.

t Test for Significance of Individual Parameters (2 of 2)

Test Statistics

$$t = \frac{b_1}{s_{b_1}} = \frac{.0611}{.0099} = 6.1717$$

$$t = \frac{b_2}{s_{b_2}} = \frac{.9234}{.2211} = 4.1764$$

Conclusions

Reject both $H_0 : \beta_1 = 0$ and $H_0 : \beta_2 = 0$. Both independent variables are significant.

Testing for Significance: Multicollinearity (1 of 2)

- The term multicollinearity refers to the correlation among the independent variables.
- When the independent variables are highly correlated $\left(\text{say, } |r| > .7\right)$,

  it is not possible to determine the separate effect of any particular independent variable on the dependent variable.

Testing for Significance: Multicollinearity (2 of 2)

- If the estimated regression equation is to be used only for predictive purposes, multicollinearity is usually not a serious problem.
- Every attempt should be made to avoid including independent variables that are highly correlated.

Using the Estimated Regression Equation for Estimation and Prediction (1 of 2)

- The procedures for estimating the mean value of y and predicting an individual value of y in multiple regression are similar to those in simple regression.
- We substitute the given values of $x_1, x_2, \ldots, x_p$ into the estimated regression equation and use the corresponding value of y as the point estimate.

Using the Estimated Regression Equation for Estimation and Prediction (2 of 2)

- The formulas required to develop interval estimates for the mean value of $\hat{y}$ and for an individual value of y are beyond the scope of the textbook.
- Software packages for multiple regression will often provide these interval estimates.

Categorical Independent Variables (1 of 6)

- In many situations we must work with categorical independent variables such as gender (male, female), method of payment (cash, check, credit card), etc.
- For example, x_2 might represent gender where $x_2 = 0$ indicates male and $x_2 = 1$ indicates female.
- In this case, x_2 is called a dummy or indicator variable.

Categorical Independent Variables (2 of 6)

Example: Johnson Filtration, Inc.

Managers of Johnson Filtration, Inc. want to predict the repair time necessary for processing its maintenance requests. Repair time is believed to be related to two factors, the number of months since last service and the type of repair problem (mechanical or electrical). Data for a sample of 10 service calls are reported in the table below.

Categorical Independent Variables (3 of 6)

Example:

Johnson Filtration, Inc.

0 = Mechanical

1 = Electrical

Service Call	Months since last service	Type of repair	Repair time (hours)
1	2	1	2.9
2	6	0	3.0
3	8	1	4.8
4	3	0	1.8
5	2	1	2.9
6	7	1	4.9
7	9	0	4.2
8	8	0	4.8
9	4	1	4.4
10	6	1	4.5

Categorical Independent Variables (4 of 6)

Example: Johnson Filtration, Inc.

Regression Equation Output

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

where:

$\hat{y}$ = Repair time in hours

x_1 = Number of months since last maintenance service

x_2 = 0 if type of repair is mechanical, 1 if the type of repair is electrical

(x_2 is a dummy variable)

Categorical Independent Variables (5 of 6)

Example: Johnson Filtration, Inc.

ANOVA Output

10	ANOVA					
11		*df*	*SS*	*MS*	*F*	*Significance F*
12	Regression	2	9.0009	4.5005	21.357	0.0010
13	Residual	7	1.4751	0.2107		
14	Total	9	10.476			

$$R^2 = 9.0009/10.476 = .8952$$

$$R_a^2 = 1-(1-.8952)\frac{10-1}{10-2-1} = .8190$$

This indicates estimated regression equation does a good job of explaining the variability in repair times.

Categorical Independent Variables (6 of 6)

Example: Johnson Filtration, Inc.

16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*
17	Intercept	0.9305	0.4670	1.9926	0.0866
18	Months	0.3876	0.0626	6.1954	0.0004
19	Type	1.2627	0.3141	4.0197	0.0051

Conclusion:

Both months since service and the type of repair are significant.

More Complex Categorical Variables (1 of 2)

- If a categorical variable has k levels, $k - 1$ dummy variables are required, with each dummy variable being coded as 0 or 1.
- For example, a variable with levels A, B, and C could be represented by x_1 and x_2 values of (0, 0) for A, (1, 0) for B, and (0, 1) for C.
- Care must be taken in defining and interpreting the dummy variables.

More Complex Categorical Variables (2 of 2)

Example: Sales regions

A variable indicating sales regions could be represented by x_1 and x_2 values as follows:

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

Residual Analysis

- For simple linear regression, the residual plot against $\hat{y}$ and the residual plot against x provide the same information.
- In multiple regression analysis it is preferable to use the residual plot against $\hat{y}$ to determine if the model assumptions are satisfied.

Standardized Residual Plot Against $\hat{y}$ (1 of 3)

- Standardized residuals are frequently used in residual plots for purposes of:
  - Identifying outliers (typically, standardized residuals < -2 or $> +2$)
  - Providing insight about the assumption that the error term ε has a normal distribution
- The computation of the standardized residuals in multiple regression analysis is too complex to be done by hand.
- Excel's Regression tool can be used.

Standardized Residual Plot Against $\hat{y}$ (2 of 3)

Example: Butler Trucking Company: Residual output

	A	B	C	D	E	F	G
37	*Observation*	*Predicted Time*	*Residuals*	*Standard Residuals*		*Predicted Time*	*Standard Residuals*
38	1	8.9385	0.3615	0.7153		8.9385	0.7153
39	2	4.9583	-0.1583	-0.3132		4.9583	-0.3132
40	3	8.9385	-0.0385	-0.0761		8.9385	-0.0761
41	4	7.0916	-0.5916	-1.1704		7.0916	-1.1704
42	5	4.0349	0.1651	0.3267		4.0349	0.3267
43	6	5.8689	0.3311	0.6550		5.8689	0.6550
44	7	6.4867	0.9133	1.8069		6.4867	1.8069
45	8	6.7987	-0.7987	-1.5802		6.7987	-1.5802
46	9	7.4037	0.1963	0.3884		7.4037	0.3884
47	10	6.4803	-0.3803	-0.7523		6.4803	-0.7523

Standardized Residual Plot Against $\hat{y}$ (3 of 3)

Example: Butler Trucking Company