Simple Linear Regression

Introduction

Managerial decisions often are based on the relationship between two or more variables.

A statistical procedure called *regression analysis* can be used to develop an equation showing how the variables are related.

- The variable being predicted is called the **dependent variable** and is denoted by y.
- The variables being used to predict the value of the dependent variable are called the **independent variables** and are denoted by x.

In this chapter we consider **simple linear regression**, the simplest type of regression:

- it involves one independent variable and one dependent variable.
- the relationship between the two variables is approximated by a straight line.

Regression analysis involving two or more independent variables is called multiple regression, and it will be the subject of the next two chapters.

Simple Linear Regression Model

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

The *simple linear regression model* is described by the following equation.

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

Where

y is the dependent variable.

x is the independent variable.

β_0 and β_1 are referred to as the parameters of the model.

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

Simple Linear Regression Equation

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

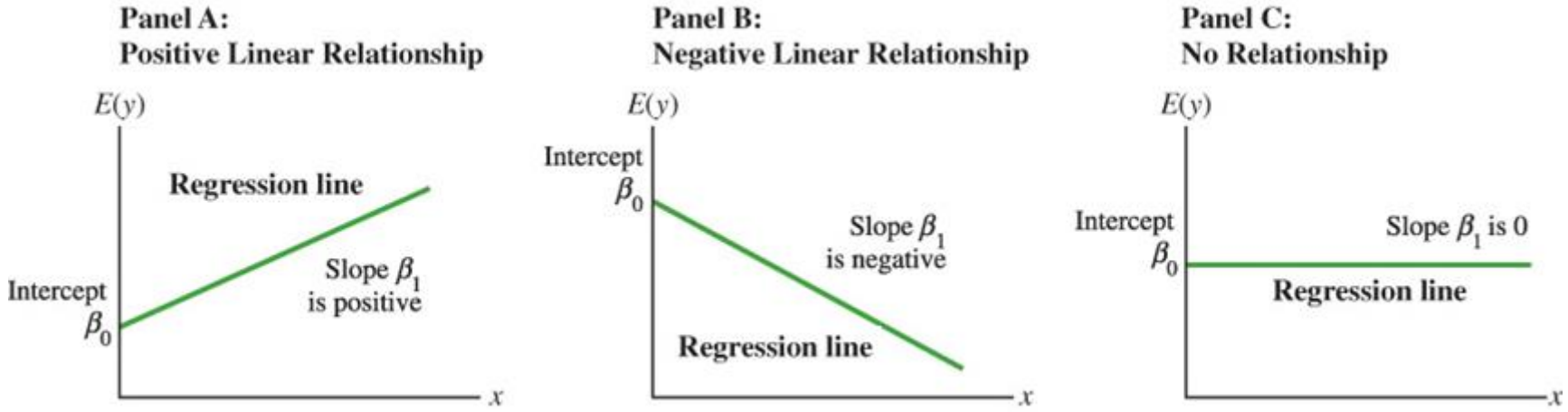
$$\boldsymbol{E(y) = \beta_0 + \beta_1 x}$$

Where

$E(y)$ is the expected value of a distribution of y values for a given x value.

β_0 is the y-intercept of the regression line.

β_1 is the slope of the regression line. Its sign dictates the type of linear relationship.



The Estimation Process in Simple Linear Regression

The *estimated simple linear regression equation* is

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

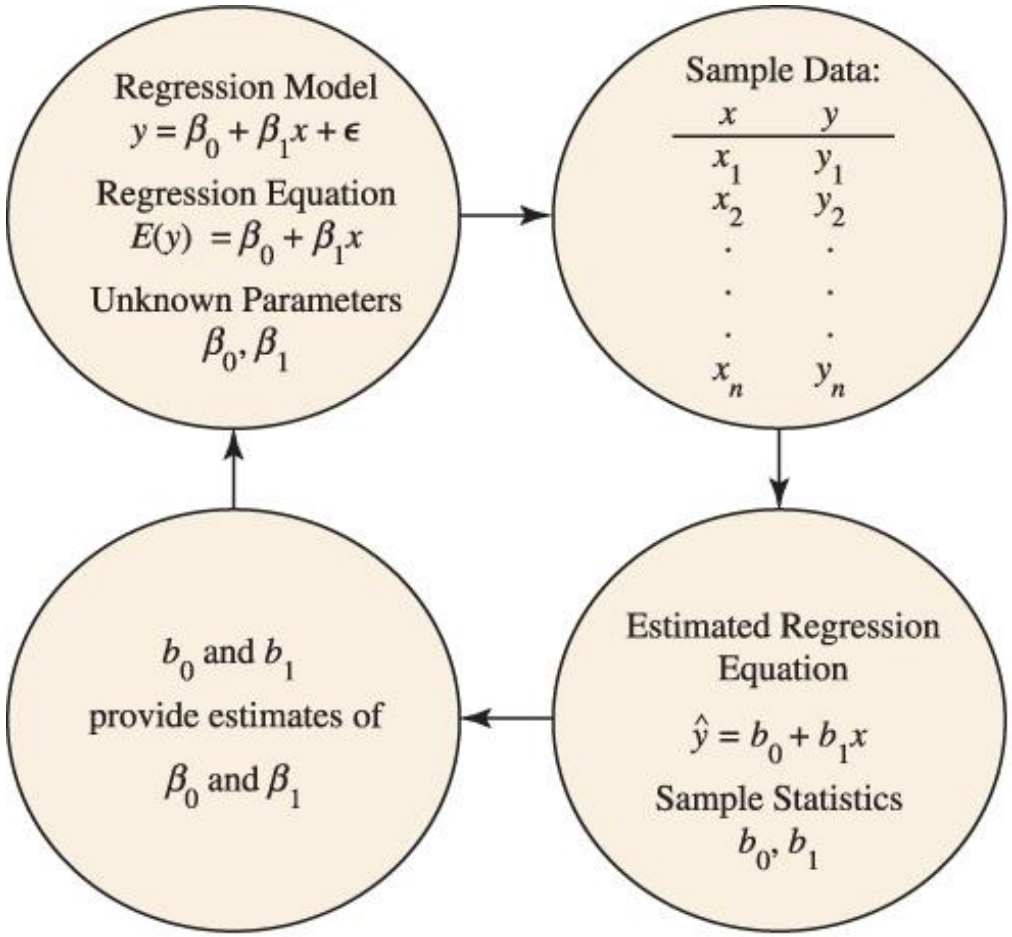
The graph of the estimated simple linear regression equation is called the *estimated regression line*, where:

$\hat{y}$ is a point estimate of $E(y)$ for a given x.

b_0 is the y-intercept of the regression.

b_1 is the slope of the regression.

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



Least Squares Method

The **least squares method** uses the sample data to provide the values of b_0 and b_1 that minimize the sum of the squares of the deviations between the observed values of the dependent variable y_i and the predicted values of the dependent variable $\hat{y}_i$.

Least Squares Criterion

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

Slope and y-Intercept for the Estimated Regression Equation

$$b_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} \qquad b_0 = \bar{y} - b_1 \bar{x}$$

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

$\bar{x}$ and $\bar{y}$ are the mean values for the independent and dependent variables.

n is the total number of observations.

The Armand's Pizza Parlors Example

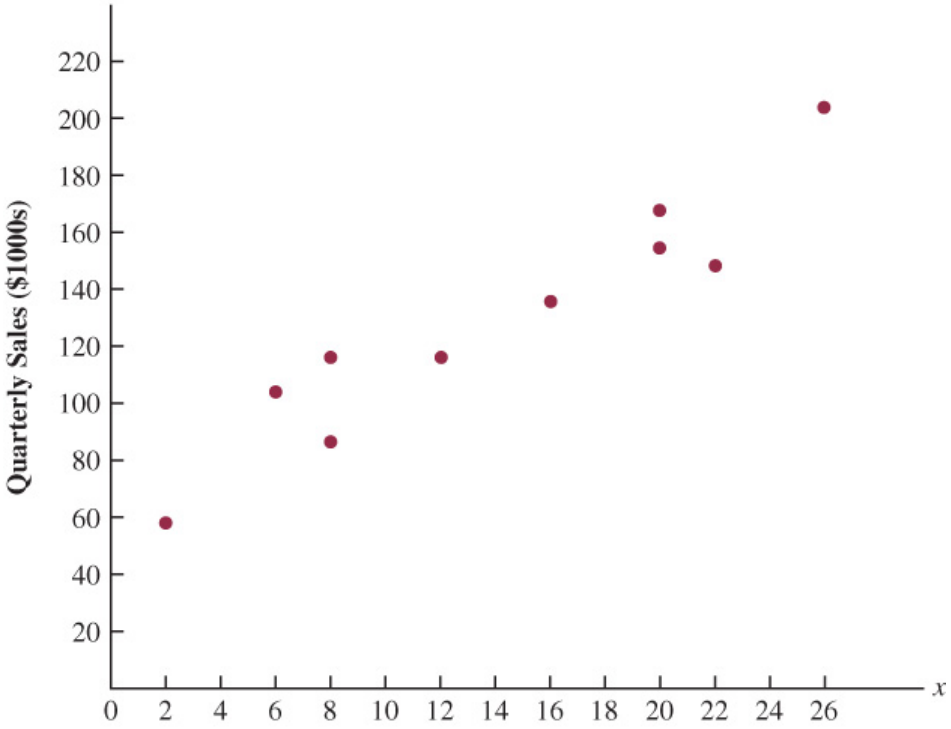
The managers at Armand's Pizza Parlors, a chain of Italian-food restaurants located near college campuses, believe that restaurants' quarterly sales (denoted by y) are positively related to the size of the student population (denoted by x.)

In this example, the chain of the Armand's restaurants constitutes the population of interest.

DATAfile: *Armands*

We use a sample of 10 randomly selected restaurants to build a scatter diagram depicting the relationship between quarterly sales and the student population.

Because the scatter diagram shows a positive linear relationship, we choose the simple linear regression model to represent the relationship between Armand's quarterly sales (y) and student population (x.)



Estimated Regression Equation for Armand's Pizza

The calculations necessary to develop the least squares estimated regression equation for the sample of $n = 10$ Armand's Pizza Parlors restaurants are shown in the table below.

We first calculate $\sum x_i = 140$ and $\sum y_i = 1{,}300$. Dividing them by $n = 10$, we obtain $\bar{x} = 14$ and $\bar{y} = 130$, which we then use to compute $\sum(x_i - \bar{x})(y_i - \bar{y}) = 2840$ and $\sum(x_i - \bar{x})^2 = 568$.

The calculations of b_1 and b_0 are :

$$b_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} = \frac{2{,}840}{568} = 5$$

$$b_0 = \bar{y} - b_1\bar{x} = 130 - 5(14) = 60$$

Thus, the estimated regression equation for the Armand's Pizza Parlors sample of 10 restaurants is

$$\hat{y} = 60 + 5x$$

Restaurant *i*	x_i	y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$
1	2	58	-12	-72	864	144
2	6	105	-8	-25	200	64
3	8	88	-6	-42	252	36
4	8	118	-6	-12	72	36
5	12	117	-2	-13	26	4
6	16	137	2	7	14	4
7	20	157	6	27	162	36
8	20	169	6	39	234	36
9	22	149	8	19	152	64
10	26	202	12	72	864	144
	140	1300			2840	568
	$\sum x_i$	$\sum y_i$			$\sum(x_i - \bar{x})(y_i - \bar{y})$	$\sum(x_i - \bar{x})^2$

Interpretation of the Estimated Regression Equation

The slope of the estimated regression equation ($b_1 = 5$) is positive, implying that as the student population increases, sales also increase.

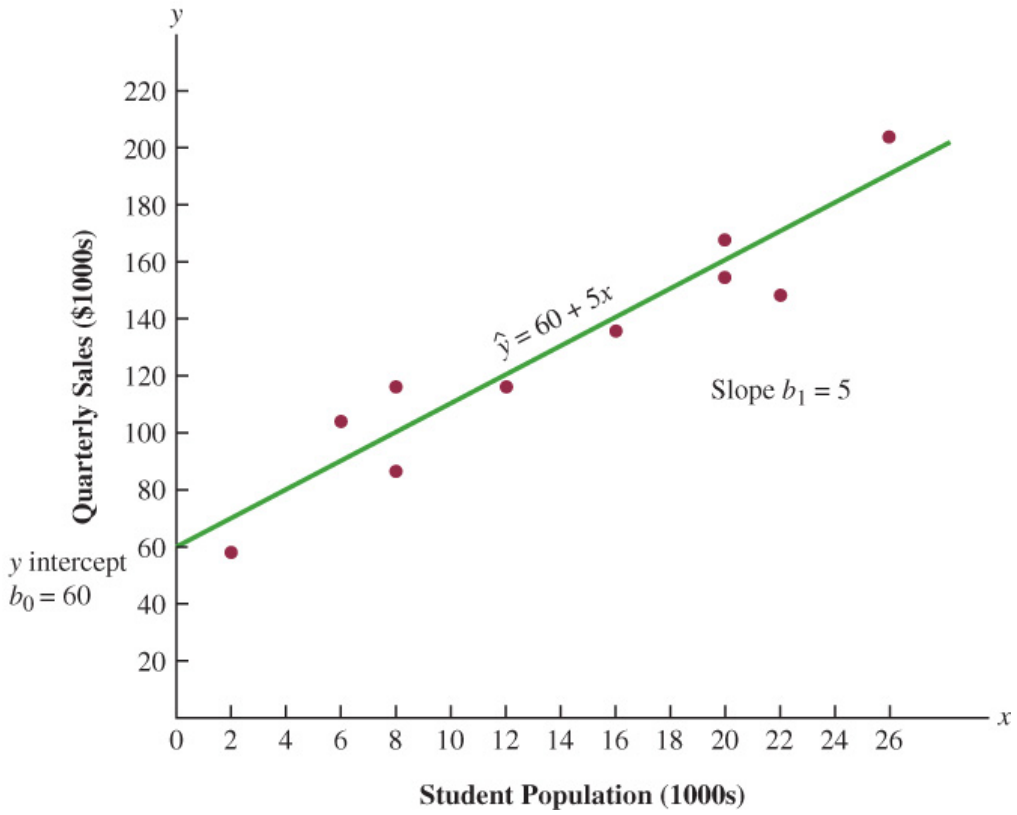
We can conclude that quarterly sales are expected to increase by

$$b_1 = 5\left[\frac{\$1{,}000}{1{,}000 \text{ students}}\right] = \$5 \text{ per student}$$

We can also use the estimated regression equation to predict the value of y for a given value of x. For a restaurant located near a 16,000 students' campus

$$\hat{y} = 60 + 5x = 60 + 5(16) = 140 \ [\$1{,}000]$$

Hence, we predict quarterly sales of $140,000 for this restaurant.



Sum of Squares Due to Error (SSE)

$y_i - \hat{y}_i$ is the **residual** for the ith observation; that is, the difference between the observed value of the dependent variable, y_i, and the predicted value of the dependent variable, $\hat{y}_i$.

The sum of squares of the residuals, also known as the **sum of squares due to error** (SSE), is the quantity minimized by the least squares method.

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

The table to the right shows the calculations required to compute SSE for the Armand's Pizza Parlors example.

Thus, $SSE = 1{,}530$ measures the error in using the estimated regression equation $\hat{y} = 60 + 5x$ to predict sales.

Restaurant *i*	x_i = Student Population (1000s)	y_i = Quarterly Sales ($1000s)	Predicted Sales $\hat{y}_i = 60 + 5x_i$	Error $y_i - \hat{y}_i$	Squared Error $(y_i - \hat{y}_i)^2$
1	2	58	70	-12	144
2	6	105	90	15	225
3	8	88	100	-12	144
4	8	118	100	18	324
5	12	117	120	-3	9
6	16	137	140	-3	9
7	20	157	160	-3	9
8	20	169	160	9	81
9	22	149	170	-21	441
10	26	202	190	12	144
					SSE = 1,530

Total Sum of Squares (SST)

For the ith restaurant in the sample, the difference $y_i - \bar{y}$ provides a measure of the error when using $\bar{y}$ to predict sales. The corresponding sum of squares, called the **total sum of squares**, is denoted by SST.

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

The table to the right shows the calculations required to compute SST for the Armand's Pizza Parlors example.

Thus, $SST = 15{,}730$ measures the error in using $\bar{y} = 130$ to predict sales.

Note how much larger the squared deviations for SST are when compared to the squared error values for SSE.

Restaurant *i*	x_i = Student Population (1000s)	y_i = Quarterly Sales ($1000s)	Deviation $y_i - \bar{y}$	Squared Deviation $(y_i - \bar{y})^2$
1	2	58	-72	5,184
2	6	105	-25	625
3	8	88	-42	1,764
4	8	118	-12	144
5	12	117	-13	169
6	16	137	7	49
7	20	157	27	729
8	20	169	39	1,521
9	22	149	19	361
10	26	202	72	5,184
				SST = 15,730

Sum of Squares Due to Regression (SSR)

The difference $\hat{y}_i - \bar{y}$ provides a measure of how much the $\hat{y}_i$ values on the estimated regression line (depicted in green in the scatter diagram) deviate from $\bar{y}$ (depicted in blue.)

The corresponding sum of squares is the **sum of squares due to regression**, SSR.

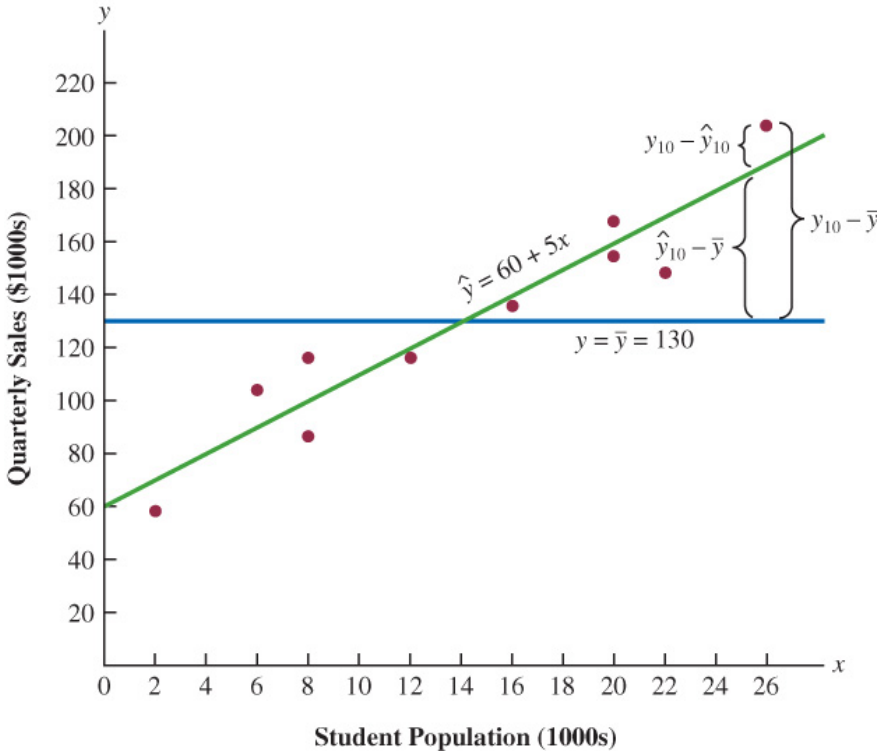
$$\boldsymbol{SSR = \sum(\hat{y}_i - \bar{y})^2}$$

The total sum of squares, SST, can be partitioned into two components: the sum of squares due to regression, SSR, and the sum of squares due to error, SSE.

$$SST = SSR + SSE$$

For instance, in the Armand's Pizza Parlors example, we can solve for SSR and obtain

$$SSR = 15{,}730 - 1530 = 14{,}200$$



Coefficient of Determination

The estimated regression equation provides a perfect fit when every value of the dependent variable y_i lies on the estimated regression line, so that all residuals $y_i - \hat{y}_i = 0$ and $SSE = 0$.

Because $SST = SSR + SSE$, for a perfect fit, $SSR = SST$ and the ratio $SSR/SST = 1$.

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 zero and one, and it is used to evaluate the goodness of fit for the estimated regression equation.

For the Armand's Pizza Parlor example, we have: $r^2 = \frac{SSR}{SST} = \frac{14{,}200}{15{,}730} = 0.9027$

Thus, about 90.3% of the variability in the quarterly sales at Armand's Pizza Parlor can be explained by the linear relationship between the population of students and quarterly sales.

Correlation Coefficient

We previously introduced the **correlation coefficient** as a descriptive measure of the strength of the linear association between two variables, x and y, and having the following properties:

- Values of the correlation coefficient are always between −1 and +1.
- A value of +1 indicates that x and y are aligned on a straight line with a positive slope.
- A value of −1 indicates that x and y are aligned on a straight line with a negative slope.
- Values of the correlation coefficient near zero indicate that x and y are not linearly related.

The sample correlation coefficient for regression equation $\hat{y} = b_0 + b_1 x$ is computed as follows.

$$r_{xy} = (\text{sign of } b_1)\sqrt{r^2}$$

For the Armand's Pizza Parlor example, we had $\hat{y} = 60 + 5x$ and $r^2 = 0.9027$. Thus, we have

$$r_{xy} = +\sqrt{0.9027} = +0.9501$$

The positive value near +1 indicates a strong positive linear association between x and y.

Assumptions About the Error Term ϵ

The tests of significance in regression analysis are based on the following assumptions about the error term, ϵ. In the regression model $y = \beta_0 + \beta_1 x + \epsilon$, we have:

1. The error term ϵ is a random variable with an expected value of zero; that is, $E(\epsilon) = 0$

   Implication: because β_0 and β_1 are constant, $E(\beta_0) = \beta_0$ and $E(\beta_1) = \beta_1$. Thus, for a given value of x, we obtain the simple linear regression equation, $E(y) = \beta_0 + \beta_1 x$.

2. The variance of ϵ, denoted by σ^2, is the same for all values of x.

   Implication: The variance of y about the regression line equals σ^2 for all x values.

3. The values of ϵ are independent.

   Implication: Because the values of ϵ for any two values of x are not related, it follows that the values of y for any two values of x are also not related.

4. The error term ϵ is a normally distributed random variable for all values of x.

   Implication: Because y is a linear function of ϵ, y is also a normally distributed random variable for all values of x.

Graphical Representation of the Assumptions for the Regression Model

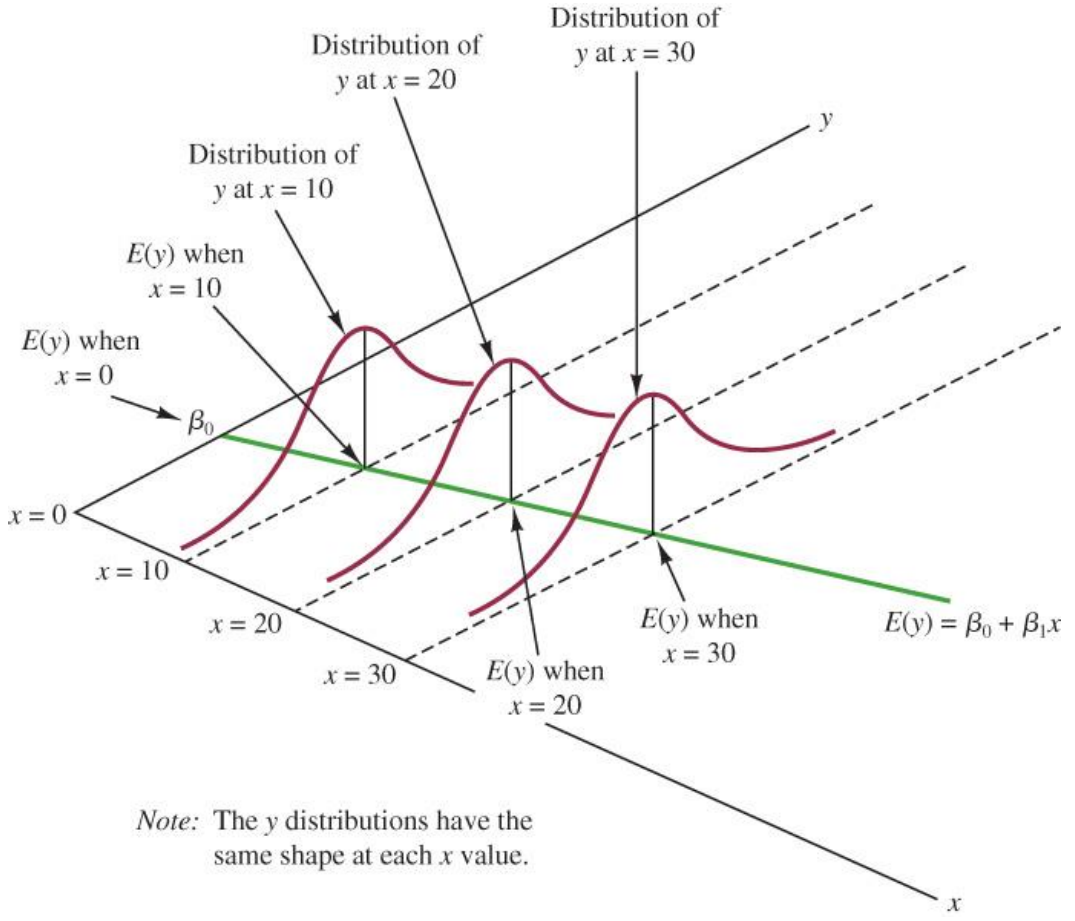
Note that the value of $E(y)$ changes according to the specific value of x.

However, the probability distribution of ϵ and the probability distribution of y are normally distributed, each with the same variance.

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

Keep in mind that when we choose a linear model, we are just making an assumption about the form of the relationship between x and y.

However, other non-linear models may very well provide a better fit for the underlying relationship.



Testing for Significance

The simple linear regression equation, $E(y) = \beta_0 + \beta_1 x$, tell us that if $\beta_1 = 0$,

$$E(y) = \beta_0 + (0)x = \beta_0.$$

- If $\beta_1 = 0$, $E(y)$ does not depend on the value of x, and x and y are not linearly related
- If $\beta_1 \neq 0$, we conclude that the two variables are related

Thus, to test for a significant regression relationship, we must conduct a hypothesis test to determine whether $\beta_1 = 0$.

Two tests are commonly used:

- the *t* test
- the *F* test.

Both *t* and *F* tests require an estimate of σ^2, the variance of ϵ, in the regression model.

Estimate of σ^2

From the assumptions on the error term, we concluded that σ^2, the variance of ϵ, also represents the variance of the residuals; that is, the deviations of the y values about the regression line.

Because SSE, the sum of squared residuals, is a measure of the variability of the y values about the estimated regression line, it follows that the **mean square error** (MSE), SSE divided by its degrees of freedom, provides s^2, the estimate of σ^2.

$$\boldsymbol{s^2 = MSE = SSE/(n-2)}$$

To estimate the standard deviation of ϵ, we take the square root of s^2. The resulting value, *s*, is referred to as the **standard error of the estimate**.

$$\boldsymbol{s = \sqrt{MSE} = \sqrt{SSE/(n-2)}}$$

For the Armand's Pizza Parlor example, we found $SSE = 1{,}530$, and $n = 10$. Thus, we have

$MSE = SSE/(n-2) = 1{,}530/(10-2) = 191.25$, and $s = \sqrt{MSE} = \sqrt{191.25} = 13.829$.

t Test for Significance in Simple Linear Regression

The purpose of the *t* test is to see whether we can conclude that $\beta_1 \neq 0$. The hypotheses are

$$\boldsymbol{H_0}: \boldsymbol{\beta_1 = 0}$$

$$\boldsymbol{H_a}: \boldsymbol{\beta_1 \neq 0}$$

b_1, the least squares estimator of β_1, follows a sampling distribution with these properties:

Expected value: $\boldsymbol{E(b_1) = \beta_1}$

Standard deviation: $\boldsymbol{\sigma_{b_1} = \sigma / \sqrt{(x_i - \bar{x})^2}}$

Distribution form: **Normal**

The test statistic follows a *t* distribution with $n - 2$ degrees of freedom.

$$\boldsymbol{t = \frac{b_1}{s_{b_1}}}$$

Where s_{b_1} is the estimated standard deviation of b_1: $\boldsymbol{s_{b_1} = s / \sqrt{(x_i - \bar{x})^2}}$

t Test for the Armand's Pizza Parlor Example

To test for a significant relationship between student population (x) and Armand's Pizza Parlor quarterly sales (y), we use a significance level, $\alpha = 0.01$.

Recollect that, for the Armand's Pizza Parlor case, we found:

$$\hat{y} = 60 + 5x$$

$$s = 13.829$$

$$\sum(x_i - \bar{x})^2 = 568$$

Hence, the estimated standard error of b_1 is

$$s_{b_1} = s/\sqrt{(x_i - \bar{x})^2} = 13.829/\sqrt{568} = 0.5803$$

And the test statistic is

$$t = \frac{b_1}{s_{b_1}} = \frac{5}{0.5803} = 8.62$$

Rejection rule for the *t* Test

Because we have a two-tailed test, the *p*-value is two times the area under the curve of the *t* distribution for $t \geq 8.62$, and with $n - 2 = 8$ degrees of freedom.

Using Table 2 in Appendix B, the *t* distribution table for 8 degrees of freedom provides

Area in upper tail	**0.20**	**0.10**	**0.05**	**0.025**	**0.01**	**0.005**
***t* Value ($df = 8$)**	0.889	1.397	1.860	2.306	2.896	3.355

With a symmetric *t* distribution, the upper tail at $t = 8.62$ is the same as the lower tail at $t = -8.62$. Thus, the *p*-value range is twice $p\text{−value} \leq 0.005$, or $p\text{−value} \leq 0.01$.

Because $p\text{−value} \leq 0.01 = \alpha$, we reject H_0 and conclude that $\beta_1 \neq 0$.

If we use a critical value approach instead, the critical values $\pm t_{\alpha/2}$ for a two-tailed test are

$$-t_{.005} = -3.355 \quad \text{and} \quad t_{.005} = 3.355$$

Because $\mathrm{t} = 8.62 \geq 3.355 = t_{.005}$, we also reject H_0 and still conclude that $\beta_1 \neq 0$.

Interval Estimate of β_1

The form of a confidence interval for β_1 is as follows.

$$\boldsymbol{b_1 \pm t_{\alpha/2} s_{b_1}}$$

Where b_1 is the point estimator of β_1, $t_{\alpha/2} s_{b_1}$ is the margin of error, $1 - \alpha$ is the confidence coefficient, and $t_{\alpha/2}$ is the *t* value providing an area of $\alpha/2$ in the upper tail of a *t* distribution with $n - 2$ degrees of freedom.

For the Armand's Pizza Parlor example, we found $b_1 = 5$ and $s_{b_1} = 0.5803$.

If we want to develop a 99% confidence interval estimate of β_1, then from the previous slide we have $t_{\alpha/2} = t_{.005} = 3.355$.

Thus, the confidence interval is

$$b_1 \pm t_{.005} s_{b_1} = 5 \pm 3.355(0.5803) = 5 \pm 1.95$$

We are 99% confident that β_1 is between $5 - 1.95 = 3.05$ and $5 + 1.95 = 6.95$.

F Test for Significance in Simple Linear Regression

With only one independent variable, the *F* test is set up just like the *t* test and it will provide the same conclusion.

$$\boldsymbol{H_0}\boldsymbol{:}\ \boldsymbol{\beta_1 = 0}$$

$$\boldsymbol{H_a}\boldsymbol{:}\ \boldsymbol{\beta_1 \neq 0}$$

However, with more than one independent variable, only the *F* test can be used to test for an overall significant relationship of the regression.

The test statistic follows an *F* distribution with one degree of freedom at the numerator, and $n - 2$ degrees of freedom at the denominator.

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

Where, $MSR = SSR/(number\ of\ independent\ variables)$. Because in simple linear regression we have only one independent variable, it follows that, in this case, $MSR = SSR$.

F Test for the Armand's Pizza Parlor Example

Recollect that, for the Armand's Pizza Parlor case, we found $SSR = 14{,}200$ and $MSE = 191.25$.

To test for a significant relationship in the linear regression example, we use $\alpha = 0.01$.

Hence, $MSR = SSR = 14{,}200$, and the test statistic is

$$F = \frac{MSR}{MSE} = \frac{14{,}200}{191.25} = 74.25$$

In simple linear regression, the *p*-value is the upper tail of an F distribution with one degree of freedom at the numerator and $n - 2 = 8$ degrees of freedom at the denominator.

Area in upper tail	**0.10**	**0.05**	**0.025**	**0.01**
F **Value ($df_1 = 1, df_2 = 8$)**	3.46	5.32	7.57	11.26

Because $F = 74.25$ is greater than $F_{.01} = 11.26$, the values in the "Area in Upper Tail" row show that $p\text{–value} \leq 0.01$. Thus, we reject H_0 and conclude that $\beta_1 \neq 0$.

The same conclusion is reached using the critical value because $F = 74.25 \geq 11.26 = F_{0.01}$.

ANOVA Table for Simple Linear Regression

The ANOVA table for simple linear regression conveniently displays the following results.

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Squares	*F*	*p*-value
Regression	SSR	1	MSR = SSR / 1	MSR / MSE	$Pr(F \geq \text{MSR} / \text{MSE})$
Error	SSE	$n - 2$	MSE = SSE / $(n - 2)$		
Total	SST	$n - 1$			

Regression, Error, and Total are the labels for the three sources of variation, with SSR, SSE, and SST appearing as the corresponding sum of squares in the second column.

The ANOVA table with the *F* test computations performed for Armand's Pizza Parlors follow.

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Squares	*F*	*p*-value
Regression	14,200	1	14,200	74.25	0.0000
Error	1,530	8	191.25		
Total	17,530	9			

Some Cautions About the Interpretation of Significance Tests

1. Rejecting H_0: $\beta_1 = 0$ and concluding that the relationship between x and y is significant does not enable us to conclude that a cause-and-effect relationship is present between x and y.
   - Concluding a cause-and-effect relationship is warranted only if the analyst can provide some type of theoretical justification that the relationship is, in fact, causal.
   - For Armand's Pizza Parlors, we cannot necessarily conclude that changes in student population *cause* changes in quarterly sales.
2. Just because we are able to reject H_0: $\beta_1 = 0$ and demonstrate statistical significance does not enable us to conclude that there is a linear relationship between x and y.
   - Given a significant relationship, we can use with confidence the estimated regression equation only for predictions within the range of the x values observed in the sample.
   - For Armand's Pizza Parlors, we should use the regression equation only to predict sales for restaurants where the associated student population is between 2,000 and 26,000.

Using the Estimated Regression Equation for Estimation and Prediction

We use the estimated regression equation to predict an individual value of y for a given value of x. However, to provide information about the precision associated with a *point estimator of* $E(y^*)$ and *predictor* $\hat{y}$, we must develop confidence intervals and prediction intervals.

A confi**dence interval** is an interval estimate of the *mean value of* y for a given value of x.

A **prediction interval** is used whenever we want to *predict an individual value of* y for a new observation corresponding to a given value of x.

The margin of error is larger for a prediction interval.

For this section we will be using the following notation to help clarify matters:

x^* = the given value of the independent variable x.

y^* = the random variable for the possible values of the dependent variable y when $x = x^*$.

$\hat{y}^*$ = the point estimator of $E(y^*)$ and the predictor of an individual value of y^* when $x = x^*$.

Confidence Interval for the Mean Value of *y*

The form of a confidence interval for $E(y^*)$ is as follows.

$$\boldsymbol{\hat{y}^* \pm t_{\alpha/2} s_{\hat{y}^*}}$$

Where $1 - \alpha$ is the confidence coefficient, and $t_{\alpha/2}$ is based on a t distribution with $n - 2$ degrees of freedom.

$s_{\hat{y}^*}$ is the estimate of the standard deviation of $\hat{y}^*$ and it is calculated as follows.

$$\boldsymbol{s_{\hat{y}^*} = s\sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum(x_i - \bar{x})^2}}}$$

Note that the estimated standard deviation of $\hat{y}^*$ is smallest when $\boldsymbol{x^* = \bar{x}}$. In such case, we have

$$s_{\hat{y}^*} = s\sqrt{\frac{1}{n} + \frac{(\bar{x} - \bar{x})^2}{\sum(x_i - \bar{x})^2}} = s\sqrt{\frac{1}{n}}$$

Confidence Interval for the Mean Quarterly Sales at a Given Value of Student Population

Let us develop a 95% confidence interval of the mean quarterly sales for all Armand's restaurants located near campuses with 10,000 students ($x^* = 10$.)

From previous calculations, we found that $n = 10$, $s = 13.829$, and $\sum(x_i - \bar{x})^2 = 568$.

At $x^* = 10$, we have:

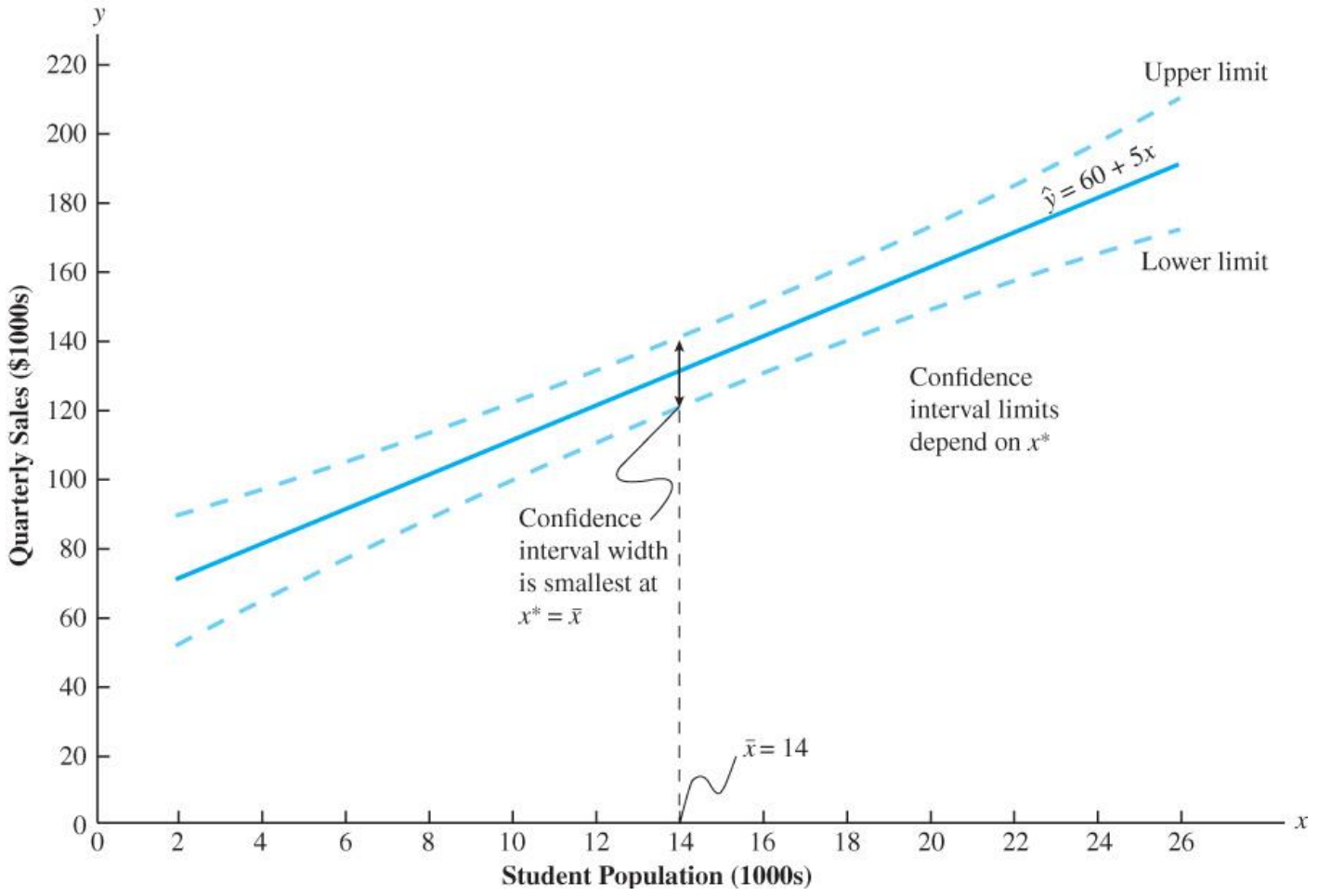
$$s_{\hat{y}^*} = s\sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum(x_i - \bar{x})^2}} = 13.829\sqrt{\frac{1}{10} + \frac{(10 - 14))^2}{568}} = 13.829\sqrt{0.1282} = 4.95$$

Using Table 2 of Appendix B, we have $t_{.025} = 2.306$, and at $x^* = 10$, the predictor is

$$\hat{y}^* = 60 + 5x^* = 60 + 5(10) = 110.$$

Thus, with $\hat{y}^* = 110$ and margin of error $t_{\alpha/2}s_{\hat{y}^*} = 2.306(4.95) = 11.415$, the 95% confidence interval for the mean quarterly sales is 110 ± 11.415.

Confidence Intervals for the Mean Value of *y* at Given Values of *x* (Armand's Pizza Example)



Prediction Interval for an Individual Value of *y*

The form of a prediction interval for y^* is as follows.

$$\hat{\boldsymbol{y}}^* \pm \boldsymbol{t_{\alpha/2} s_{pred}}$$

Where $1 - \alpha$ is the confidence coefficient, and $t_{\alpha/2}$ is based on a t distribution with $n - 2$ degrees of freedom.

s_{pred} is the the standard deviation corresponding to the prediction of the value of y when $x = x^*$.

$$\boldsymbol{s_{pred} = s\sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum (x_i - \bar{x})^2}}}$$

Note that s_{pred} is also smallest when $\boldsymbol{x^* = \bar{x}}$. In such case, we have

$$s_{\hat{y}^*} = s\sqrt{1 + \frac{1}{n} + \frac{(\bar{x} - \bar{x})^2}{\sum (x_i - \bar{x})^2}} = s\sqrt{1 + \frac{1}{n}}$$

Prediction Interval for the Quarterly Sales at an Individual Restaurant

Let us develop a 95% prediction interval of the quarterly sales for a new restaurant that Armand's is opening near a campus with 10,000 students ($x^* = 10$.)

From previous calculations, we found that $n = 10$, $s = 13.829$, and $\sum(x_i - \bar{x})^2 = 568$.

At $x^* = 10$, we have:

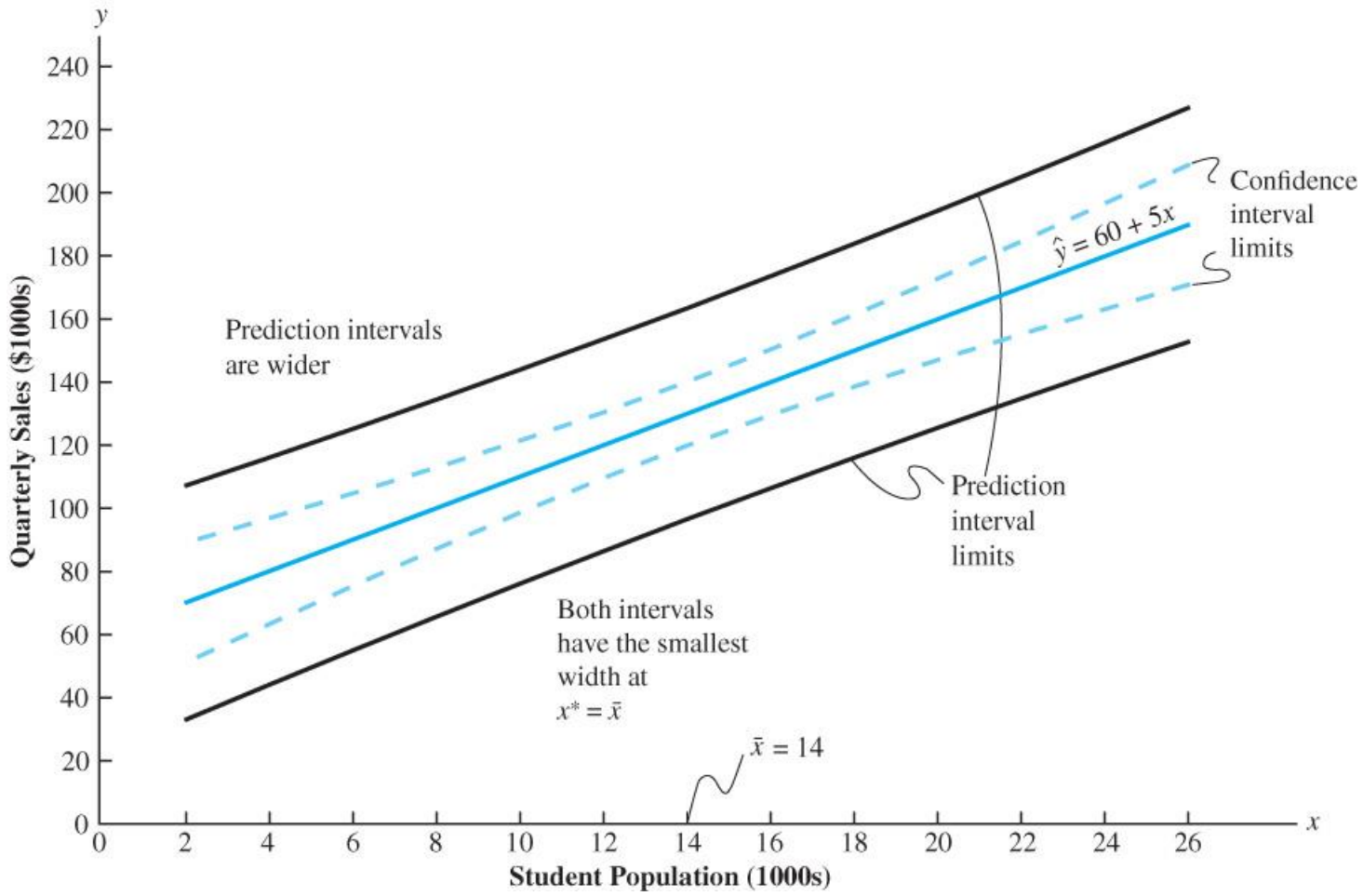
$$s_{\hat{y}^*} = s\sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum(x_i - \bar{x})^2}} = 13.829\sqrt{1 + \frac{1}{10} + \frac{(10 - 14))^2}{568}} = 13.829\sqrt{1.282} = 14.69$$

Using Table 2 of Appendix B, we have $t_{.025} = 2.306$, and at $x^* = 10$, the predictor is

$$\hat{y}^* = 60 + 5x^* = 60 + 5(10) = 110.$$

Thus, with $\hat{y}^* = 110$ and margin of error $t_{\alpha/2}s_{pred} = 2.306(14.69) = 33.875$, the 95% prediction interval for the quarterly sales at the new restaurant is 110 ± 33.875.

Confidence and Prediction Intervals for *y* at Given Values of *x* (Armand's Pizza Example)



Excel Output for the Armand's Pizza Parlors Problem

Performing regression analysis computations without the help of a computer can be quite time-consuming.

Most statistical software packages, such as Excel, MiniTab or JMP, have regression tools that can be used to perform a complete regression analysis.

In a typical Excel output for the Armand's Pizza example shown to the right, the independent variable was named *Population*, and the dependent variable was named *Sales,* so that:

$$\widehat{\text{Sales}} = 60 + 5\ \text{Population}$$

Regression Statistics	
Multiple R	95.01%
R Square	90.27%
Adj. R Square	89.06%
Standard Error	13.829
Observations	10

ANOVA

	df	SS	MS	F	P-Value
Regression	1	14200	14200	74.248	0.000
Residual	8	1530	191.25		
Total	9	15730			

	Coefficients	Std. Err.	t Stat	P-value	Lower 95%	Upper 95%
Intercept	60	9.226	6.503	0.000	38.725	81.275
Population	5	0.580	8.617	0.000	3.662	6.338

Residual Analysis

If the assumptions about the error term ϵ appear questionable, the hypothesis tests about the significance of the regression relationship and the interval estimation results may not be valid.

The **residual** for observation i is defined as:

$$\boldsymbol{y_i - \hat{y}_i}$$

Where y_i and $\hat{y}_i$ are the observed and predicted values of the dependent variable.

The residuals provide the best information about ϵ, and much of the residual analysis is based on an examination of graphical plots.

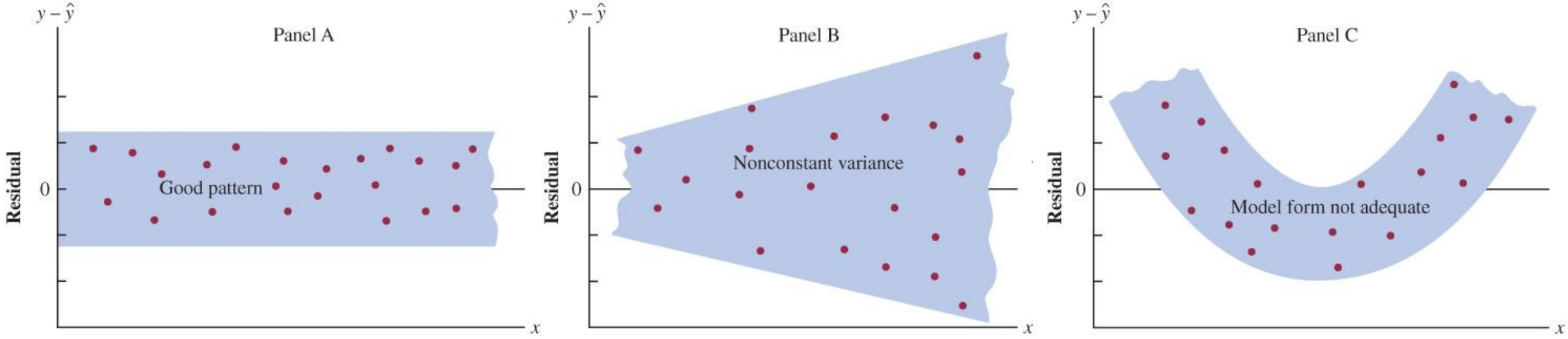
In this section, we discuss the following residual plots:

1. A plot of the residuals against values of the independent variable x
2. A plot of residuals against the predicted values of the dependent variable, $\hat{y}$
3. A standardized residual plot
4. A normal probability plot

Residual Plot Against *x*

The **residual plot** against x represents the values of the independent variable on the horizontal axis and the residual values on the vertical axis. The residual plot gives an overall impression of a horizontal band of points (Panel A) if the following assumptions are valid:

1. the variance of ϵ is the same for all values of x.
   Implication: non-consistent variance may generate the pattern displayed in Panel B.
2. the regression model adequately represents the relationship between the variables.
   Implication: an inadequate model form may create the pattern displayed in Panel C.

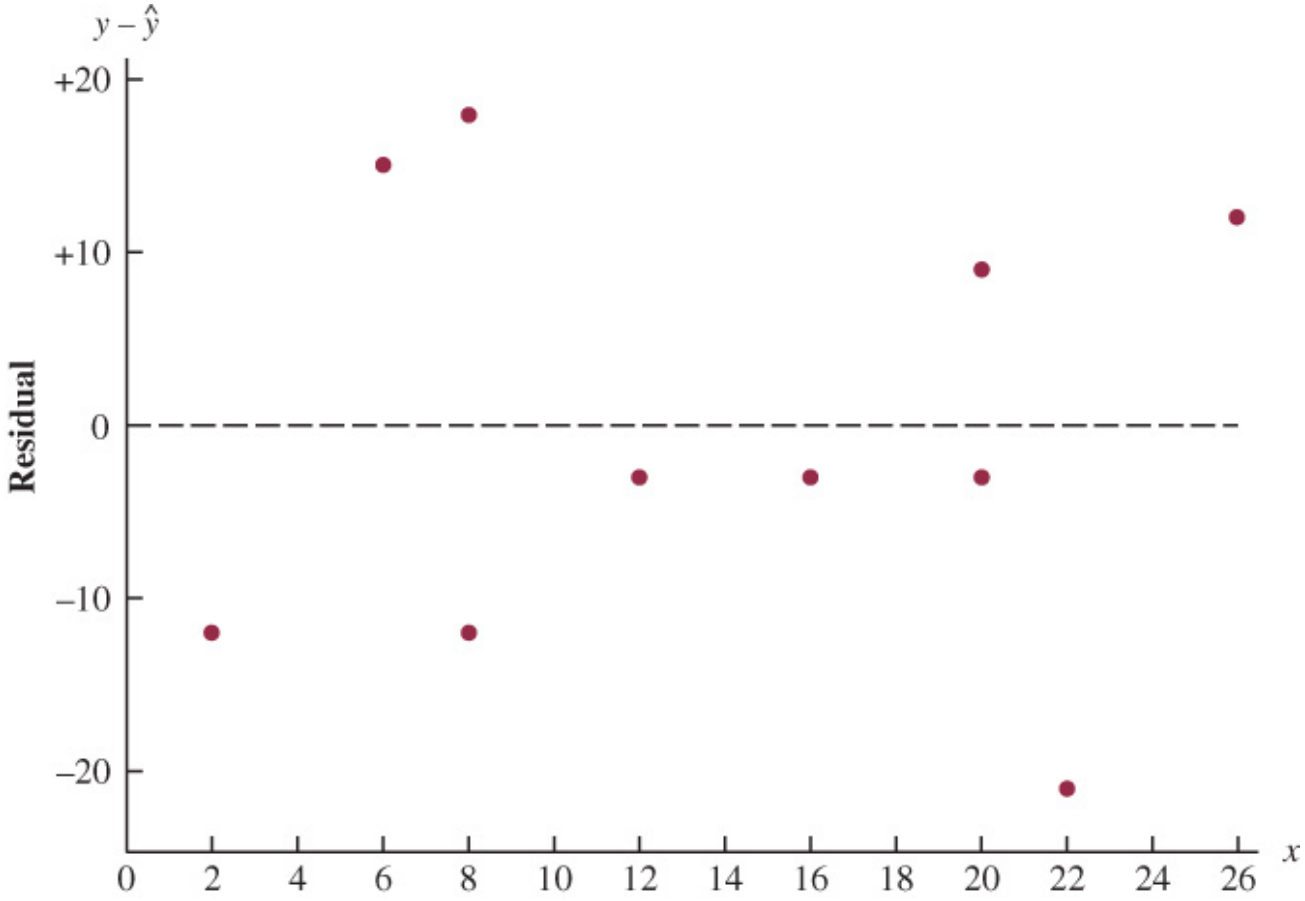


Residual Plot Against *x* for Armand's Pizza Parlors

The residuals for the Armand's Pizza Parlor sample data appear to approximate the horizontal pattern in Panel A from the previous slide.

Hence, we conclude that the residual plot does not provide evidence that the assumptions made for Armand's regression model should be challenged.

At this point, we are confident in the conclusion that Armand's simple linear regression model is valid.

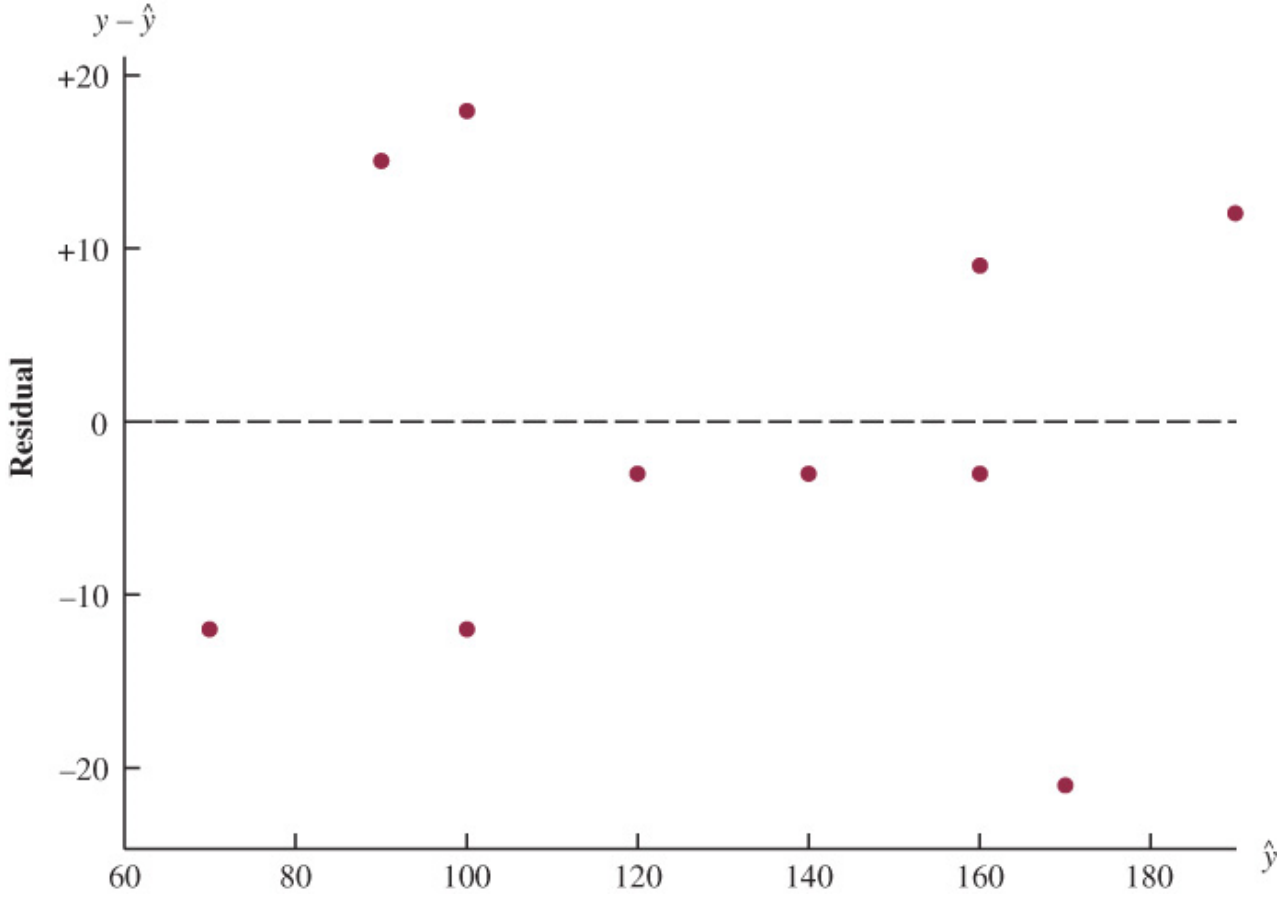


Residual Plot Against $\hat{y}$ for Armand's Pizza Parlors

Another residual plot represents the predicted value of the dependent variable $\hat{y}$ on the horizontal axis in place of the independent variable x.

The residual plot against $\hat{y}$ is used for multiple regression analysis in the presence of more than one independent variable.

The pattern of the residual plot against $\hat{y}$ is the same as the pattern of the residual plot against x. Thus, it is not a pattern that would lead us to question the model assumptions.



Standardized Residuals

Many of the residual plots provided by computer software packages use a standardized version of the residuals.

The **standardized residual for observation i** is obtained by dividing residual i by its standard deviation.

$$\frac{y_i - \hat{y}_i}{s_{y_i - \hat{y}_i}}$$

Where $s_{y_i - \hat{y}_i}$ is the **standard deviation of residual i**.

$$s_{y_i - \hat{y}_i} = s\sqrt{1 - h_i}$$

Where s is the standard error of the estimate, and h_i is the **leverage of residual i** (*see notes.)

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum(x_i - \bar{x})^2}$$

Standardized Residuals for Armand's Pizza Parlors

The table below shows the calculation of the standardized residuals for the Armand's Pizza Parlors example. Recall that previous calculations showed $s = 13.829$.

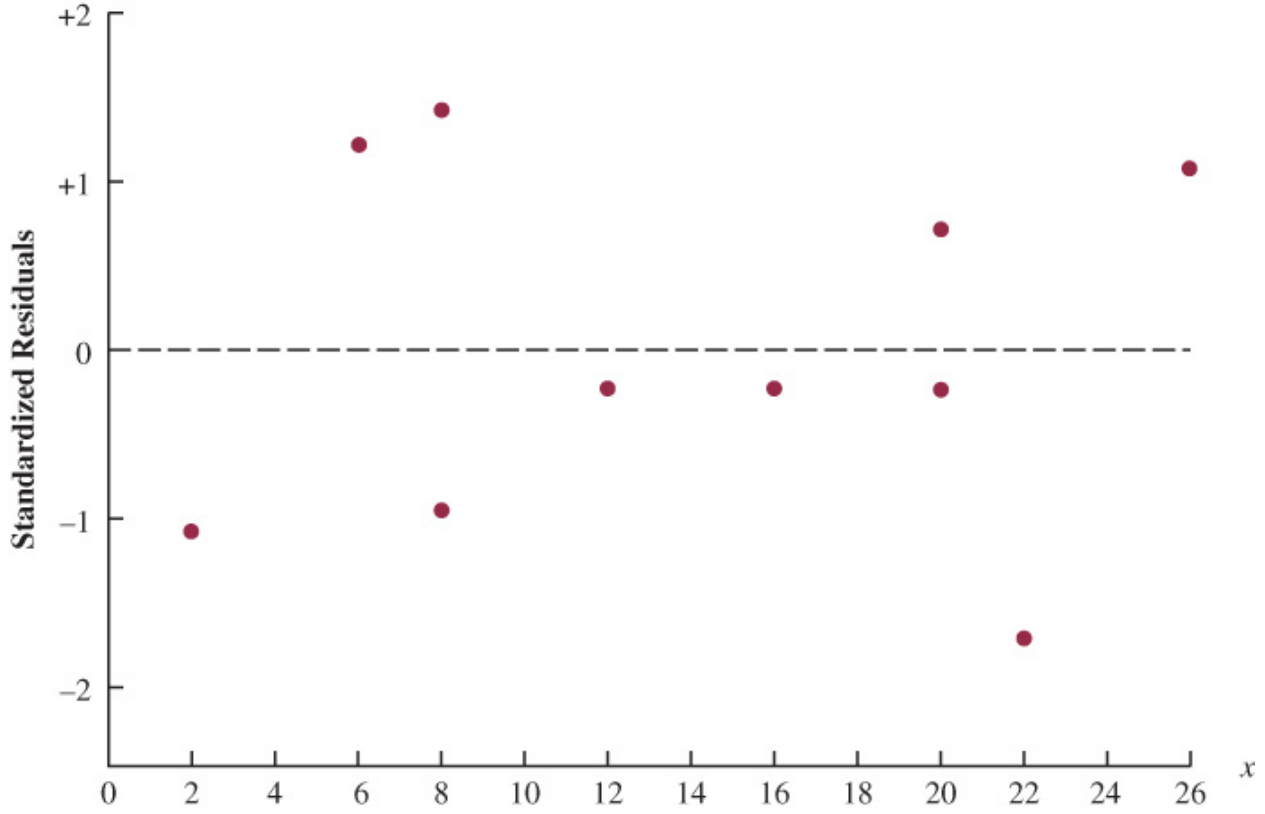
Due to their computational complexity, statistical packages provide estimated values of $\hat{y}$, the residuals, and the standardized residuals as an optional regression output.

Restaurant i	x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	$\frac{(x_i - \bar{x})^2}{\sum(x_i - \bar{x})^2}$	h_i	$s_{y_i - \hat{y}_i}$	$y_i - \hat{y}_i$	Standardized Residual
1	2	-12	144	0.2535	0.3535	11.1193	-12	-1.0792
2	6	-8	64	0.1127	0.2127	12.2709	15	1.2224
3	8	-6	36	0.0634	0.1634	12.6492	-12	-0.9487
4	8	-6	36	0.0634	0.1634	12.6492	18	1.4230
5	12	-2	4	0.0070	0.1070	13.0682	-3	-0.2296
6	16	2	4	0.0070	0.1070	13.0682	-3	-0.2296
7	20	6	36	0.0634	0.1634	12.6492	-3	-0.2372
8	20	6	36	0.0634	0.1634	12.6492	9	0.7115
9	22	8	64	0.1127	0.2127	12.2709	-21	-1.7114
10	26		144	0.2535	0.3535	11.1193	12	1.0792
			568					

Standardized Residual Plot Against *x* for Armand's Pizza Parlors

If the error term ϵ follows a normal distribution, the distribution of the standardized residuals should appear as a standard normal probability distribution.

Thus, on a standardized residual plot, we should expect to see approximately 95% of the standardized residuals between −2 and +2, as it is the case for the standardized residual plot of the Armand's Pizza Parlor example.



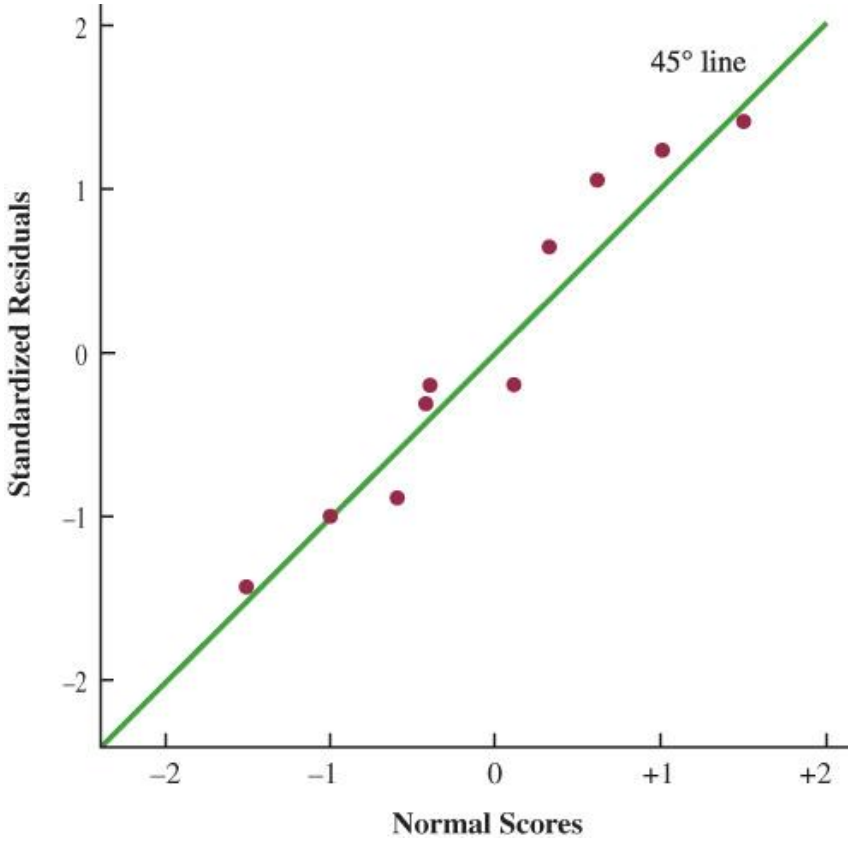
Normal Probability Plot

A **normal probability plot** displays the ordered standardized residuals against the values of the *normal scores.*

With a linear pattern close to the 45º line, we can assume that the error term has a normal probability distribution.

Below are the 10 observations for the Armand's Pizza Parlor normal probability plot that is shown to the right.

Order Statistic	Normal Scores	Ordered Standardized Residuals
1	-1.55	-1.7114
2	-1.00	-1.0792
3	-0.65	-0.9487
4	-0.37	-0.2372
5	-0.12	-0.2296
6	0.12	0.2296
7	0.37	0.7115
8	0.65	1.0792
9	1.00	1.2224
10	1.55	1.4230



Detecting Outliers

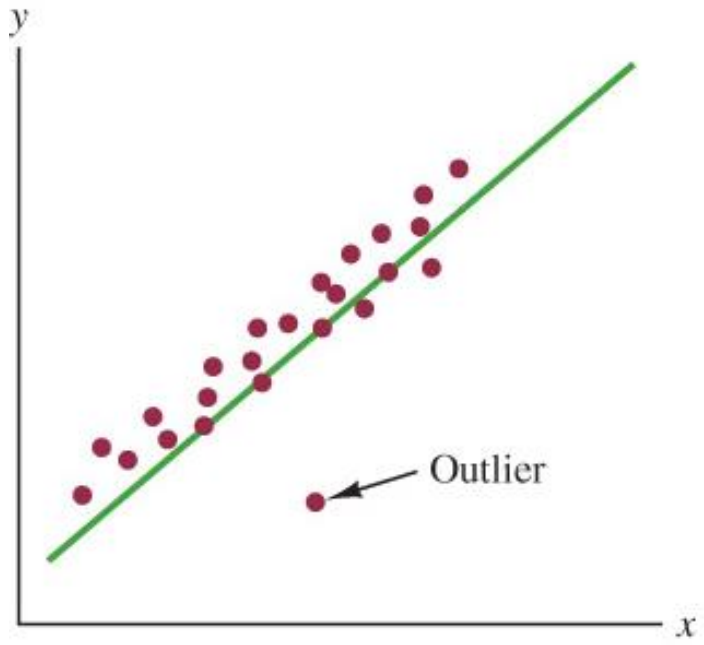
An **outlier** is a data point (observation) that does not fit the trend shown by the remaining data.

Outliers represent observations that are suspect and warrant careful examination.

The presence of an outlier may be due to one of the following reasons:

- Erroneous data.
  Action: the data should be corrected.
- Signal of a violation in the model assumptions
  Action: another model should be considered.
- Unusual values that occurred by chance
  Action: the data should be retained.

Standardized residuals can also be used to identify outliers. If an observation deviates greatly from the pattern of the remaining data, the corresponding standardized residual will be large in absolute value.



Data Set with an Outlier

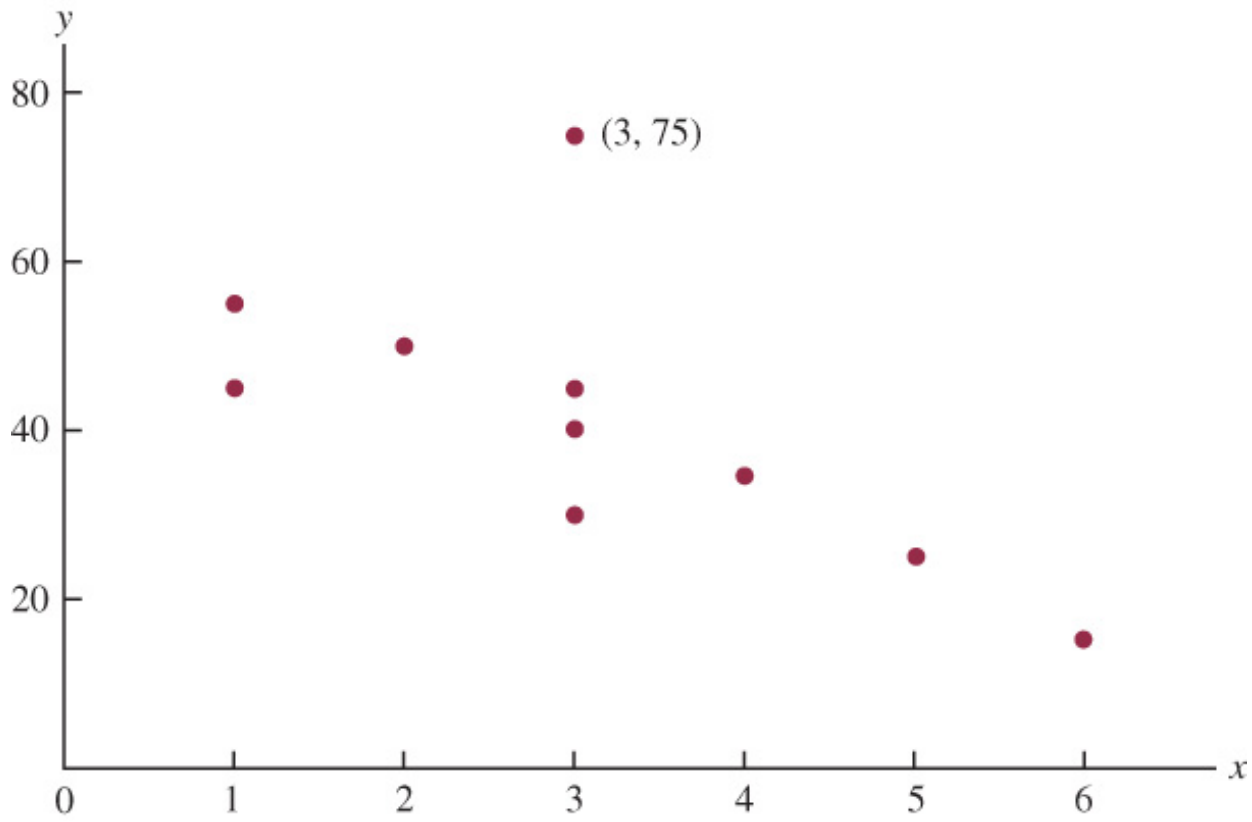
For the case of simple linear regression, one can often detect outliers by simply examining the scatter diagram.

Consider the data set displayed in the scatter diagram to the right:

> Except for observation 4 ($x_4 = 3$, $y_4 = 75$), the pattern suggests a negative linear relationship.

As stated before, standardized residuals may also be used to identify outliers.

Observation 4 has a standardized residual of 2.68, as shown in the next slide's regression computer output.



Regression Output for the Outlier Data Set

Regression Statistics	
Multiple R	70.48%
R Square	49.68%
Adj. R Sq.	43.39%
Std. Error	12.670
Observations	10

ANOVA

	df	*SS*	*MS*	*F*	*P-Value*
Regression	1	1268.2	1268.2	7.90	0.023
Residual	8	1284.3	160.5		
Total	9	2552.5			

	Coefficients	*Std. Err.*	*t Stat*	*P-value*
Intercept	64.96	9.26	7.02	0.000
x	-7.33	2.60	-2.81	0.023

Regression Equation

y = 64.96 -7.33 x

Observation	Predicted y	Residuals	Standard Residuals
1	57.6271	-12.6271	-1.0570
2	57.6271	-2.6271	-0.2199
3	50.2966	-0.2966	-0.0248
4	42.9661	32.0339	2.6816
5	42.9661	-2.9661	-0.2483
6	42.9661	2.0339	0.1703
7	42.9661	-5.6356	-0.4718
8	35.6356	-0.6356	-0.0532
9	28.3051	-3.3051	-0.2767
10	20.9746	-5.9746	-0.5001

Regression Output for the Revised Outlier Data Set

In deciding how to handle an outlier, we should first check to see whether it is a valid observation.

Suppose that in checking the outlier data set, we find an error for observation 4, with the correct value being ($x_4 = 3$, $y_4 = 30$).

The regression output to the right, obtained after correction of the value of y_4, shows how using the correct data value substantially improved the goodness of fit.

- r^2 increased from 49.68% to 83.80%
- s decreased from 12.6704 to 5.2481
- b_0 changed from 64.96 to 59.24
- b_1 changed from -7.33 to -6.95

Regression Statistics	
Multiple R	91.54%
R Square	83.80%
Adj. R Sq.	81.77%
Std. Error	5.2481
Observations	10

ANOVA

	df	*SS*	*MS*	*F*	*P-Value*
Regression	1	1139.66	1139.66	41.38	0.000
Residual	8	220.34	27.54		
Total	9	1360.00			

	Coefficients	*Std. Err.*	*t Stat*	*P-value*
Intercept	59.24	3.83	15.45	0.000
x	-6.95	1.08	-6.43	0.000

Detecting Influential Observations

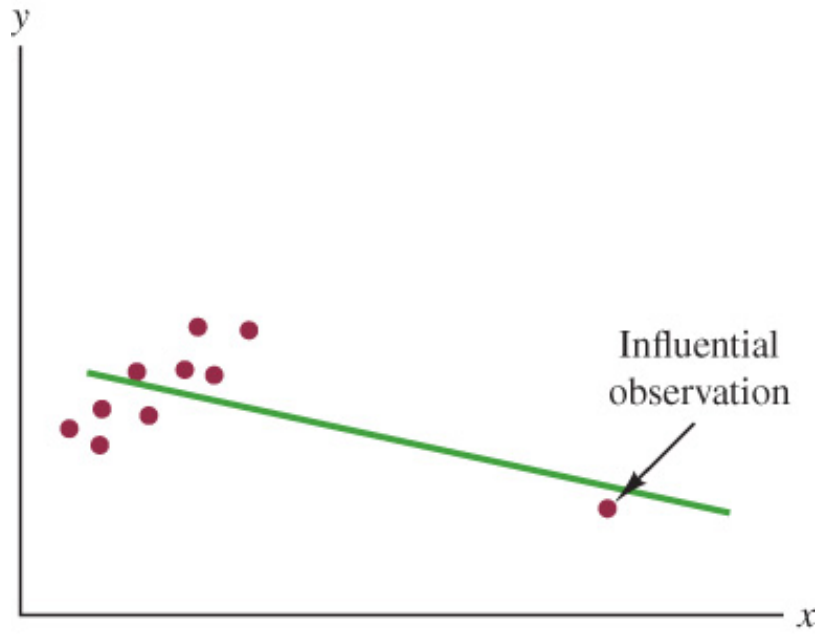
An **influential observation** exerts a strong influence on the results of linear regression.

The estimated regression line shown in the diagram has a negative slope. However, if the influential observation were dropped from the data set, the slope of the estimated regression line would change from negative to positive and the y-intercept would be smaller.

Influential observations may be identified from a scatter diagram when only one independent variable is present.

An influential observation may be:

- an outlier (an off-trend y value)
- **a high leverage point**: an observation with an extreme x value that is far away from $\bar{x}$.
- a combination of the two (a somewhat off-trend y value and a somewhat extreme x value)



Data Set with a High Leverage Observation

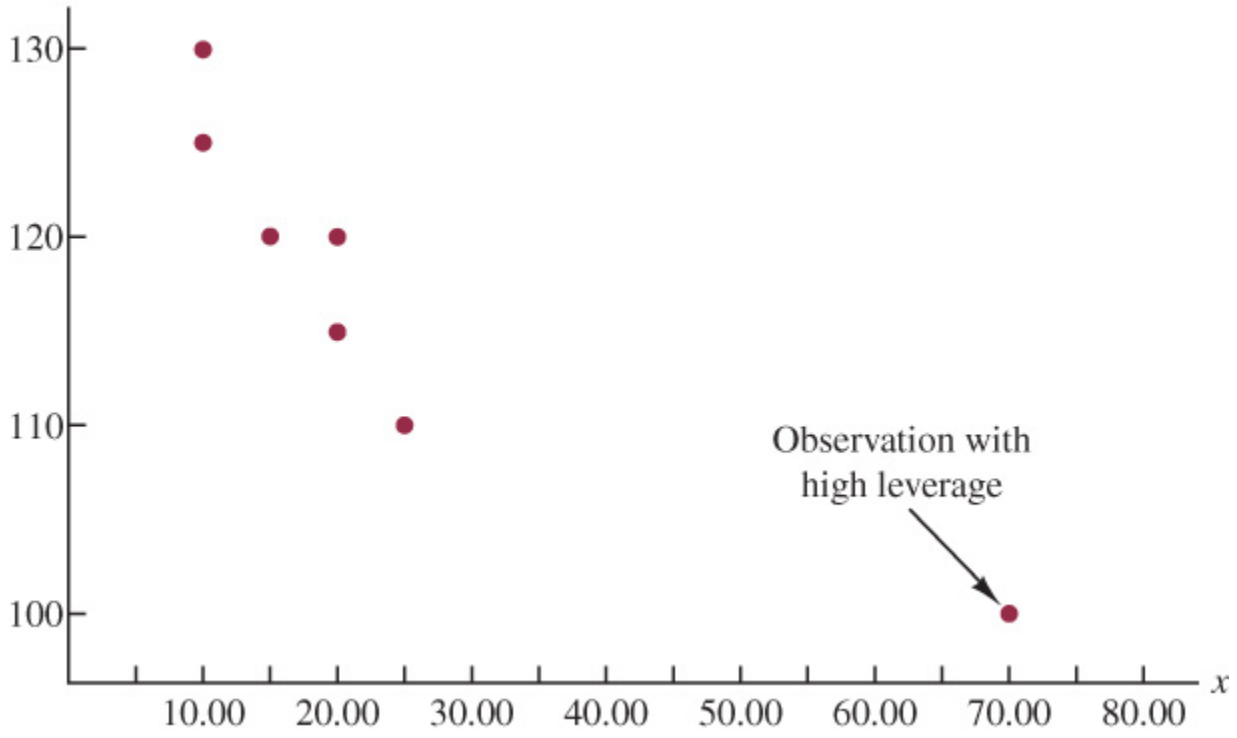
For the single-independent-variable case, the leverage of observation i, denoted h_i, can be computed by using the following equation.

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum(x_i - \bar{x})^2}$$

From the formula, it is clear that the farther x_i is from its mean $\bar{x}$, the higher the leverage of observation i.

Many statistical packages automatically identify observations with high leverage as part of the standard regression output.

As an illustration of points with high leverage, let us consider observation 7 at coordinates ($x_7 = 70, y_7 = 100$).



Calculation of h_i in the High Leverage Data Set

The leverage for observation ($x_7 = 70$, $y_7 = 100$) of the high leverage data set is computed as follows.

$$h_7 = \frac{1}{n} + \frac{(x_7 - \bar{x})^2}{\sum(x_i - \bar{x})^2} = \frac{1}{7} + \frac{(70 - 24.286)^2}{2{,}621.43} = 0.94$$

For the case of simple linear regression, observation i has high leverage if $h_i > 6/n$ or 0.99, whichever is smaller.

For the high leverage data set, $6/n = 6/7 = 0.86$. Because $h_7 = 0.94 > 0.86$, we will identify observation 7 as a highly influential observation due to an extreme value of x.

Influential observations that are caused by an interaction of large residuals and high leverage can be difficult to detect.

Diagnostic procedures are available that take into account when an observation is influential. One such measure, called Cook's *D* statistic, is discussed in the next chapter.

Practical Advice: Big Data and Hypothesis Testing in Simple Linear Regression

In simple linear regression, as the sample size increases,

- the *p*-value for the *t* test, used to determine whether a significant relationship exists between the dependent and the independent variables, decreases.
- the confidence interval for the slope parameter associated with the independent variable narrows.
- the confidence interval for the mean value of y narrows.
- the prediction interval for an individual value of y narrows.

The fact that with larger sample sizes it becomes more likely to reject H_0, and the precision of estimates and prediction increases, does not necessarily make these results more reliable.

No matter how large is the sample used to estimate the simple linear regression equation, we must be concerned about the potential presence of nonsampling error in the data, and carefully consider whether a random sample of the population of interest has been taken.