Linear Regression

Outline

- Simple Linear Regression Model
- Least Squares Method
- Assessing the Fit of the Simple Linear Regression Model
- The Multiple Linear Regression Model
- Inference and Linear Regression
- Categorical Independent Variables
- Modeling Nonlinear Relationships
- Model Fitting
- Big Data and Linear Regression
- Prediction with Linear Regression
- Summary

Learning Objectives (1 of 2)

After completing this chapter, you will be able to:

- Construct an estimated simple linear regression model that estimates how a dependent variable is related to an independent variable.
- Construct an estimated multiple linear regression model that estimates how a dependent variable is related to multiple independent variables.
- Compute and interpret the estimated coefficient of determination for a linear regression model.
- Assess whether the conditions necessary for valid inference in a least squares linear regression model are satisfied and test hypotheses about the parameters.
- Test hypotheses about the parameters of a linear regression model and interpret the results of these hypotheses tests.

Learning Objectives (2 of 2)

- Compute and interpret confidence intervals for the parameters of a linear regression model.
- Use dummy variables to incorporate categorical independent variables in a linear regression model and interpret the associated estimated regression parameters.
- Use a quadratic regression model, a piecewise linear regression model, and interaction between independent variables to account for curvilinear relationships between independent variables and the dependent variable in a regression model and interpret the estimated parameters.
- Use an estimated linear regression model to predict the value of the dependent variable given values of the independent variables.

Introduction

Managerial decisions are often based on the relationship between variables.

Regression analysis is a statistical procedure that uses data to develop an equation showing how the variables are related.

- **Dependent** (or *response*) **variable** is the variable being predicted.
- **Independent** (or *predictor*) **variables** (or *features*) are variables used to predict the value of the dependent variable.
- **Simple linear regression** is a form of regression analysis in which a ("simple") single independent variable, x is used to develop a "linear" relationship (straight line) with the dependent variable, y.
- **Multiple linear regression** is a more general form of regression analysis involving two or more independent variables.

Simple Linear Regression Model

The **simple linear regression model** is an equation that describes how the dependent variable y is related to the independent variable x and error term ε.

$$\boldsymbol{y = \beta_0 + \beta_1 x + \varepsilon}$$

where

y is the dependent variable.

x is the independent variable.

β_0 and β_1 are referred to as the **population parameters**.

ε is the error term. It accounts for the variability in y that cannot be explained by the linear relationship between x and y.

Estimated Simple Linear Regression Equation

The **estimated simple linear regression equation** is described as follows.

$$\hat{y} = b_0 + b_1 x$$

Where $\hat{y}$ is the **point estimator** of $E(y|x)$, the mean of y for a given x.

b_0 is the point estimator of β_0 and the y-intercept of the regression.

The y-intercept b_0 is the estimated value of the dependent variable y when the independent variable x is equal to 0.

b_1 is the point estimator of β_1 and the slope of the regression.

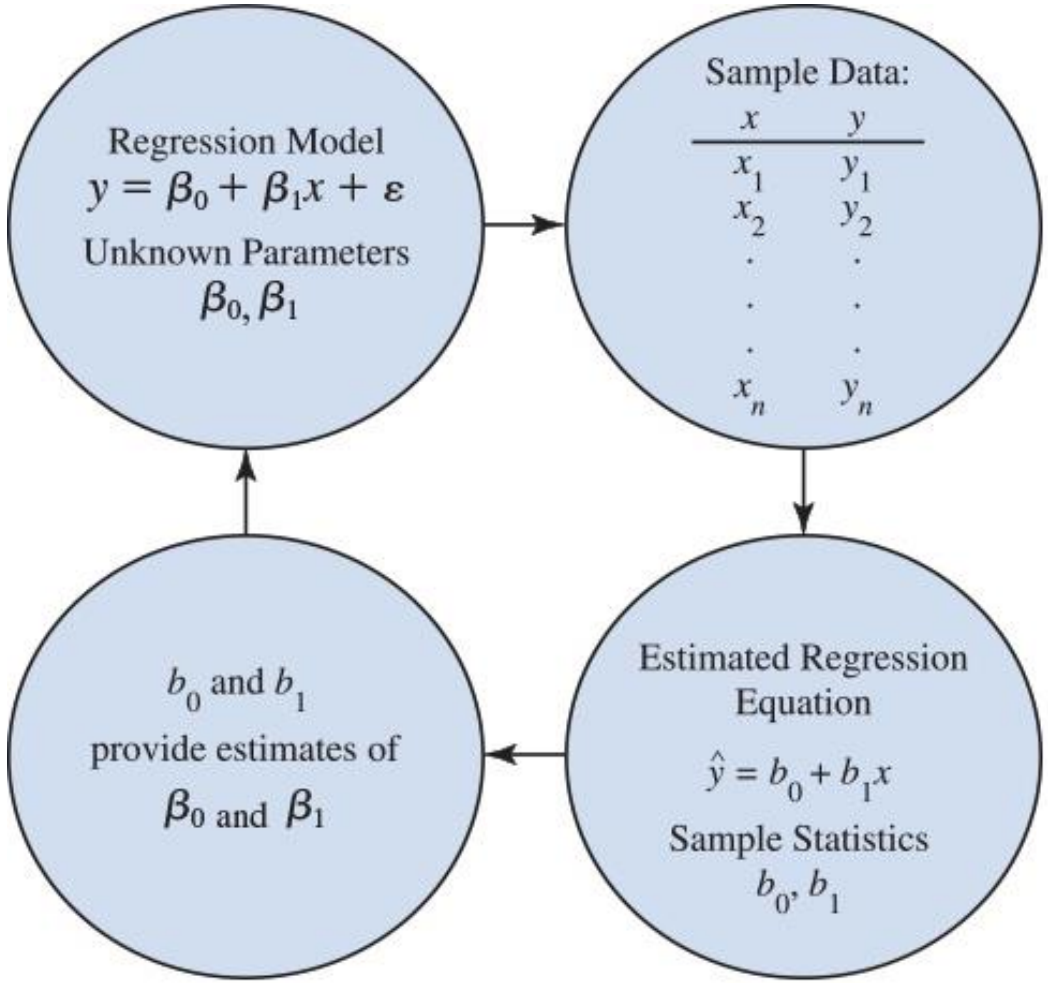
The slope b_1 is the estimated change in the value of the dependent variable y that is associated with a one unit increase in the independent variable x.

The Estimation Process in Simple Linear Regression

The estimation of b_0 and b_1 is a statistical process much like the estimation of the population mean μ described in Chapter 7.

β_0 and β_1 are the unknown parameters of interest, and b_0 and b_1 are the sample statistics used to estimate the parameters.

The flow chart to the right provides a summary of the estimation process for simple linear regression



Least Squares Method

The **least squares method** is a procedure for using sample data to find the estimated linear regression equation (see notes for b_0 and b_1 equations.)

$$\min \sum e_i^2 = \min \sum (y_i - \hat{y}_i)^2 = \min \sum (y_i - b_0 - b_1 x_i)^2$$

Where

$e_i = y_i - \hat{y}_i$ is referred to as the ith **residual**: the error made in estimating the value of the dependent variable for the ith observation.

x_i and y_i are the values of independent and dependent variables for the ith observation.

$\hat{y}_i$ is the predicted value of the dependent variable for the ith observation.

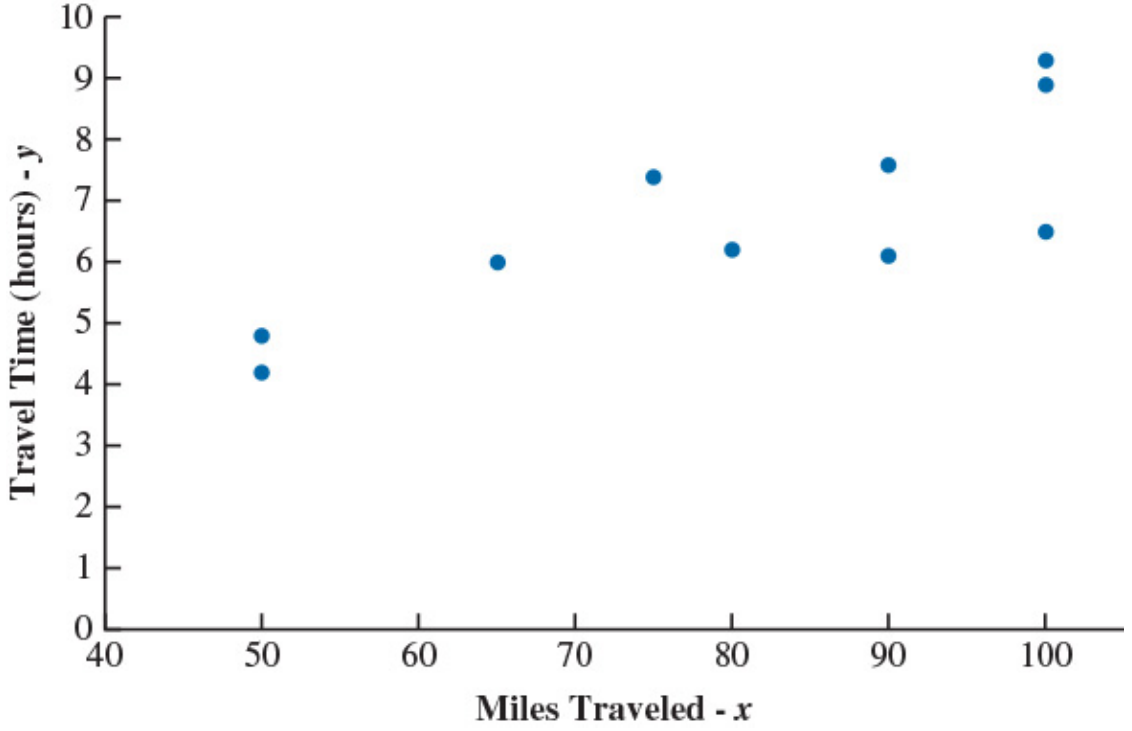
n is the total number of observations.

The Butler Trucking Company Example

We use a sample of 10 randomly selected driving assignments made by the Butler Trucking Company to build a scatter chart depicting the relationship between the travel time (in hours) and the miles traveled.

Data file: *butler.xlsx*

Because the scatter diagram shows a positive linear relationship, we choose the simple linear regression model to represent the relationship between travel time (y) and miles traveled (x.)



Regression Equation for Butler Trucking Co.

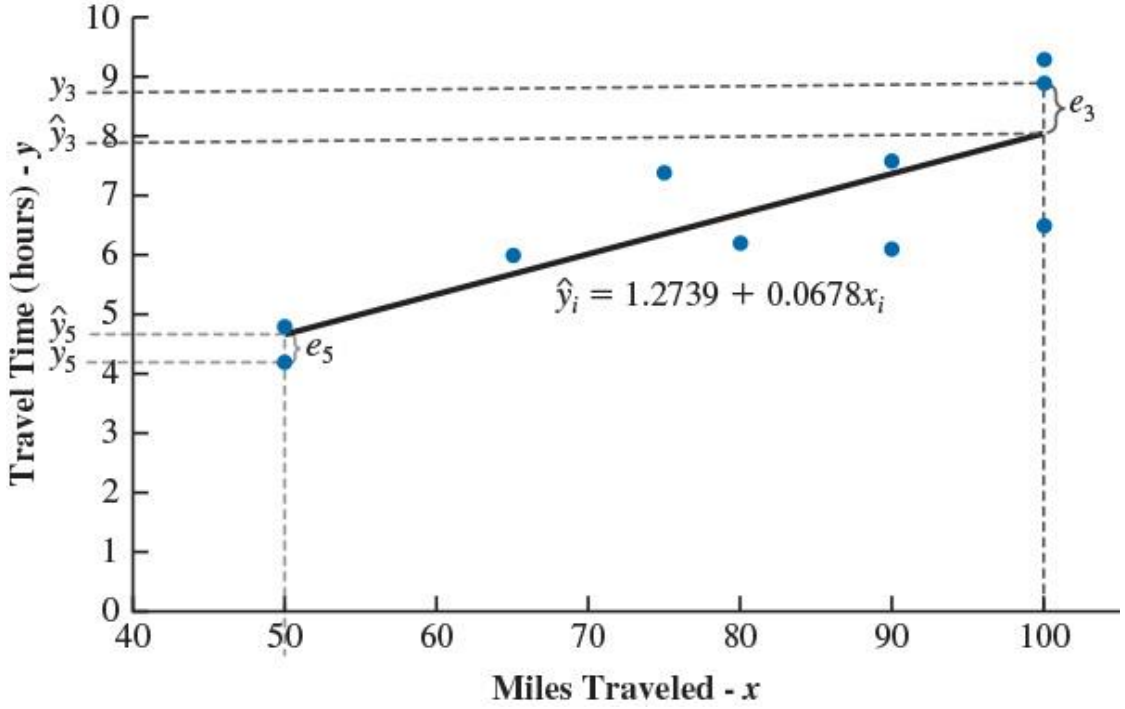
Computer software produces the following simple linear regression equation.

$$\hat{y} = 1.2739 + 0.0678x$$

The slope and intercept of the regression are $b_1 = 0.0678$ and $b_0 = 1.2739$.

Thus, we estimate that if the length of a driving assignment were 1 mile longer, the mean travel time would be 0.0678 hours or ~4 minutes longer.

Also, if the length of a driving assignment were 0 miles, the mean travel time would be 1.2739 hours or ~76 minutes. See notes for Excel.



Experimental Region and Extrapolation

The regression model is valid only over the **experimental region**, defined as the range of values of the independent variables in the data used to estimate the model.

Extrapolation, the prediction of the value of the dependent variable outside the experimental region, is risky and should be avoided unless we have empirical evidence dictating otherwise.

The experimental region for the Butler Trucking data is from 50 to 100 miles.

- Any prediction made outside the travel time for a driving distance less than 50 miles or greater than 100 miles is not a reliable estimate.
- Thus, for this model the estimate of β_0 is meaningless.

Estimating Travel Time for the Butler Trucking Co.

We can use the estimated model for the Butler Trucking Company example, and the known values for miles traveled for a driving assignment to estimate the mean travel time in hours.

For example, the first driving assignment in the data set has a value for miles traveled of $x = 100$, and a value for travel time of $y = 9.3$ hours.

The mean travel time for this driving assignment is estimated to be

$$\hat{y}_i = 1.2739 + 0.0678(100) = 8.0539 \text{ hours}$$

The resulting residual of the estimate is

$$e_i = y_i - \hat{y}_i = 9.3 - 8.0539 = 1.2461 \text{ hours}$$

The next slide shows the calculations for the 10 observations in the data set.

Predicted Travel Time and Residuals

Driving Assignment i	x_i = Miles Traveled	y_i = Travel Time (hours)	$\hat{y}_i = b_0 + b_1 x_i$	$e_i = y_i - \hat{y}_i$	$e_i^2 = (y_i - \hat{y}_i)^2$
1	100	9.3	8.0539	1.2461	1.5528
2	50	4.8	4.6639	0.1361	0.0185
3	100	8.9	8.0539	0.8461	0.7159
4	100	6.5	8.0539	-1.5539	2.4146
5	50	4.2	4.6639	-0.4639	0.2152
6	80	6.2	6.6979	-0.4979	0.2479
7	75	7.4	6.3589	1.0411	1.0839
8	65	6.0	5.6809	0.3191	0.1018
9	90	7.6	7.3759	0.2241	0.0502
10	90	6.1	7.3759	-1.2759	1.6279
		$\sum y_i = 67.0$	$\sum \hat{y}_i = 67.0000$	$\sum e_i = 0.0000$	$\sum e_i^2 = 8.0288$

The Sums of Squares

The value of the **sum of squares due to error** (**SSE**) is a measure of the error that results from using the $\hat{y}_i$ values to predict the y_i values.

$$\boldsymbol{SSE = \sum(y_i - \hat{y}_i)^2}$$

The value of the **total sum of squares** (**SST**) is a measure of the error that results from using the sample mean $\bar{y}$ to predict the y_i values.

$$\boldsymbol{SST = \sum(y_i - \bar{y})^2}$$

The value of the **sum of squares due to regression** (**SSR**) is a measure of how much the $\hat{y}_i$ values deviate from the sample mean $\bar{y}$.

$$\boldsymbol{SSR = \sum(\hat{y}_i - \bar{y})^2}$$

The relationship between these three sums of squares is $\boldsymbol{SST = SSR + SSE}$

Total Sum of Squares

Driving Assignment *i*	x_i = Miles Traveled	y_i = Travel Time (hours)	$\hat{y}_i - \bar{y}$	$(\hat{y}_i - \bar{y})^2$
1	100	9.3	2.6	6.76
2	50	4.8	-1.9	3.61
3	100	8.9	2.2	4.84
4	100	6.5	-0.2	0.04
5	50	4.2	-2.5	6.25
6	80	6.2	-0.5	0.25
7	75	7.4	0.7	0.49
8	65	6.0	-0.7	0.49
9	90	7.6	0.9	0.81
10	90	6.1	-0.6	0.36
		$\sum y_i = 67.0$	$\sum \hat{y}_i - \bar{y} = 0$	$SST = 23.90$

Coefficient of Determination

The ratio SSR/SST is called the **coefficient of determination**, denoted by r^2.

$$\boldsymbol{r^2 = \frac{SSR}{SST}}$$

The coefficient of determination can only assume values between 0 and 1 and is used to evaluate the goodness of fit for the estimated regression equation.

A perfect fit exists when y_i is identical to $\hat{y}_i$ for every observation i so that all residuals $y_i - \hat{y}_i = 0$.

- In such case, $SSE = 0$, $SSR = SST$, and $r^2 = SSR/SST = 1$.

Poorer fits between y_i and $\hat{y}_i$ result in larger values of SSE and lower r^2 values.

- The poorest fit happens when $SSE = SST$, $SSR = 0$, and $r^2 = 0$.

Goodness of Fit

From our previous calculations for the sum of squares due to error, we already know that

$$SSE = \sum(y_i - \hat{y}_i)^2 = \sum e_i^2 = 8.0288$$

Similar calculations for the total sum of squares reveal that

$$SST = \sum(y_i - \bar{y})^2 = 23.90$$

Because of the sum of squares relationship, we can write

$$r^2 = SSR/SST = 1 - SSE/SST = 1 - 8.0288/23.90 = 0.6641$$

Thus, we can conclude that 66.41% of the variability in the values of travel time can be explained by the linear relationship between the miles traveled and travel time. See notes for Excel instructions.

Multiple Linear Regression Model

The **multiple linear regression model** describes how the dependent variable y is related to the independent variables $x_1, x_2, \ldots, x_q$ and an error term ε.

$$\boldsymbol{y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_q x_q + \varepsilon}$$

where

$\beta_0, \beta_1, \beta_2, \ldots, \beta_q$ are the parameters of the model.

ε is the error term that accounts for the variability in y that cannot be explained by the linear effect of the q independent variables.

The coefficient β_j (with $j = 1 \ldots q$) represents the change in the mean value of y that corresponds to a one unit increase in the independent variable x_j, *holding the values of all other independent variables in the model constant*.

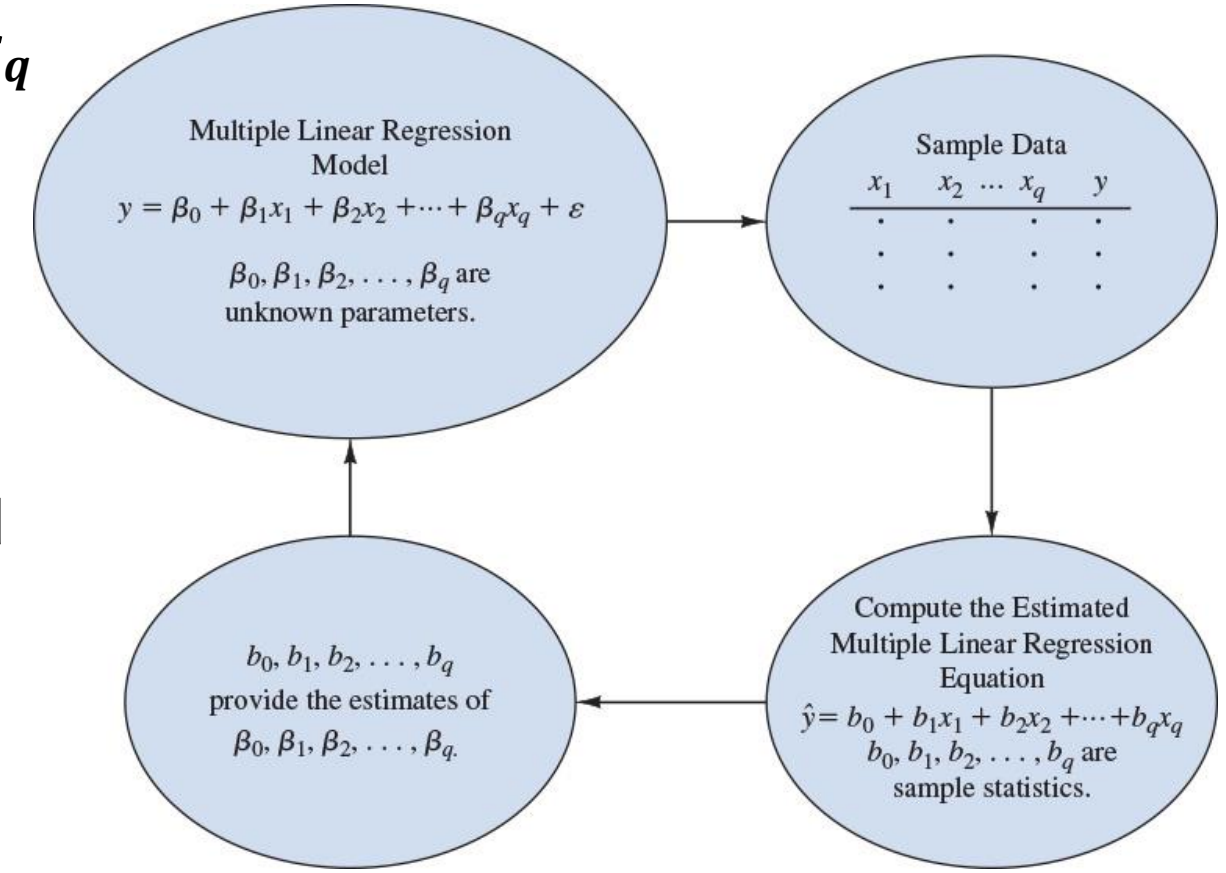
The Estimation Process in Multiple Regression

The **estimated multiple linear regression equation** is

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

Where:

- $\hat{y}$ is a point estimate of $E(y)$ for a given set of p independent variables, $x_1, x_2, \ldots, x_q$.
- A simple random sample is used to compute the sample statistics $b_0, b_1, b_2, \ldots, b_q$ that are used as estimates of $\beta_0, \beta_1, \beta_2, \ldots, \beta_q$.



Least Squares Method and Multiple Regression

The least squares method uses the sample data to provide the values of the sample statistics $b_0, b_1, b_2, \ldots, b_q$ that minimize the sum of the square errors between the y_i and the $\hat{y}_i$.

$$\min \sum (y_i - \hat{y}_i)^2 = min \sum \left(y_i - \left(b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_q x_q \right) \right)^2 = \min \sum e_i^2$$

Where, y_i is the value of dependent variable for the ith observation.

$\hat{y}_i$ is the predicted value of dependent variable for the ith observation.

Because the formulas for the regression coefficients involve the use of matrix algebra, we 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 with Two Independent Variables

Data file: *butlerwithdeliveries.xlsx*

We add a second independent variable, the number of deliveries made per driving assignment, which also contributes to the total travel time.

The estimated multiple linear regression with two independent variables is

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

Where

$\hat{y}$ = estimated mean travel time

x_1 = distance traveled (miles)

x_2 = number of deliveries

The SSE, SST, SSR and coefficient of determination (denoted R^2 in multiple linear regression) are computed as we saw for simple linear regression.

Butler Trucking Co. and Multiple Regression

The estimated multiple linear regression equation (see notes for Excel instructions), after rounding the sample coefficients to four decimal places, is

$$\hat{y} = 0.1273 + 0.0672x_1 + 0.6900x_2$$

- For a fixed number of deliveries, the mean travel time is expected to increase by 0.0672 hours (~4 minutes) when the distance traveled increases by 1 mile.
- For a fixed distance traveled, the mean travel time is expected to increase by 0.69 hours (~41 minutes) for each additional delivery.
- The interpretation of the estimated y-intercept is not meaningful because it results from extrapolation.
- We can now explain $R^2 = 81.73\%$ of the variability in total travel time.

Butler Trucking Co. Excel Regression Output

	A	B	C	D	E	F	G	H	I
1	SUMMARY OUTPUT								
2									
3	*Regression Statistics*								
4	Multiple R	0.90407397							
5	R Square	0.817349743							
6	Adjusted R Square	0.816119775							
7	Standard Error	0.829967216							
8	Observations	300							
9									
10	ANOVA								
11		*df*	*SS*	*MS*	*F*	*Significance F*			
12	Regression	2	915.5160626	457.7580313	664.5292419	2.2419E-110			
13	Residual	297	204.5871374	0.68884558					
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	0.127337137	0.20520348	0.620540826	0.53537766	-0.276499931	0.531174204	-0.404649592	0.659323866
18	Miles	0.067181742	0.002454979	27.36551071	3.5398E-83	0.062350385	0.072013099	0.06081725	0.073546235
19	Deliveries	0.68999828	0.029521057	23.37308852	2.84826E-69	0.631901326	0.748095234	0.613465414	0.766531147

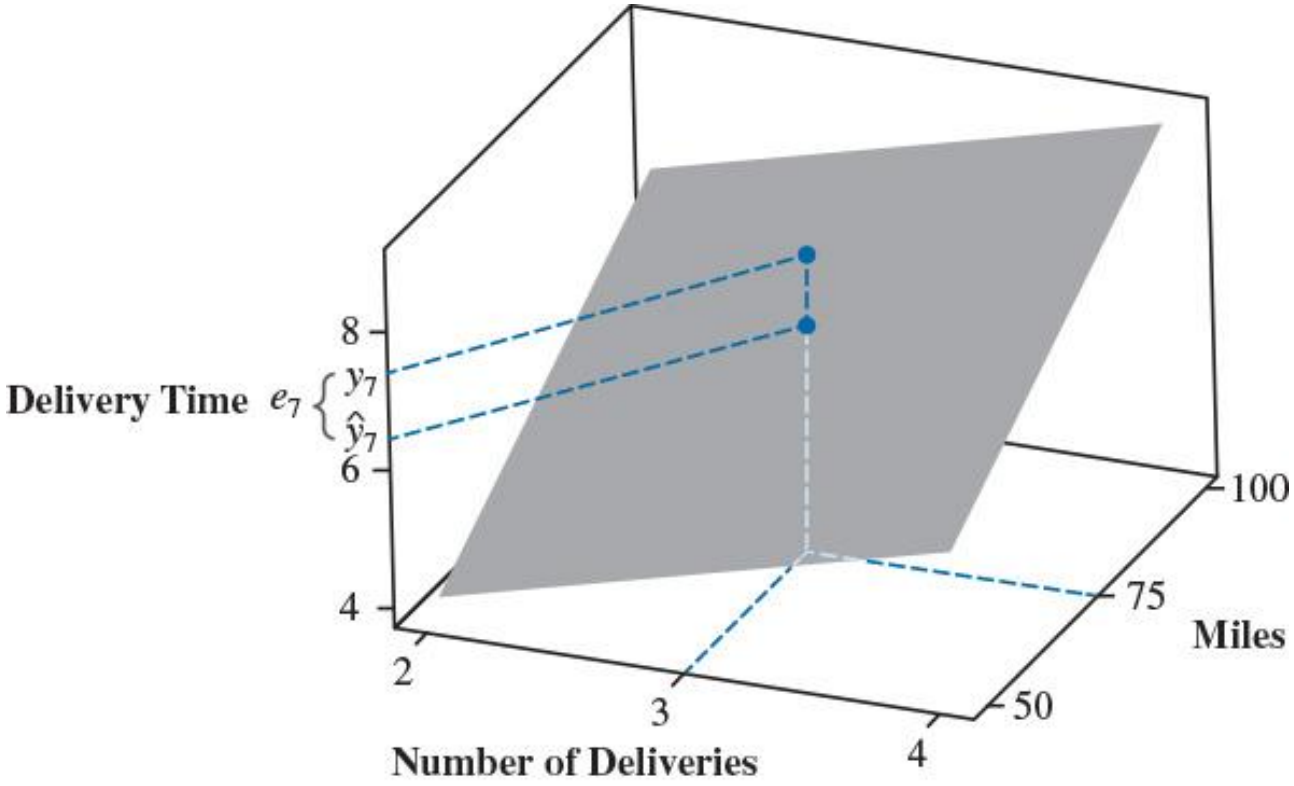
Graph of the Multiple Linear Regression Equation

With two independent variables x_1 and x_2, we now generate a predicted value of y for every combination of values of x_1 and x_2.

- Instead of a regression line, we now create a 3-D regression plane.

The graph of the estimated regression plane shows the seventh driving assignment for the Butler Trucking Company example.

*See notes for details on the interpretation of the graph.



Conditions for Valid Inference in Regression

Given a multiple linear regression model expressed as

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_q x_q + \varepsilon$$

The least squares method is used to develop estimates of the model parameters resulting in the estimated multiple linear regression equation.

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

The validity of inferences depends on the two conditions about the error term ε.

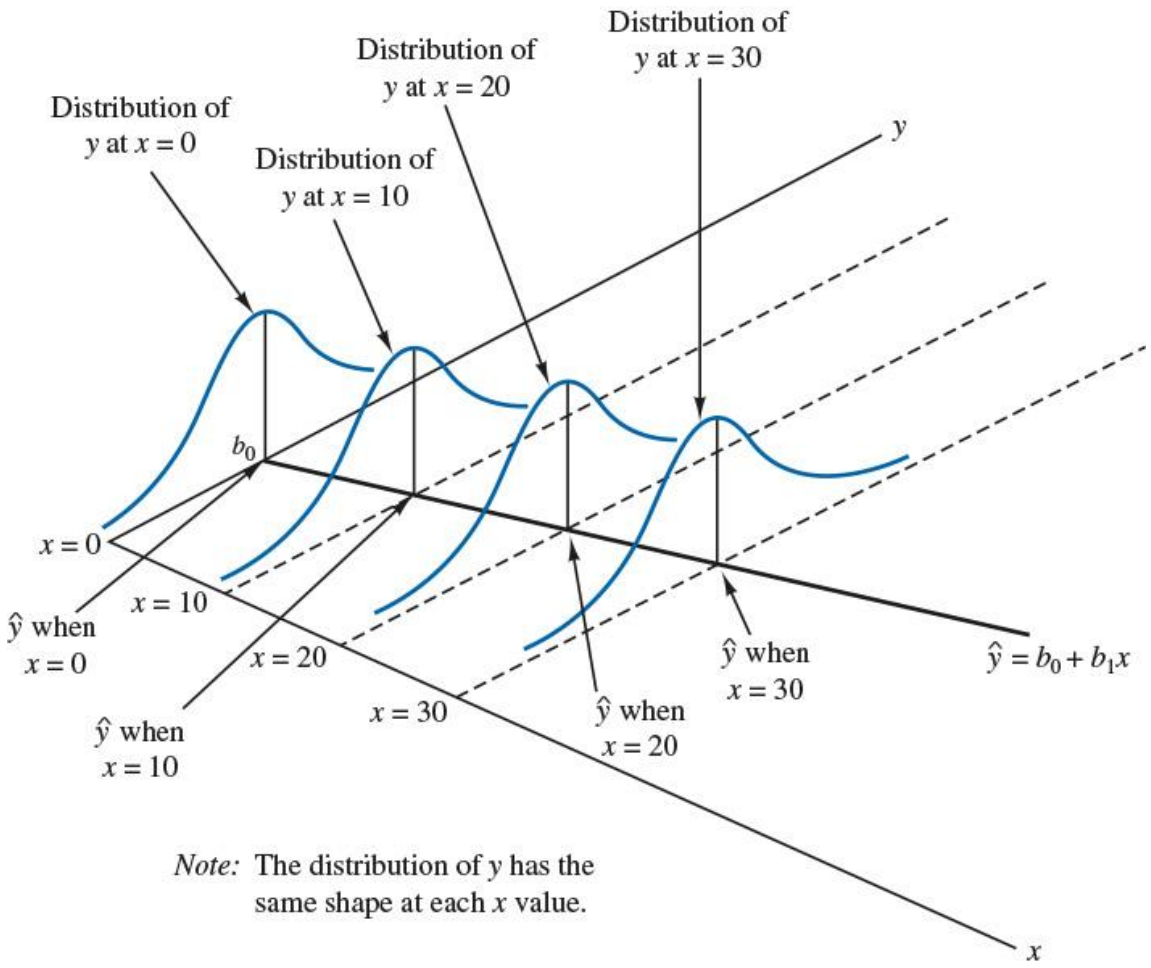
1. For any given combination of values of the independent variables $x_1, x_2, \ldots, x_q$, the population of potential error terms ε is normally distributed with a mean of 0 and a constant variance.
2. The values of ε are statistically independent.

Illustration of the Conditions for Valid Inference

The value of $E(y|x)$ changes linearly according to the specific value of x considered, and so the mean error is zero at each value of x.

The error term ε and hence the dependent variable y are normally distributed with the same variance.

The specific value of the error term ε at any particular point depends on whether the actual value of y is greater or less than $E(y|x)$.



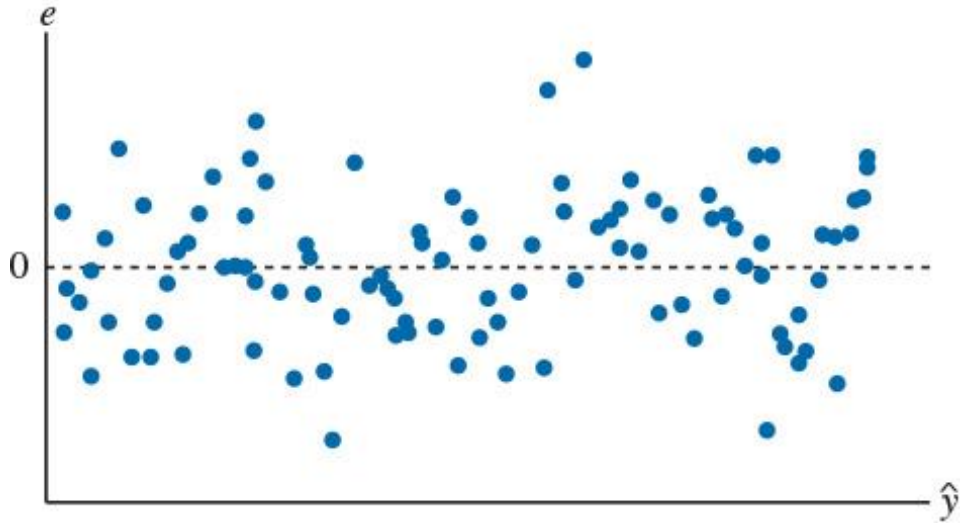
Scatter Chart of the Residuals

A simple scatter chart of the residuals is an extremely effective method for assessing whether the error term conditions are violated.

The example to the right displays a random error pattern for a scatter chart of residuals versus the predicted values of the dependent variable.

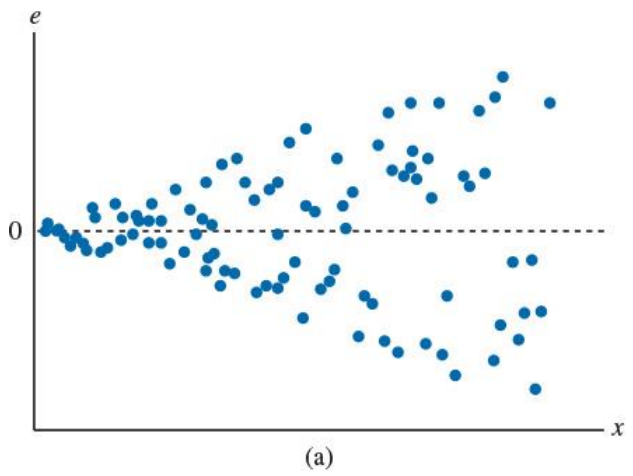
For proper inference, the scatter chart must exhibit a random pattern with

- residuals centered around zero,
- a constant spread of the residuals throughout, and
- residuals symmetrically distributed with the values near zero occurring more frequently than those outside.

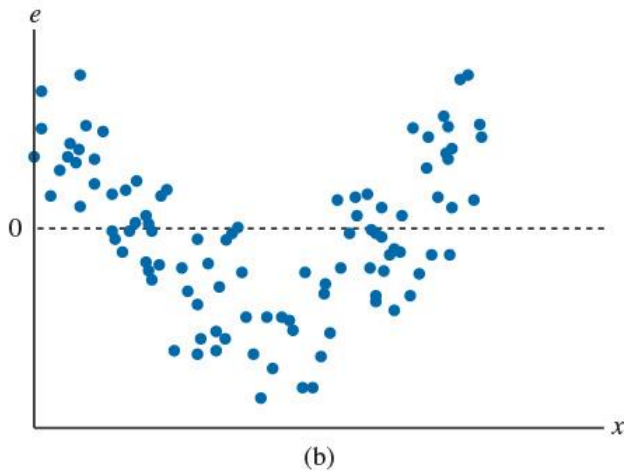


Common Error Term Violations

Non-constant spread



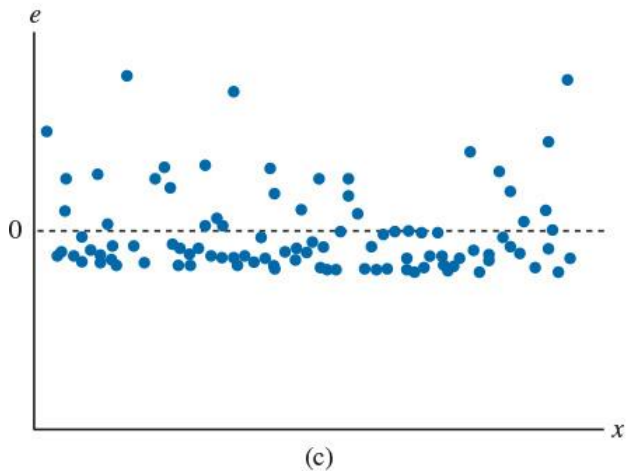
(a)



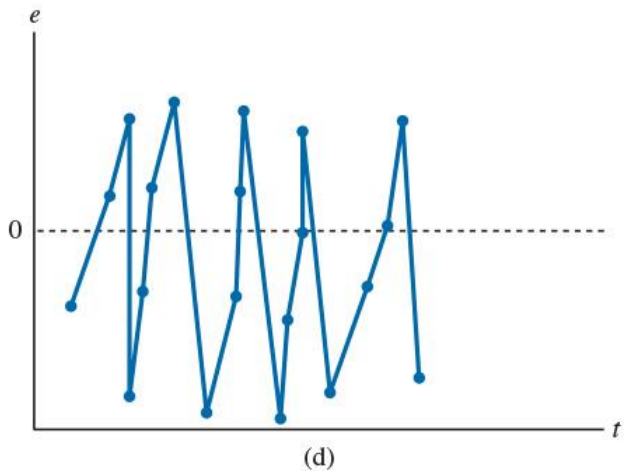
(b)

Nonlinear pattern

Non-normal residuals



(c)

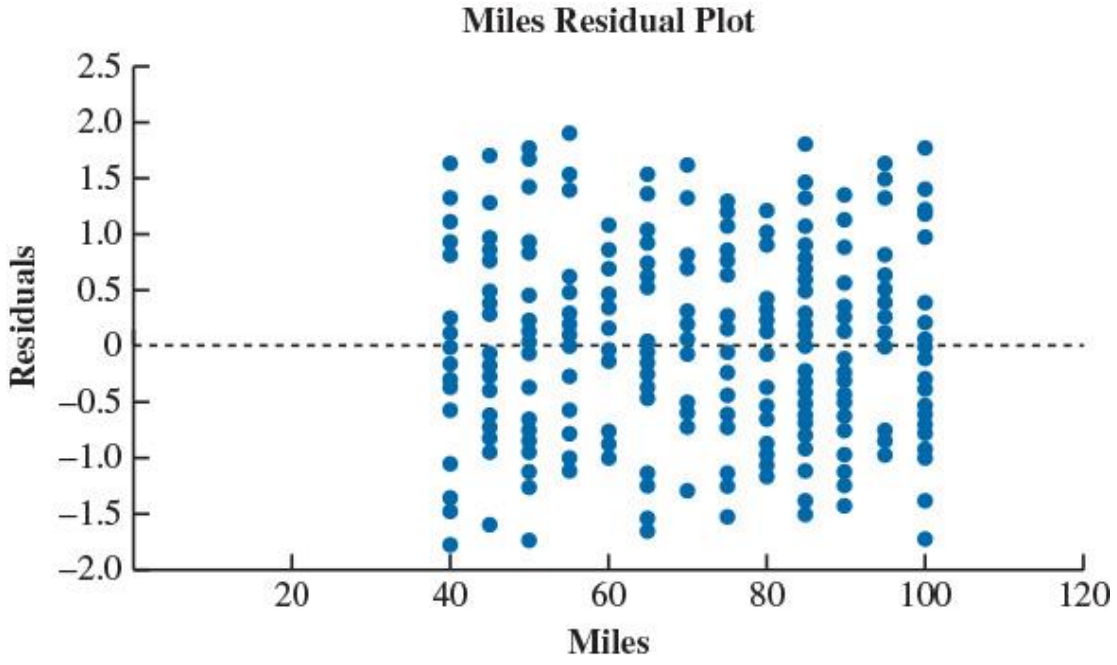


(d)

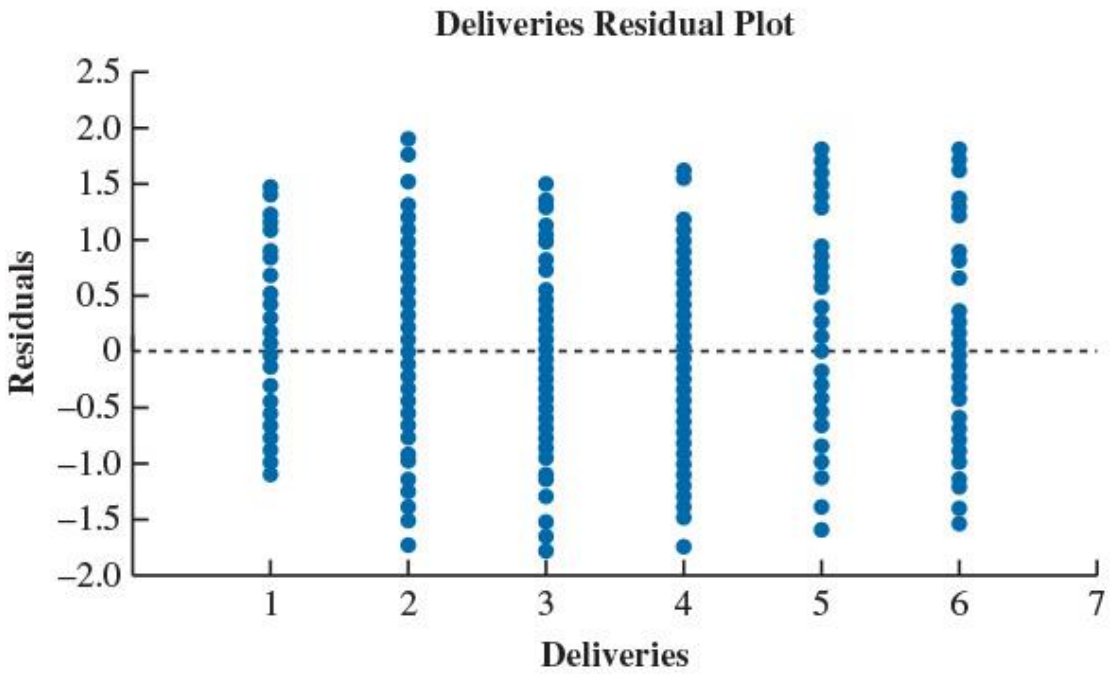
Non-independent residuals

Excel Residual Plots for the Butler Trucking Co.

Residuals vs. Miles



Residuals vs. Deliveries



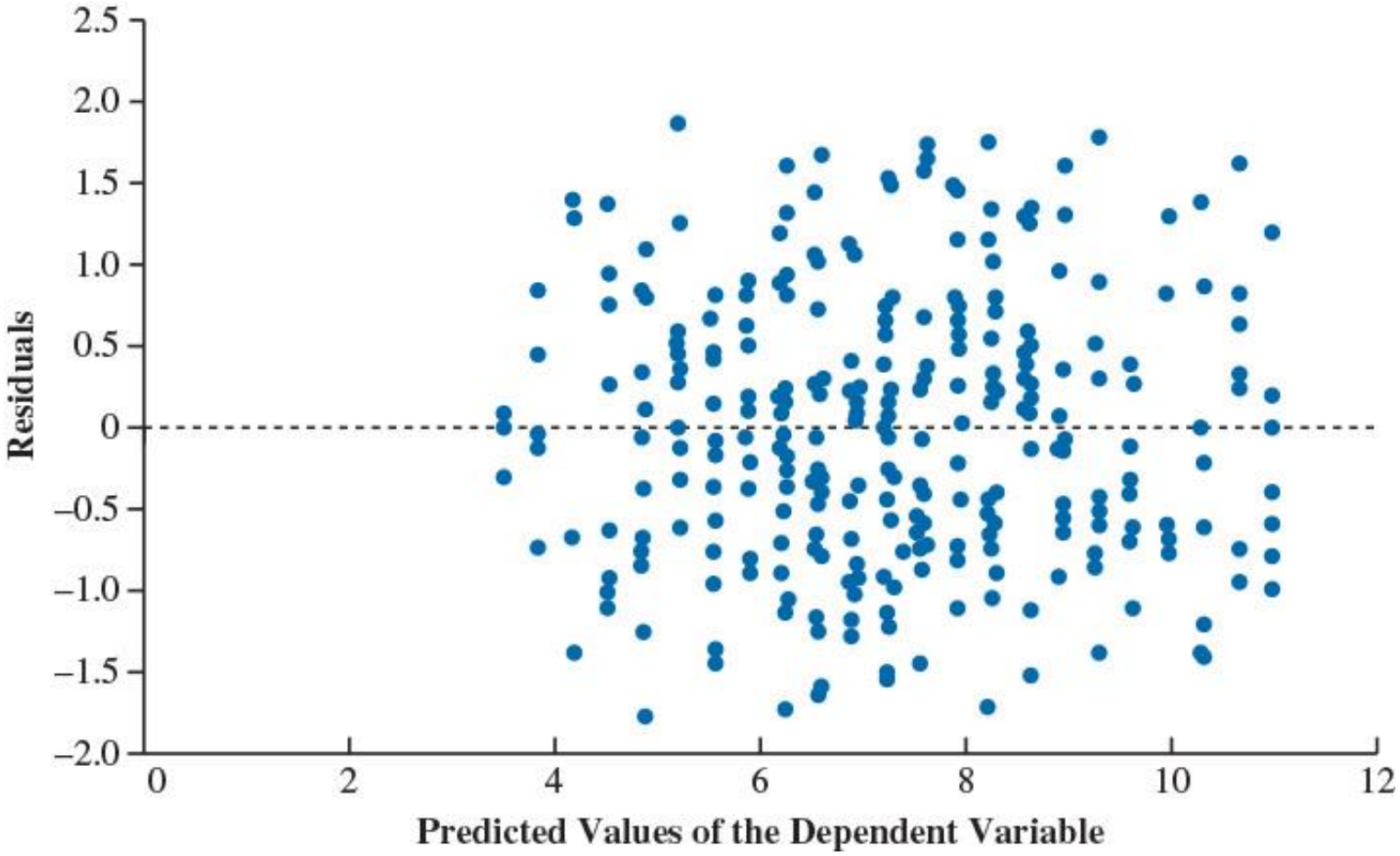
Both Residual plots show valid conditions for inference. In Excel, select the **Residual Plots** option in the **Residuals** area of the **Regression** dialog box.

Scatter Chart of Residuals vs. Predicted Variable

A scatter chart of the residuals against the predicted values $\hat{y}$ is also commonly used.

The scatter chart to the right for the Butler Trucking Company data shows valid conditions for inference.

See notes to create the data and this chart in Excel.



t Test for Individual Significance

In a multiple regression model with *p* independent variables, for each parameter β_j $(j = 1 \ldots q)$, we use a ***t* test** to test the hypothesis that parameter β_j is zero.

$$\boldsymbol{H_0}: \boldsymbol{\beta_j = 0}$$
$$\boldsymbol{H_a}: \boldsymbol{\beta_j \neq 0}$$

The test statistic follows a *t* distribution with $n - q - 1$ degrees of freedom.

$$\boldsymbol{t = b_j / s_{b_j}}$$

Where s_{b_j} is the estimated standard deviation of the regression coefficient b_j.

If $p\text{–value} \leq \alpha$, we reject H_0 and conclude that there is a linear relationship between the dependent variable y and the independent variable x_j.

Statistical software will generally report a $p\text{–value}$ for each test statistic.

Individual *t* Tests

The multiple regression output for the Butler Trucking Company example shows the *t*-ratio calculations. The values of b_1, b_2, s_{b_1}, and s_{b_1} are as follows.

Variable Miles: $b_1 = 0.0672 \quad s_{b_1} = 0.00245$

Variable Deliveries: $b_2 = 0.6900 \quad s_{b_1} = 0.02952$

Calculation of the *t*-ratios provide the test statistic for the hypotheses involving parameters β_1 and β_2, also provided by the computer output.

$$b_1/s_{b_1} = 0.0672/0.00245 = 27.37$$

$$b_2/s_{b_2} = 0.6900/0.02952 = 23.37$$

Using $\alpha = 0.01$, the *p*-values of 0.000 in the output indicate that we can reject H_0: $b_1 = 0$ and H_0: $b_2 = 0$. Hence, both parameters are statistically significant.

Testing Regression Coefficients with Confidence Intervals

Confidence interval can be used to test whether each of the regression parameters $\beta_1, \beta_2, \ldots, \beta_q$ is equal to zero.

To test that β_j is zero (i.e., there is no linear relationship between x_j and y) at some predetermined level of significance (say 0.05), first build a confidence interval at the $(1 - 0.05)100\%$ confidence level.

If the resulting confidence interval does not contain zero, we conclude that β_j differs from zero at the predetermined level of significance.

The multiple regression output for the Butler Trucking Company example shows each regression coefficient's confidence intervals at 95% and 99% confidence levels.

Addressing Nonsignificant Independent Variables

If practical experience dictates that a nonsignificant independent variable x_j is related to the dependent variable y, the independent variable x_j should be left in the model.

If the model sufficiently explains the dependent variable y without the nonsignificant independent variable x_j, then consider rerunning the regression without the nonsignificant independent variable x_j.

At times, the estimates of the other regression coefficients and their p−values may change considerably when we remove the nonsignificant independent variable x_j from the model.

The appropriate treatment of the inclusion or exclusion of the y-intercept when b_0 is not statistically significant may require special consideration (*see notes.)

Multicollinearity

Multicollinearity refers to the correlation among the independent variables in multiple regression analysis (*see notes.)

- Multicollinearity increases the standard errors of the regression estimates of $\beta_1, \beta_2, \ldots, \beta_q$ and the predicted values of the dependent variable y so that inference based on these estimates is less precise than it should be.

In *t* tests for the significance of individual parameters, it is possible to conclude that a parameter associated with one of the multicollinear independent variables is not significantly different from zero when the independent variable has a strong relationship with the dependent variable instead.

The presence of multicollinearity is excluded when there is little correlation among the independent variables.

Multicollinearity in the Butler Trucking Co. Data

Data file: *butlergasconsumption.xlsx*

The regression output to the right has miles driven (x_1) and gasoline consumption (x_2) as independent variables.

The two variables are highly correlated, as gas consumption increases with total miles driven, with a correlation coefficient of 0.9572.

Because of multicollinearity, the regression coefficient β_2 is not significant, with $p-value = 0.6588$.

	A	B	C	D	E
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*
12	Regression	2	539.5808158	269.7904079	138.0269794
13	Residual	297	580.5223842	1.954620822	
14	Total	299	1120.1032		
15					
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*
17	Intercept	2.493095385	0.33669895	7.404523781	1.36703E-12
18	Miles	0.074701825	0.014274552	5.233216928	3.15444E-07
19	Gasoline Consumption	-0.067506102	0.152707928	-0.442060235	0.658767336

Dummy Variables

Thus far, the regression examples we have considered involved quantitative independent variables such as distance traveled, gas consumption, and number of deliveries.

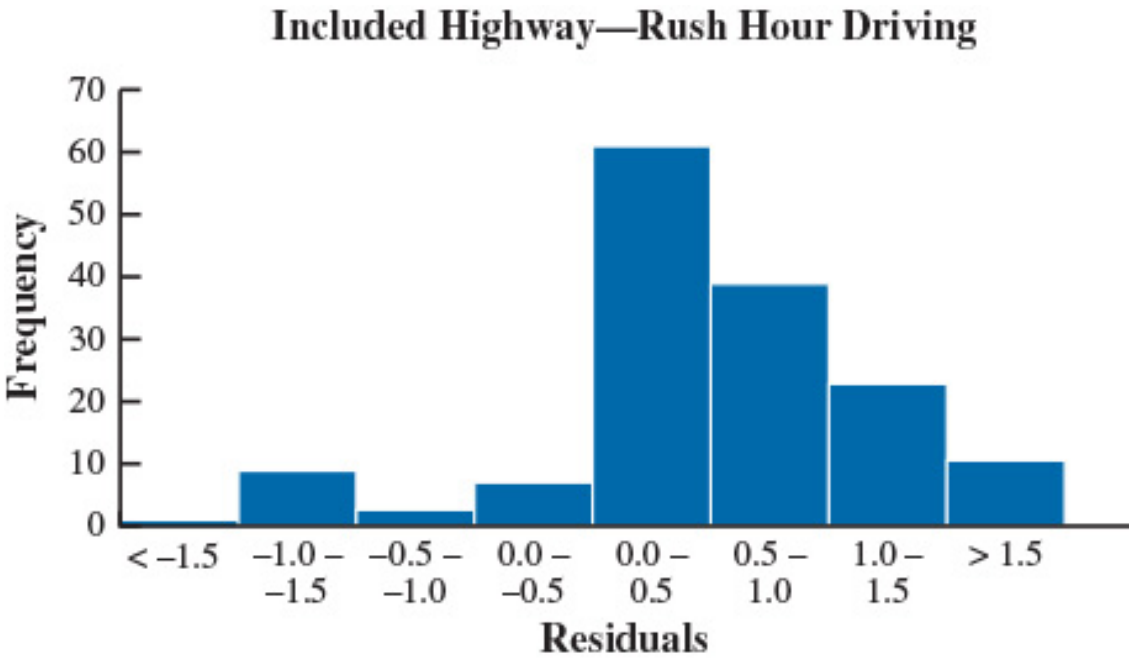
Often, we must work with **categorical independent variables**, such as:

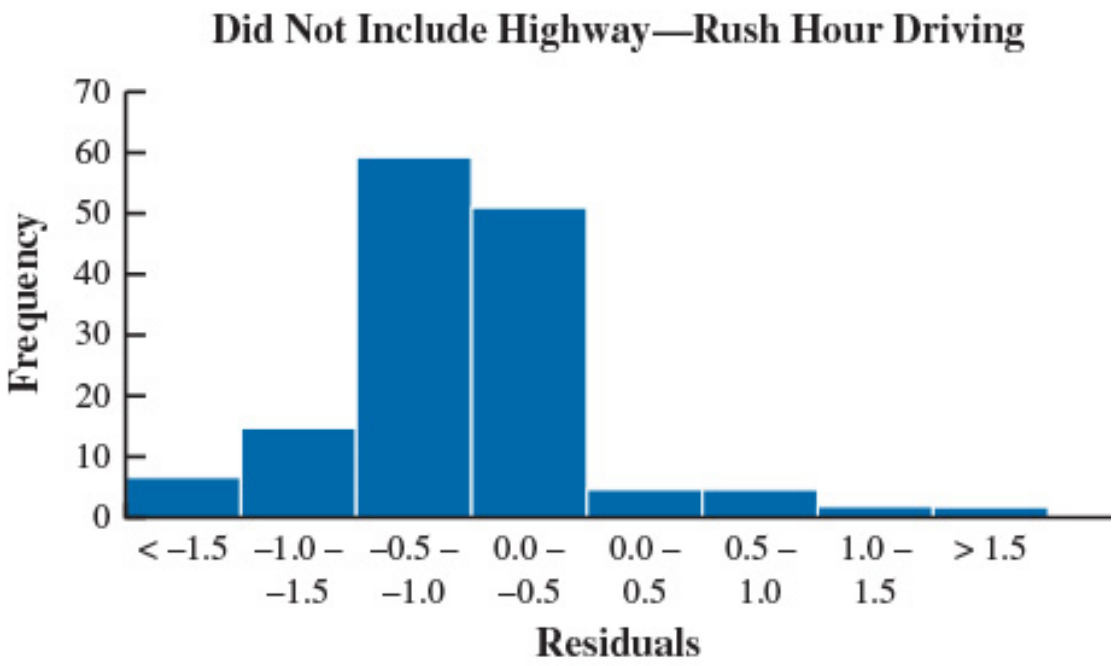
- gender (male, female)
- method of payment (cash, credit card, check)

To add a two-level categorical independent variable into a regression model, such as whether a driver should take the highway during afternoon rush hour in the Butler Trucking Company problem, we define a **dummy variable** as follows:

$$x_3 = \begin{cases} 0 \text{ if an assignment does not include driving on the highway during rush hour} \\ 1 \text{ if an assignment includes driving on the highway during rush hour} \end{cases}$$

Effect of Afternoon Rush Hour on Travel Time





A review of the residuals for the current model with miles traveled (x_1) and the number of deliveries (x_2) as independent variables reveals that driving on the highway during afternoon rush hour affects the total travel time (*see notes.)

Regression Output with Highway Rush Hour Variable

Data file: *butlerhighway.xlsx*

Excel regression output for the Butler Trucking Company regression model including the independent variables:

- miles traveled (x_1)
- number of deliveries (x_2)
- highway rush hour (x_3)

All independent variables are significant, and they explain ($R^2 = 0.8838$) about 88.4% of the total travel time variability.

	A	B	C	D	E
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*
12	Regression	3	989.9490008	329.9830003	750.455757
13	Residual	296	130.1541992	0.439710132	
14	Total	299	1120.1032		
15					
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*
17	Intercept	-0.330229304	0.167677925	-1.969426232	0.04983651
18	Miles	0.067220302	0.00196142	34.27125147	4.7852E-105
19	Deliveries	0.67351584	0.023619993	28.51465081	6.74797E-87
20	Highway	0.9980033	0.076706582	13.0106605	6.49817E-31

Interpreting the Parameters for the Butler Example

$$\hat{y} = -0.3302 + 0.0672x_1 + 0.6735x_2 + 0.9980x_3$$

The model estimates that travel time increases by:

1. 0.0672 hours (about 4 minutes) for every increase of 1 mile traveled, holding constant the number of deliveries and whether the driver uses the highway during afternoon rush hour.
2. 0.6735 hours (about 40 minutes) for every delivery, holding constant the number of miles traveled and whether the driver uses the highway during afternoon rush hour.
3. 0.9980 hours (about 60 minutes) if the driver uses the highway during afternoon rush hour, holding constant the number of miles traveled and the number of deliveries.

More Complex Categorical Variables

If an independent categorical variable has k levels, $k-1$ dummy variables are required, with each dummy variable being coded as 0 or 1.

Consider the situation faced by a manufacturer of vending machines that sells its products to three sales territories: region A, B, and C.

To code the sales regions in a regression model that explains the dependent variable (y) number of units sold, we need to define $k-1=2$ dummy variables as follows:

$$x_1 = \begin{cases} 1 \text{ if sales region B} \\ 0 \text{ otherwise} \end{cases}$$

$$x_2 = \begin{cases} 1 \text{ if sales region C} \\ 0 \text{ otherwise} \end{cases}$$

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

The regression equation relating the expected number of units sold to the sales region can be written as $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$

Interpretation of the Parameters for a Categorical Variable with Three Levels

To interpret β_0, β_1, and β_2, in the sales territory example, consider the following variations of the regression equation.

$$E(y \mid \text{region A}) = \beta_0 + \beta_1(0) + \beta_2(0) = \beta_0$$

$$E(y \mid \text{region B}) = \beta_0 + \beta_1(1) + \beta_2(0) = \beta_0 + \beta_1$$

$$E(y \mid \text{region C}) = \beta_0 + \beta_1(0) + \beta_2(1) = \beta_0 + \beta_2$$

Thus, the regression parameters are interpreted as follows.

β_0 is the mean or expected value of sales for region A.

β_1 is the difference between the mean number of units sold in region B and the mean number of units sold in region A.

β_2 is the difference between the mean number of units sold in region C and the mean number of units sold in region A.

Overfitting

Overfitting generally results from creating an overly complex regression model to explain idiosyncrasies in the sample data.

An overfit model will overperform on the sample data used to fit the model and underperform on other data from the population.

To avoid overfitting a model:

- Use only real and meaningful independent variables.
- Only use complex models when you have reasonable expectations about them.
- Use variable selection procedures only for guidance.
- If you have sufficient data, consider **cross-validation**, in which you assess the model on data other than the sample data used to generate the model.
  - One example of cross-validation is the **holdout** method, which divides the data set between a training set and a validation set.

Inference and Very Large Samples

Virtually all regression coefficients will be statistically significant if the sample is sufficiently large.

Data file: *largecredit.xlsx*

The regression output to the right shows a modest R^2 for a data set with $n = 3{,}000$.

All the regression coefficients are significant.

See notes for details.

	A	B	C	D	E
1	SUMMARY OUTPUT				
2					
3	*Regression Statistics*				
4	Multiple R	0.603724482			
5	R Square	0.36448325			
6	Adjusted R Square	0.36363448			
7	Standard Error	4830.393498			
8	Observations	3000			
9					
10	ANOVA				
11		*df*	*SS*	*MS*	*F*
12	Regression	4	40078588918	10019647230	429.4250838
13	Residual	2995	69881440529	23332701.35	
14	Total	2999	1.0996E+11		
15					
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*
17	Intercept	1712.552073	371.7837807	4.606311953	4.26973E-06
18	Annual Income ($1000)	121.6120724	3.164453912	38.43066631	4.943E-263
19	Household Size	531.213362	42.82435656	12.40446803	1.71315E-34
20	Years of Post-High School Education	-539.8345703	58.57519443	-9.216095235	5.64208E-20
21	Hours Per Week Watching Television	12.55178379	5.109759992	2.456433142	0.014088759

Model Selection

When dealing with large samples, it is often difficult to discern the most appropriate model.

- If developing a regression model for explanatory purposes, the practical significance of the estimated regression coefficients should be considered when interpreting the model and considering which variables to keep.
- If developing a regression model to make future predictions, selecting the independent variables to include in the model should be based on the predictive accuracy of observations that have not been used to train it.

For example, the credit card data set could be split into two data sets:

1. a training data set with $n = 2{,}000$, and
2. a validation data set with $n = 1{,}000$.

Summary

- We discussed how linear regression analysis is used to determine how a dependent variable y is related to one or more independent variables.
- We used sample data and the least squares method to develop the estimated simple linear regression equation, interpreted its coefficients, and presented the coefficient of determination as a measure of its goodness of fit.
- We then extended our discussion to include multiple independent variables and reviewed how to use Excel to find the estimated multiple linear regression equation, build estimates in the form of prediction and confidence intervals, and the ramifications of multicollinearity.
- We discussed the necessary conditions for the linear regression model and its associated error term to conduct valid inference for regression.
- We showed how to incorporate categorical independent variables into a regression model and discussed how to fit nonlinear relationships.
- Finally, we discussed various variable selection procedures, the problem of overfitting, and the implication of big data on regression analysis.