# RWTH BUSINESS SCHOOL

Mathematics & Statistics
M.Sc. Data Analytics and Decision Science

Prof. Dr. Thomas S. Lontzek

CENGAGE

© 2021 Cengage Learning. All Rights Reserved. May not be scanned, copied or duplicated, or posted to a publicly accessible website, in whole or in part.

BUSINESS SCHOOL | RWTH AACHEN UNIVERSITY

# Simple Linear Regression

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

# Simple Linear Regression

- Simple linear regression 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.

# Simple Linear Regression Model

- The equation that describes how *y* is related to *x* and an error term is called the regression model.
- The simple linear regression model is:

$$y = \beta_0 + \beta_1 x + \varepsilon$$

where:

$\beta_0$ and $\beta_1$ are called parameters of the model,

$\varepsilon$ is a random variable called the error term.

# Simple Linear Regression Equation

- The Simple Linear Regression Equation is:

$$E(y) = \beta_0 + \beta_1 x$$

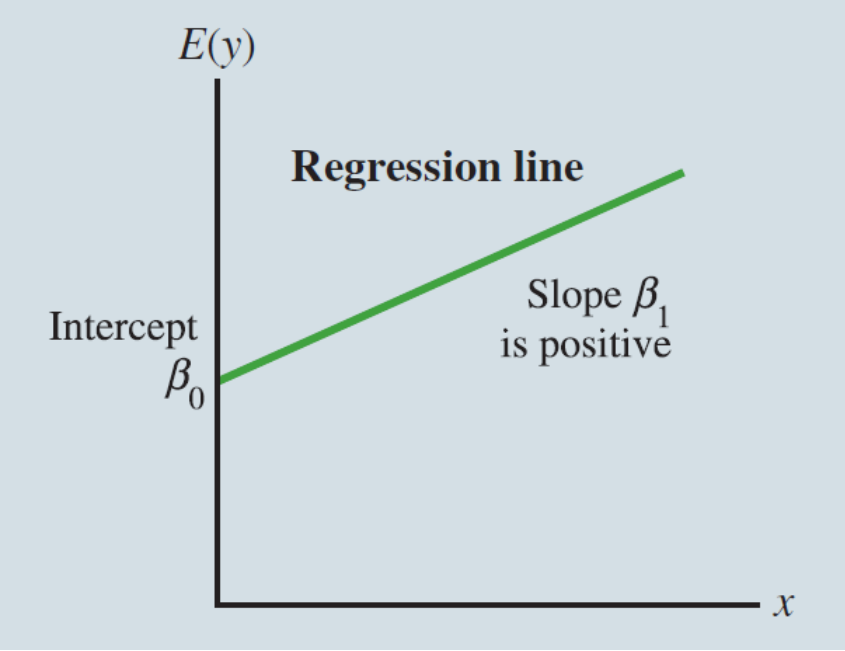
Graph of the regression equation is a straight line.

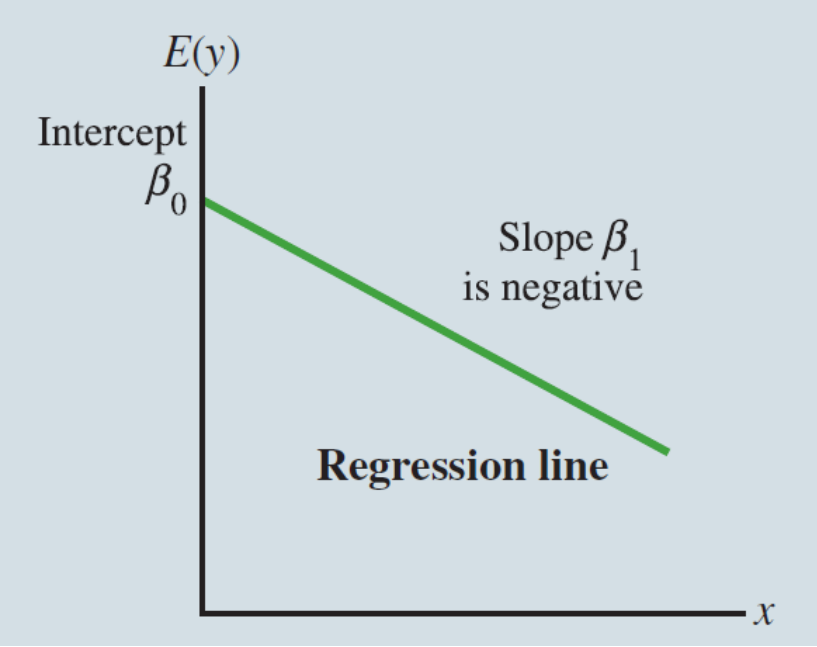
$\beta_0$ is the $y$ intercept of the regression line.

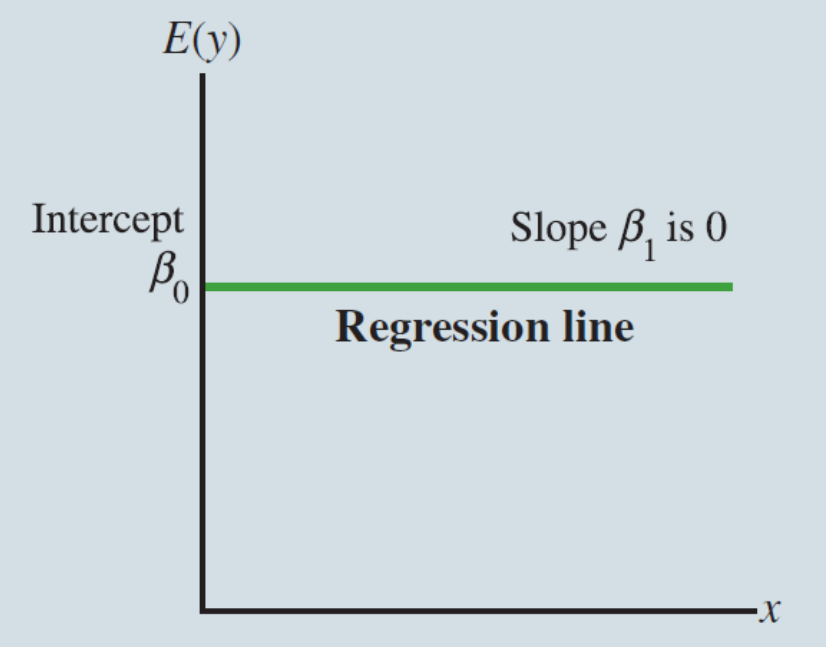
$\beta_1$ is the slope of the regression line.

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

# Simple Linear Regression Equation



# Estimated Simple Linear Regression Equation

The estimated simple linear regression equation

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

The graph is called the estimated regression line.

$b_0$ is the $y$ intercept of the line.

$b_1$ is the slope of the line.

$\hat{y}$ is the estimated value of $y$ for a given $x$ value.

# Least Squares Method

- Least Squares Criterion

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

where:

$y_i$ = observed value of the dependent variable for the $i$th observation.

$\hat{y}$ = estimated value of the dependent variable for the $i$th observation.

# Least Squares Method

- Slope for the Estimated Regression Equation

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

where:

$x_i$ = value of independent variable for $i$th observation

$y_i$ = value of dependent variable for $i$th observation

$\bar{x}$ = mean value for dependent variable

$\bar{y}$ = mean value for independent variable

- $y$-Intercept for the Estimated Regression Equation

$$b_0 = \bar{y} - b_1\bar{x}$$

# Simple Linear Regression

**Example:** Armand's Pizza Parlor Restaurants

Data was collected from a sample of 10 Armand's Pizza Parlor Restaurants near college campuses. For the *i*th observation or restaurant in the sample, $x_i$ is the size of the student population and $y_i$ is the quarterly sales.

# Simple Linear Regression

**Example:** Armand's Pizza Parlor Restaurants

| Restaurant | Student population (1000s) | Quarterly sales ($1000s) |
|---|---|---|
| 1 | 2 | 58 |
| 2 | 6 | 105 |
| 3 | 8 | 88 |
| 4 | 8 | 118 |
| 5 | 12 | 117 |
| 6 | 16 | 137 |
| 7 | 20 | 157 |
| 8 | 20 | 169 |
| 9 | 22 | 149 |
| 10 | 26 | 202 |

# Simple Linear Regression

**Example**:
Armand's
Pizza Parlor
Restaurants

| $i$ | $X_i$ | $Y_i$ | $x_i - \bar{x}$ | $y_i - \bar{y}$ | $(x_i - \bar{x})(y_i - \bar{y})$ | $(x_i - \bar{x})^2$ |
|---|---|---|---|---|---|---|
| 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 |
| Totals | **140** | **1300** | | | **2840** | **568** |

# Simple Linear Regression

**Example:** Armand's Pizza Parlor Restaurants

- Slope for the Estimated Regression Equation

$$b_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{2840}{568} = 5$$

- *y*-Intercept for the Estimated Regression Equation

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

- Estimated Regression Equation

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

# Estimated Regression Equation

**Example:** Armand's Pizza Parlor Restaurants

- Excel Worksheet

| | A | B | C | D |
|---|---|---|---|---|
| 1 | Restaurant | Population | Sales | |
| 2 | 1 | 2 | 58 | |
| 3 | 2 | 6 | 105 | |
| 4 | 3 | 8 | 88 | |
| 5 | 4 | 8 | 118 | |
| 6 | 5 | 12 | 117 | |
| 7 | 6 | 16 | 137 | |
| 8 | 7 | 20 | 157 | |
| 9 | 8 | 20 | 169 | |
| 10 | 9 | 22 | 149 | |
| 11 | 10 | 26 | 202 | |

# Using Excel's Chart Tools for Scatter Diagram & Estimated Regression Equation (1 of 4)

**Example:** Armand's Pizza Parlor Restaurants

- Producing a Scatter Diagram
  - Step 1 Select cells B2:C11
  - Step 2 Click the **Insert** tab on the Ribbon
  - Step 3 In the Charts group, click **Insert Scatter (X,Y) or Bubble Chart**
  - Step 4 When the list of scatter diagram subtypes appears,

    Click **Scatter** (chart in upper left corner)

# Using Excel's Chart Tools for Scatter Diagram & Estimated Regression Equation (2 of 4)

**Example:** Armand's Pizza Parlor Restaurants

- Editing a Scatter Diagram
  - Step 1 Click the **Chart Title** and replace it with *Armand's Pizza Parlors*
  - Step 2 Click the **Chart Elements** button
  - Step 3 When the list of chart elements appears:

    Click **Axis Titles** (creates placeholders for titles)

    Click **Gridlines** (to deselect gridlines option)

    Click **Trendline**

# Using Excel's Chart Tools for Scatter Diagram & Estimated Regression Equation (3 of 4)

**Example:** Armand's Pizza Parlor Restaurants

- Editing a Scatter Diagram (continued)
  - Step 4 Click the horizontal **Axis Title** and replace it with *Student population (1000s)*
  - Step 5 Click the **Vertical (Value) Axis Title** and replace it with *Quarterly Sales ($1000s)*
  - Step 6 Select the **Format Trendline** option
  - Step 7 When the Format Trendline dialog box appears:

    Select **Display equation on chart**

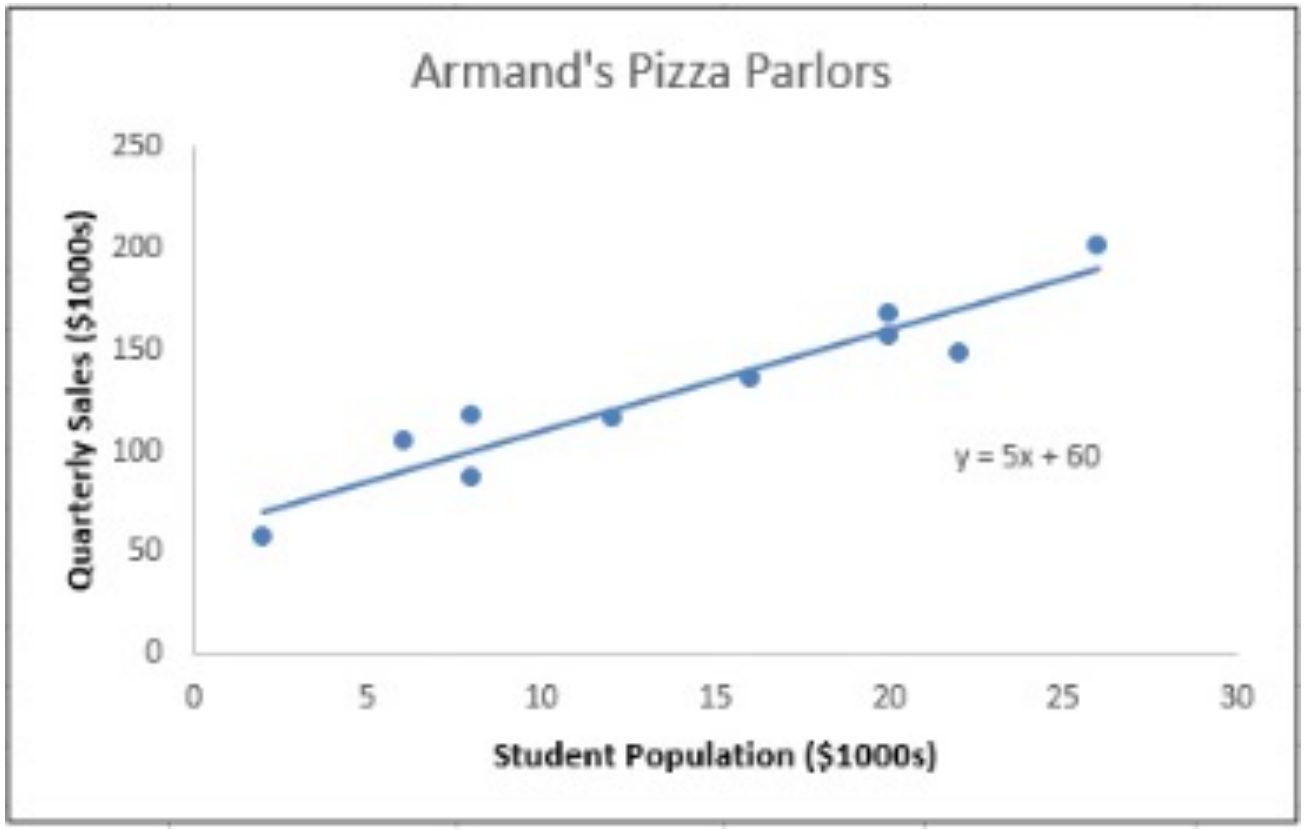
    Click the **Fill & Line** button

    In the **Dash type** box, select **Solid**

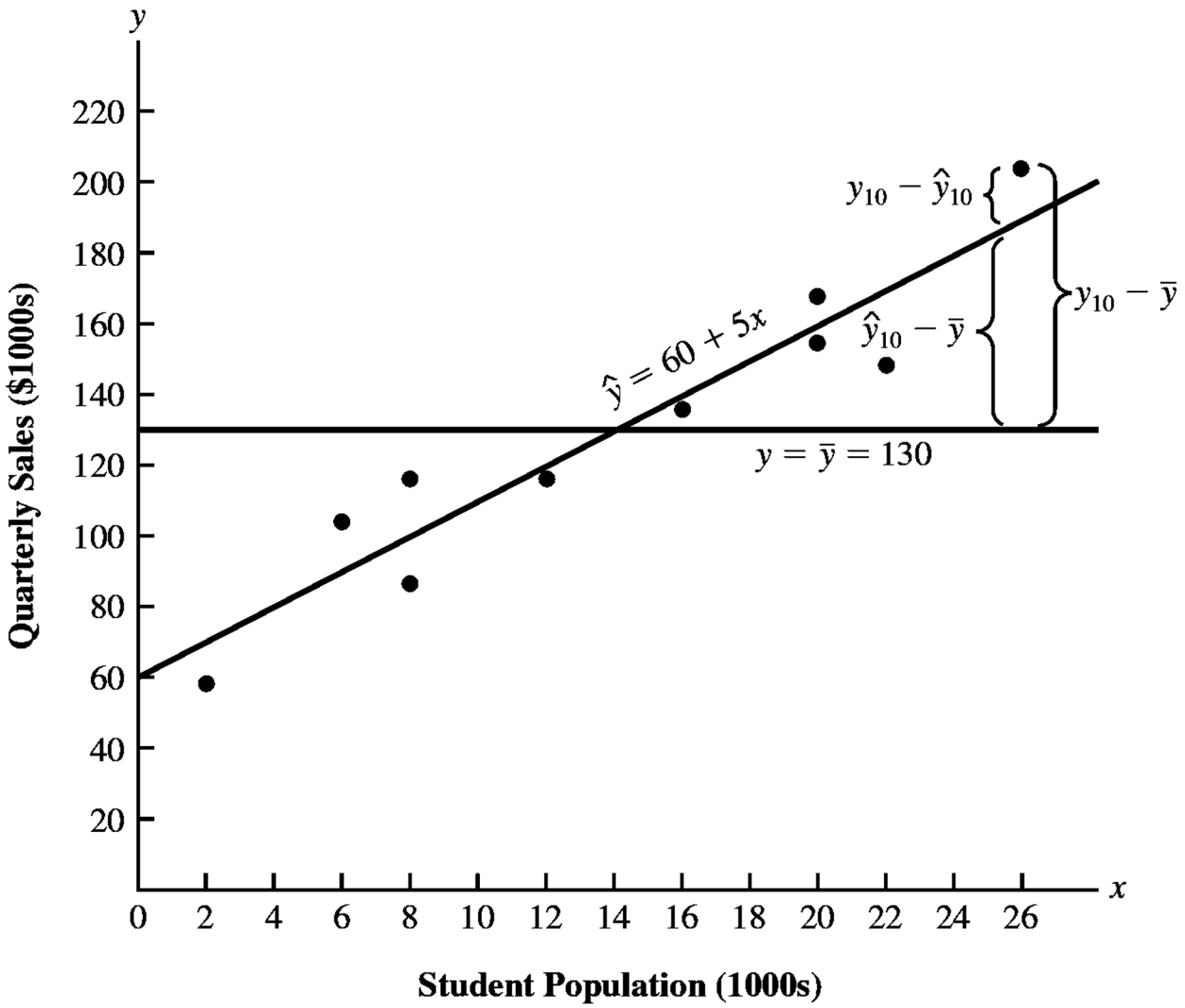
    Close the **Format Trendline** dialog box

# Using Excel's Chart Tools for Scatter Diagram & Estimated Regression Equation (4 of 4)

**Example:** Armand's Pizza Parlor Restaurants



# Deviations about the estimated regression line and the y mean



# SSE – Sum of Squared Errors (residuals)

The value of SSE is a measure of the error in using the estimated regression equation to predict the values of the dependent variable in the sample.

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

| 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 = 1530 |

# SST – Total Sum of Squares

The value of SST is a measure of the error in using the mean of the dependent variable to predict the values of the dependent variable in the sample.

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

SST < SSE
15,730 < 1,530

| 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 | 5184 |
| 2 | 6 | 105 | −25 | 625 |
| 3 | 8 | 88 | −42 | 1764 |
| 4 | 8 | 118 | −12 | 144 |
| 5 | 12 | 117 | −13 | 169 |
| 6 | 16 | 137 | 7 | 49 |
| 7 | 20 | 157 | 27 | 729 |
| 8 | 20 | 169 | 39 | 1521 |
| 9 | 22 | 149 | 19 | 361 |
| 10 | 26 | 202 | 72 | 5184 |
| | | | | SST = 15,730 |

# SSR – Sum of Squares due to Regression

To measure how much the $\hat{y}$ values on the estimated regression line deviate from $\bar{y}$, another sum of squares is computed. This sum of squares, called the *sum of squares due to regression,* is denoted SSR.

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

The total sum of squares can be partitioned into two components, the sum of squares due to regression and the sum of squares due to error. Hence, if the values of any two of these sum of squares are known, the third sum of squares can be computed easily. For instance, in the Armand's Pizza Parlors example, we already know that SSE = 1530 and SST = 15,730; therefore, solving for SSR in equation we find that the sum of squares due to regression is SSR = 14,200.

# SSR

- Relationship Among SST, SSR, SSE

$$SST = SSR + SSE$$

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

where:

SST = total sum of squares

SSR = sum of squares due to regression

SSE = sum of squares due to error

SSR can be thought of as the explained portion of SST, and SSE can be thought of as the unexplained portion of SST.

# Coefficient of Determination

- The coefficient of determination tells us how much of the variation in the dependent variable can be explained if we include the independent variable.

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

where:

SSR = sum of squares due to regression

SST = total sum of squares

# Coefficient of Determination

**Example:** Armand's Pizza Parlor Restaurants

$$r^2 = \frac{\text{SSR}}{\text{SST}} = \frac{14{,}200}{15{,}730} = .9027$$

The regression relationship is very strong; 90.27% of the variability in the sales can be explained by the linear relationship between the size of the student population and sales.

# Using Excel to Compute the Coefficient of Determination (1 of 2)

- Adding $r^2$ Value to Scatter Diagram
  - Step 1 Right-click on the trendline and select the

    **Format Trendline** option
  - Step 2 When the Format Trendline dialog box appears:

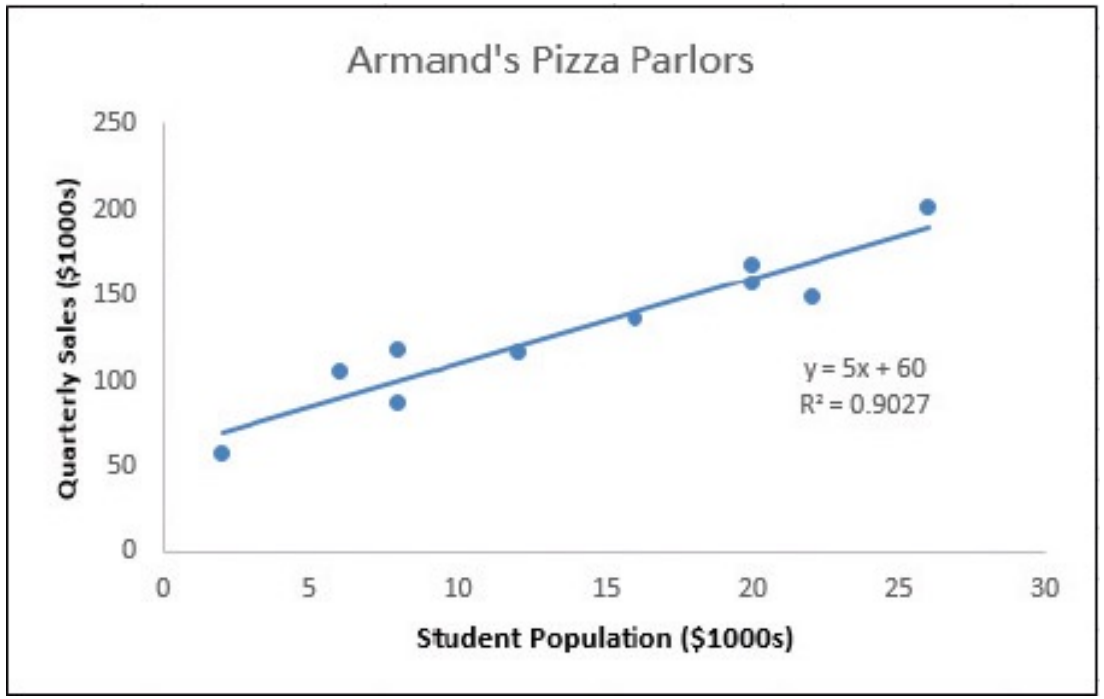
    Select **Display R-squared on chart**

    Close the **Format Trendline** dialog box

# Using Excel to Compute the Coefficient of Determination (2 of 2)

**Example:** Armand's Pizza Parlor Restaurants

- Adding $r^2$ Value to Scatter Diagram



# $R^2$ - Problems and Solutions

- Problem 1: $\boldsymbol{R^2}$ (most likely) increases each time we include a variable.
- Solution: use $R^2$ adjusted if >1 independent (explanatory) variable

- Problem 2: $R^2$ close to 1 is great – but we need a measure of $R^2$
- Solution: p value (F-test)

$$R^2 = \frac{\text{Variation in y explained by x}}{\text{Variation in y without considering x}}$$

$$F = \frac{\text{Variation in y explained by x}}{\text{Variation in y not explained by x}}$$


© 2021 Cengage Learning. All Rights Reserved. May not be scanned, copied or duplicated, or posted to a publicly accessible website, in whole or in part.

# Sample Correlation Coefficient

$$r_{xy} = (\text{sign of } b_1)\sqrt{\text{Coefficient of Determination}}$$

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

where:

$b_1$ = the slope of the estimated regression equation $\hat{y} = b_0 + b_1 x$

# Sample Correlation Coefficient

**Example:** Armand's Pizza Parlor Restaurants

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

The sign of $b_1$ in the equation $\hat{y} = 10 + 5x$ is "+".

$$r_{xy} = +\sqrt{.9027}$$

$$r_{xy} = +.9501$$

The sample correlation coefficient is restricted to a linear relationship between two variables, the coefficient of determination can be used for nonlinear relationships and for relationships that have two or more independent variables. Thus, the coefficient of determination provides a wider range of applicability.

# Testing for Significance

- To test for a significant regression relationship, we must conduct a hypothesis test to determine whether the value of $\beta_1$ is zero.
- Two tests are commonly used:

*t* Test and *F* test

- Both the *t* test and *F* test require an estimate of $\sigma^2$, the variance of $\varepsilon$ in the regression model.

# Testing for Significance

- An Estimate of $\sigma^2$

  The mean square error (MSE) provides the estimate of $\sigma^2$, and the notation $s^2$ is also used.

$$s^2 = \text{MSE} = \frac{\text{SSE}}{(n-2)}$$

where:

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

- SSE has n-2 degrees of freedom because two parameters (coefficient and slope) must be estimated to compute SSE

# Testing for Significance (3 of 3)

- An Estimate of $\sigma$
  - To estimate $\sigma$, we take the square root of $s^2$.
  - The resulting $s$ is called the standard error of the estimate.

$$s = \sqrt{MSE} = \sqrt{\frac{SSE}{n-2}}$$

# Testing for Significance: *t* Test (1 of 4)

- Hypotheses

$$H_0 : \beta_1 = 0$$
$$H_a : \beta_1 \neq 0$$

- Test Statistic

$$t = \frac{b_1}{s_{b_1}} \text{ where } s_{b_1} = \frac{s}{\sqrt{\sum (x_i - \bar{x})^2}}$$

# Testing for Significance: *t* Test (2 of 4)

- Rejection Rule

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

where: $t_{\alpha/2}$ is based on a $t$ distribution with $n-2$ degrees of freedom

**p-VALUE**

A *p*-value is a probability that provides a measure of the evidence against the null hypothesis provided by the sample. Smaller *p*-values indicate more evidence against $H_0$.

The p value is the probability that random chance generated data, or something else that is equal or rarer.

Check out: https://www.graphpad.com/quickcalcs/pvalue1.cfm

# Testing for Significance: *t* Test (3 of 4)

1. Determine the hypotheses.

$$H_0 : \beta_1 = 0$$
$$H_a : \beta_1 \neq 0$$

2. Specify the level of significance.

$$\alpha = .01$$

3. Select the test statistic.

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

4. State the rejection rule.

Reject $H_0$ if $p$-value $\leq .01$ or $|t| > 3.355$
(with 2 degrees of freedom)

# Testing for Significance: *t* Test (4 of 4)

5. Compute the value of the test statistic.

$$t = \frac{b_1}{s_{b_1}} = \frac{5}{.5803} = 8.62$$

6. Determine whether to reject $H_0$.

   $t = 3.355$ provides an area of .005 in the upper tail. Hence the *p*-value is less than .005. Also, $t = 8.62 > 3.355$. We can reject $H_0$.

# Confidence Interval for $\beta_1$ (1 of 3)

We can use a 99% confidence interval for $\beta_1$ to test the hypotheses just used in the *t* test.

$H_0$ is rejected if the hypothesized value of $\beta_1$ is not included in the confidence interval for $\beta_1$.

# Confidence Interval for $\beta_1$ (2 of 3)

- The form of a confidence interval for $\beta_1$ is :

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

where:

$b_1$ is the point estimator,

$t_{\alpha/2} s_{b_1}$ is the margin of error, 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

# Confidence Interval for $\beta_1$ (3 of 3)

- Rejection Rule

  Reject $H_0$ if 0 is not included in the confidence interval for $\beta_1$.

- 99% Confidence Interval for $\beta_1$.

$$b_1 \pm t_{\alpha/2} s_{b_1} = 5 \pm 3.355(.5803) = 5 \pm 1.95$$

or 3.05 to 6.95

- Conclusion

  0 is not included in the confidence interval. Reject $H_0$

# Testing for Significance: *F* Test

An $F$ test, based on the $F$ probability distribution, can also be used to test for significance in regression. With only one independent variable, the $F$ test will provide the same conclusion as the $t$ test; that is, if the $t$ test indicates $\beta_1 \neq 0$ and hence a significant relationship, the $F$ test will also indicate a significant relationship. But with more than one independent variable, only the $F$ test can be used to test for an overall significant relationship.

The logic behind the use of the $F$ test for determining whether the regression relationship is statistically significant is based on the development of two independent estimates of $\sigma^2$. We explained how MSE provides an estimate of $\sigma^2$. If the null hypothesis $H_0$: $\beta_1 = 0$ is true, the sum of squares due to regression, SSR, divided by its degrees of freedom provides another independent estimate of $\sigma^2$. This estimate is called the *mean square due to regression,* or simply the *mean square regression,* and is denoted MSR. In general,

$$\text{MSR} = \frac{\text{SSR}}{\text{Regression degrees of freedom}}$$

# Testing for Significance: *F* Test (1 of 4)

- Hypotheses

$$H_0 : b_1 = 0$$
$$H_a : b_1 \neq 0$$

- Test Statistic

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

$$F = \frac{\text{Variation in y explained by x}}{\text{Variation in y not explained by x}}$$

# Testing for Significance: *F* Test (2 of 4)

- Rejection Rule

$$\text{Reject } H_0 \text{ if } p\text{-value} \leq \alpha \text{ or } F \geq F_\alpha$$

where:

$F_\alpha$ is based on an $F$ distribution with 1 degree of freedom in the numerator and $n-2$ degrees of freedom in the denominator.

1 degree of freedom in the numerator because MSR ist the variance explained by adding one more parameter (slope)
N-2 df in the denominator because we need exactly two observations to fit 2 parameters (slope/intercept)

# Testing for Significance: *F* Test (3 of 4)

1. Determine the hypotheses.

$$H_0 : \beta_1 = 0$$
$$H_a : \beta_1 \neq 0$$

2. Specify the level of significance.

$$\alpha = .01$$

3. Select the test statistic.

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

4. State the rejection rule.

Reject $H_0$ if $p$-value $\leq .01$ or $F \geq 11.26$ (with 1 d.f. in numerator and 8 d.f. in denominator).

# Testing for Significance: *F* Test (4 of 4)

5. Compute the value of the test statistic.

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

6. Determine whether to reject $H_0$.

$F = 11.26$ provides an area of .01 in the upper tail. Thus, the *p*-value corresponding to $F = 74.25$ is less than .01. Hence, we reject $H_0$.

Remember: p-value is the number of more (or equally extreme values) by more values

The statistical evidence is sufficient to conclude that a significant relationship exists between the size of the student population and quarterly sales

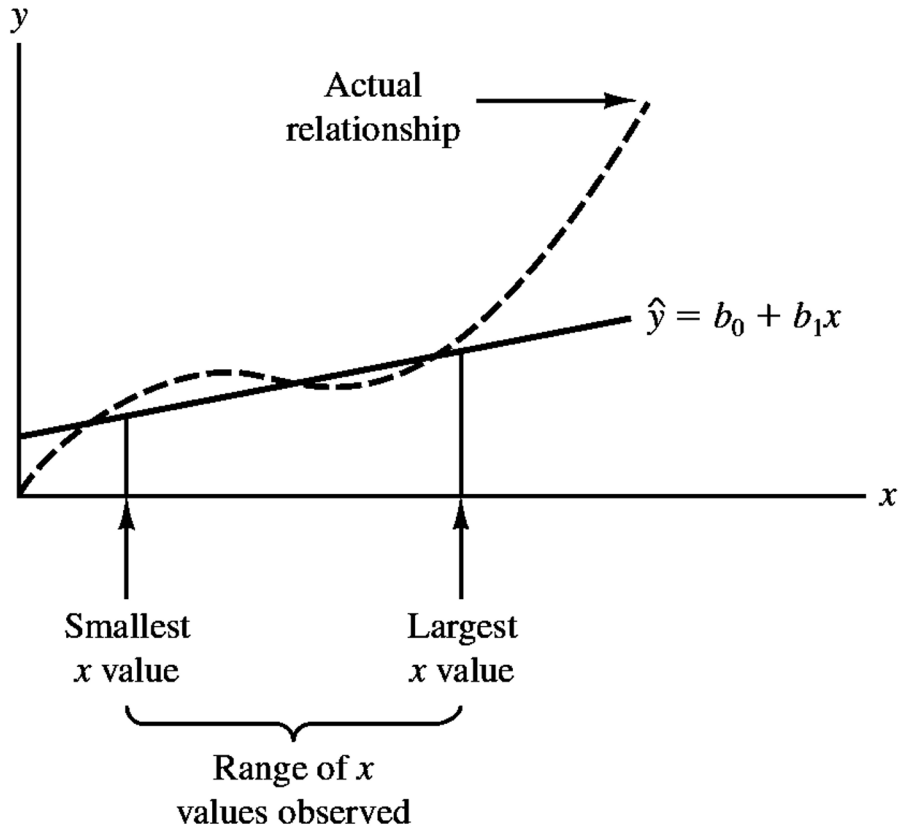
# Testing for Significance: *F* Test (4 of 4)

| Source of Variation | Sum of Squares | Degrees of Freedom | Mean Square | *F* | *p*-Value |
|---|---|---|---|---|---|
| Regression | SSR | 1 | $\text{MSR} = \frac{\text{SSR}}{1}$ | $F = \frac{\text{MSR}}{\text{MSE}}$ | |
| Error | SSE | $n - 2$ | $\text{MSE} = \frac{\text{SSE}}{n-2}$ | | |
| Total | SST | $n - 1$ | | | |

| Source of Variation | Sum of Squares | Degrees of Freedom | Mean Square | *F* | *p*-Value |
|---|---|---|---|---|---|
| Regression | 14,200 | 1 | $\frac{14{,}200}{1} = 14{,}200$ | $\frac{14{,}200}{191.25} = 74.25$ | .000 |
| Error | 1530 | 8 | $\frac{1530}{8} = 191.25$ | | |
| Total | 15,730 | 9 | | | |

# Some Cautions about the Interpretation of Significance Tests

- 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*.



# Simple Linear Regression

- Using the Estimated Regression Equation for Estimation and Prediction
- Excel’s Regression Tool
- Residual Analysis: Validating Model Assumptions
- Outliers and Influential Observations

# Excel's Regression Tool

- Up to this point, you have seen how Excel can be used for various parts of a regression analysis.
- Excel also has a comprehensive tool in its Data Analysis package called Regression.
- The Regression tool can be used to perform a complete regression analysis.

# Using Excel's Regression Tool (1 of 4)

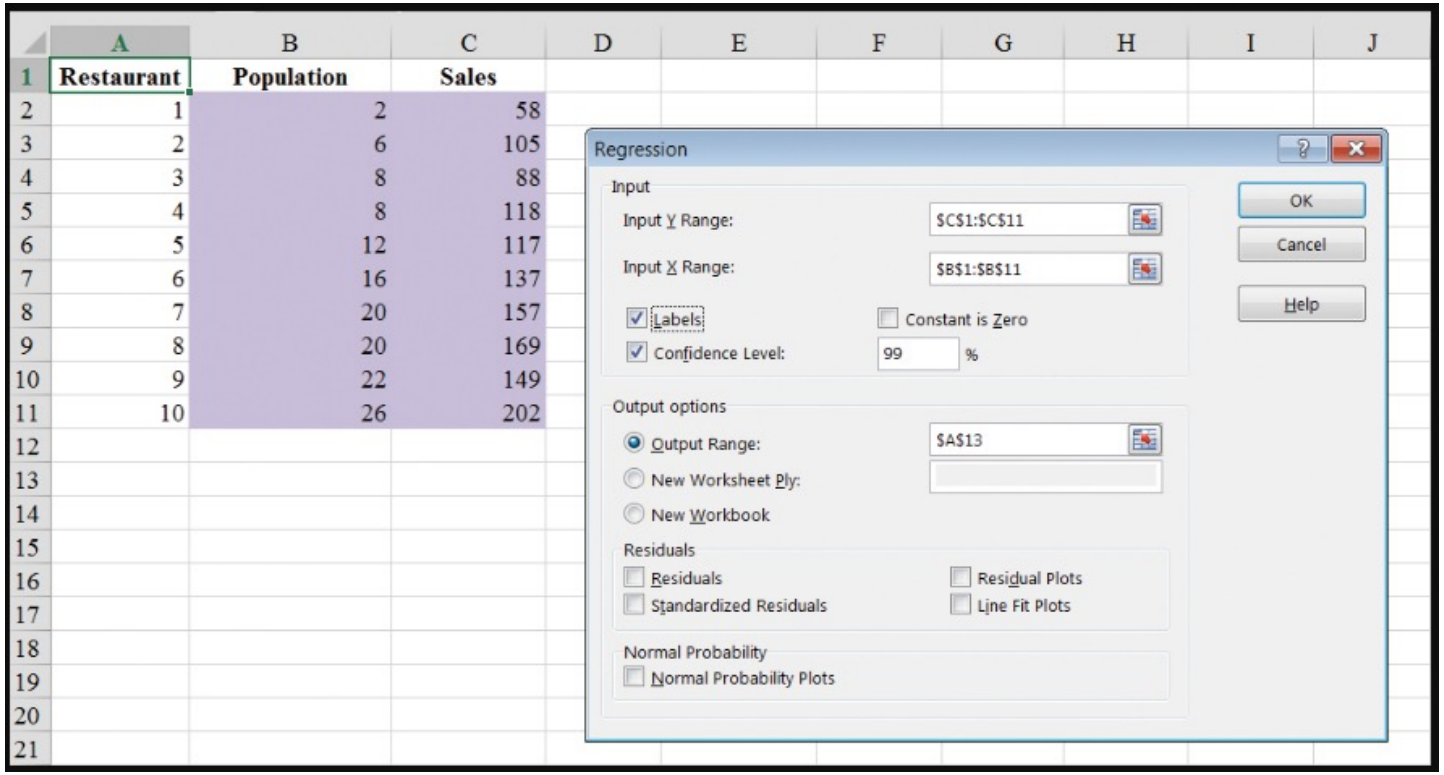
- Performing the Regression Analysis
  - Step 1  Click the **DATA** tab on the Ribbon
  - Step 2  In the **Analyze** group, click **Data Analysis**
  - Step 3  Choose **Regression** from the list of Analysis Tools

# Using Excel's Regression Tool (2 of 4)

- Performing the Regression Analysis
  - Step 4  When the Regression dialog box appears:
    - Enter *C1:C11* in the **Input Y Range** box
    - Enter *B1:B11* in the **Input X Range** box
    - Select the check box for **Labels**
    - Select the check box for **Confidence Level**
    - Enter *99* in the **Confidence Level** box
    - Select **Output Range**
    - Enter *A13* in the **Output Range** box
    - Click **OK**

# Using Excel's Regression Tool (3 of 4)

**Example:** Armand's Pizza Parlors: Regression tool dialog box



| | A | B | C |
|---|---|---|---|
| 1 | Restaurant | Population | Sales |
| 2 | 1 | 2 | 58 |
| 3 | 2 | 6 | 105 |
| 4 | 3 | 8 | 88 |
| 5 | 4 | 8 | 118 |
| 6 | 5 | 12 | 117 |
| 7 | 6 | 16 | 137 |
| 8 | 7 | 20 | 157 |
| 9 | 8 | 20 | 169 |
| 10 | 9 | 22 | 149 |
| 11 | 10 | 26 | 202 |

# Using Excel's Regression Tool (4 of 4)

**Example:** Armand's Pizza Parlors: Regression tool output

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 13 | SUMMARY OUTPUT | | | | | | | | |
| 14 | | | | | | | | | |
| 15 | *Regression Statistics* | | | | | | | | |
| 16 | Multiple R | 0.9501 | | | | | | | |
| 17 | R Square | 0.9027 | | | | | | | |
| 18 | Adjusted R Square | 0.8906 | | | | | | | |
| 19 | Standard Error | 13.8293 | | | | | | | |
| 20 | Observations | 10 | | | | | | | |
| 21 | | | | | | | | | |
| 22 | ANOVA | | | | | | | | |
| 23 | | *df* | *SS* | *MS* | *F* | *Significance F* | | | |
| 24 | Regression | 1 | 14200 | 14200 | 74.2484 | 2.55E-05 | | | |
| 25 | Residual | 8 | 1530 | 191.25 | | | | | |
| 26 | Total | 9 | 15730 | | | | | | |
| 27 | | | | | | | | | |
| 28 | | *Coefficients* | *Standard Error* | *t Stat* | *P-value* | *Lower 95%* | *Upper 95%* | *Lower 99.0%* | *Upper 99.0%* |
| 29 | Intercept | 60 | 9.2260 | 6.5033 | 0.0002 | 38.7247 | 81.2753 | 29.0431 | 90.9569 |
| 30 | Population | 5 | 0.5803 | 8.6167 | 2.55E-05 | 3.6619 | 6.3381 | 3.0530 | 6.9470 |
| 31 | | | | | | | | | |

Standard Error: the average distance that the data points fall from the regression line (precision of the analysis)

If the lower/upper X%c.i. do not contain zero, your p is lower than 100-X%

With a p value of 0.00005 there is a 100%- 0.005% chance of being correct that x has an effect on y

t values test whether the coefficients are zero

t=coef/stderror

# Residual Analysis: Validating Model Assumptions

- If the assumptions about the error term $\varepsilon$ appear questionable, the hypothesis tests about the significance of the regression relationship and the interval estimation results may not be valid.
- The residuals provide the best information about $\varepsilon$.
- Residual for observation $i$

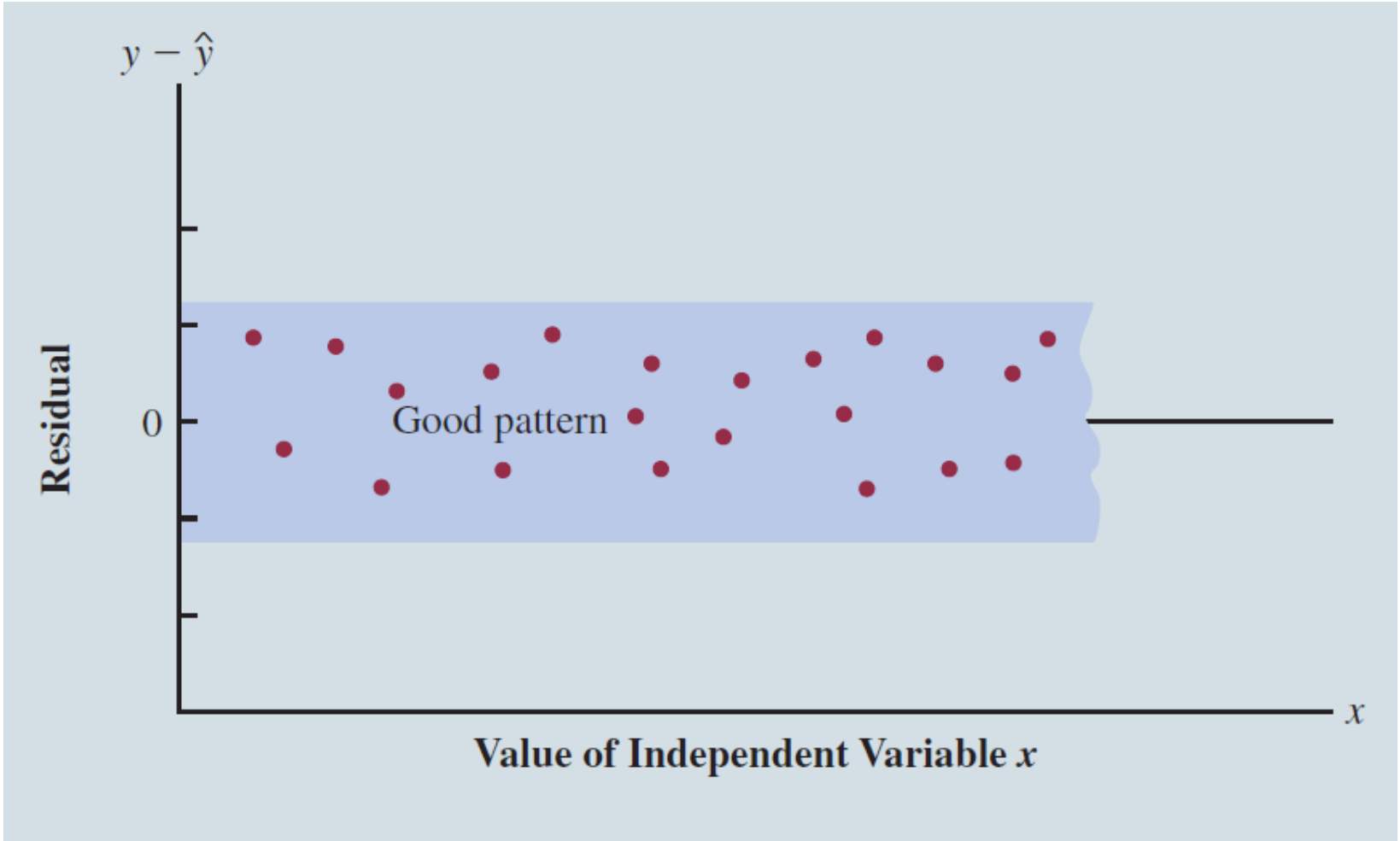
$$y_i - \hat{y}_i$$

- Much of the residual analysis is based on an examination of graphical plots.

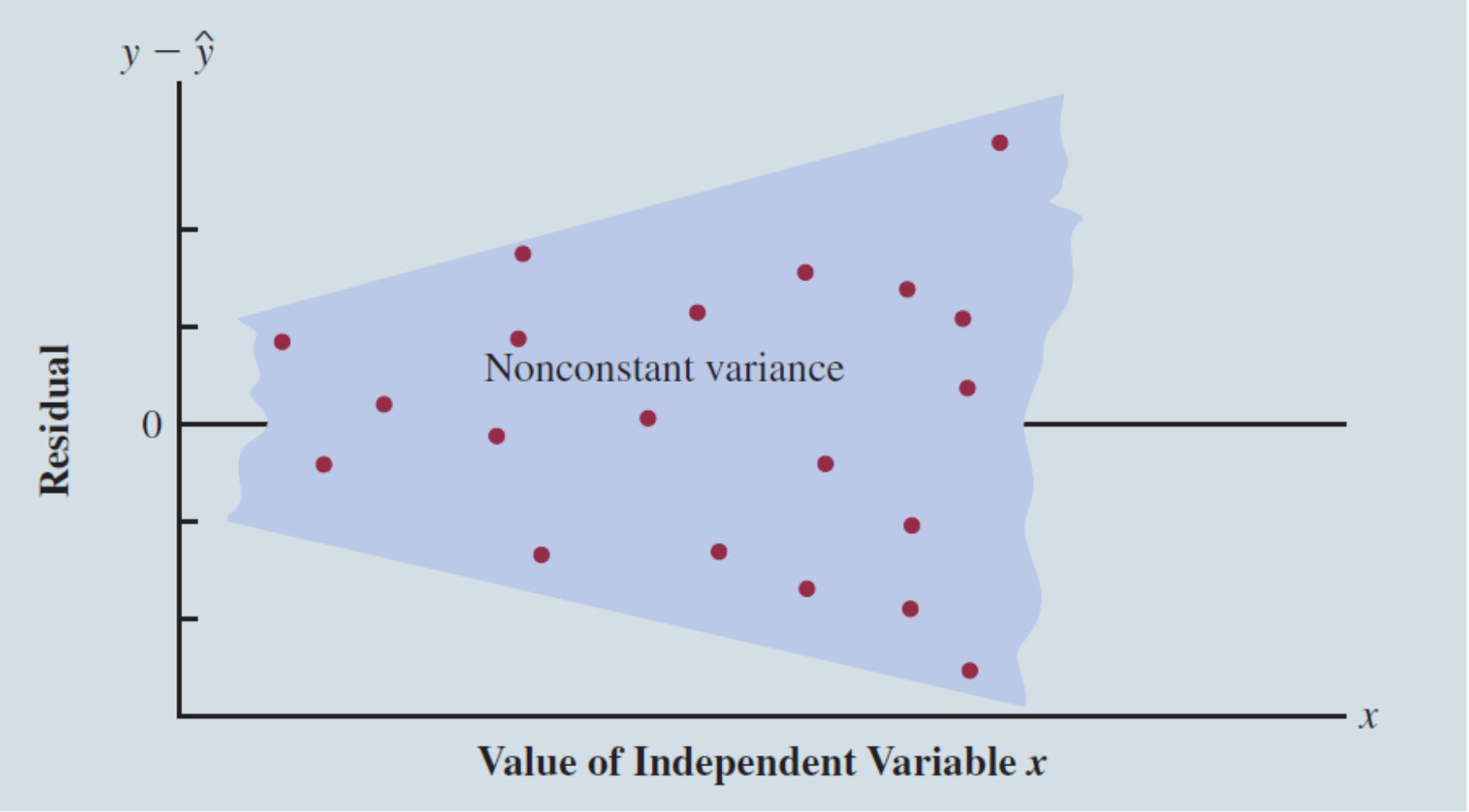
# Residual Plot Against *x* (1 of 7)

- If the assumption that the variance of $\varepsilon$ is the same for all values of $x$ is valid, and the assumed regression model is an adequate representation of the relationship between the variables, then the residual plot should give an overall impression of a horizontal band of points.

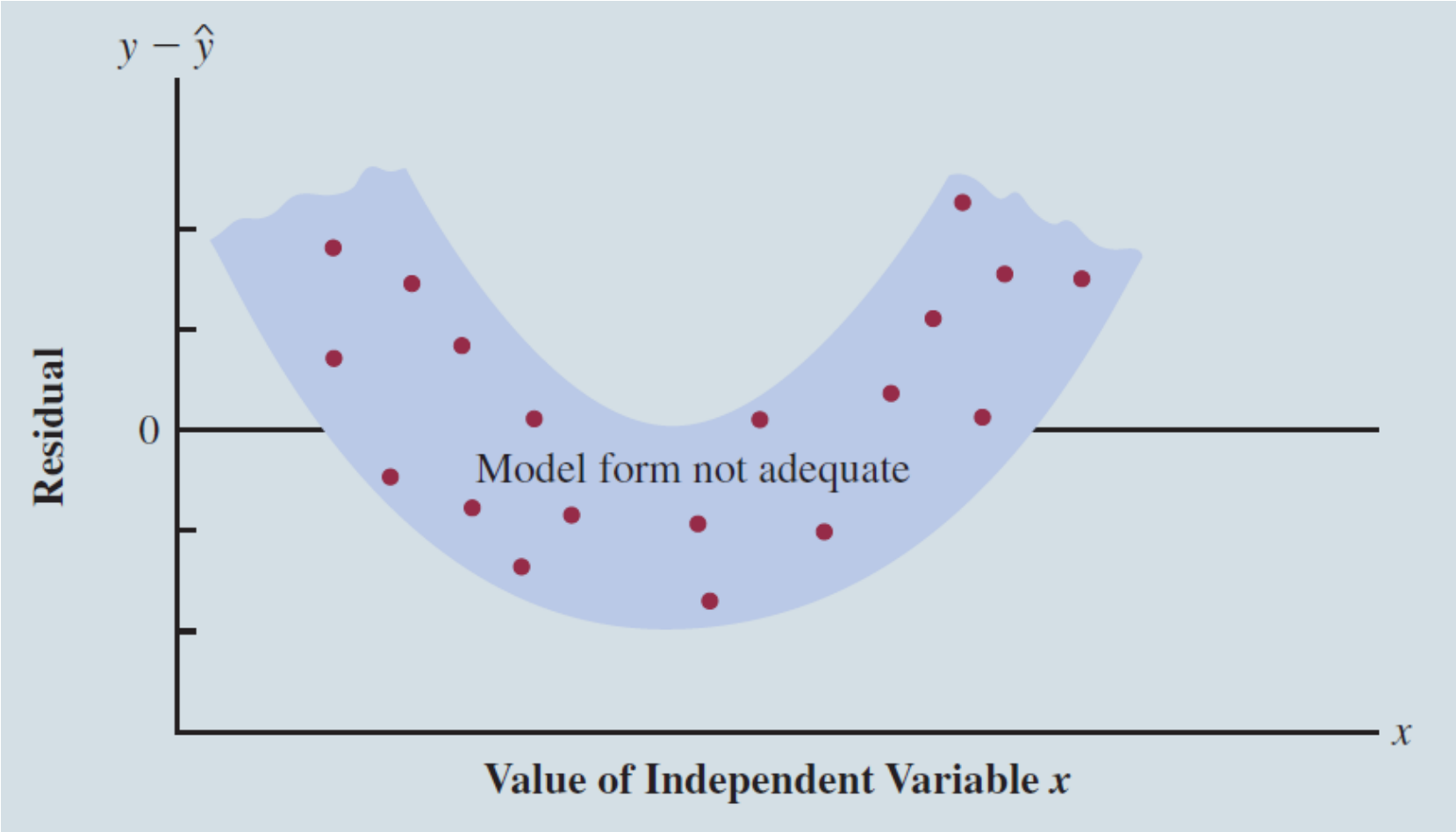
# Residual Plot Against *x* (2 of 7)



# Residual Plot Against *x* (3 of 7)



# Residual Plot Against *x* (4 of 7)



# Residual Plot Against *x* (5 of 7)

**Example:**

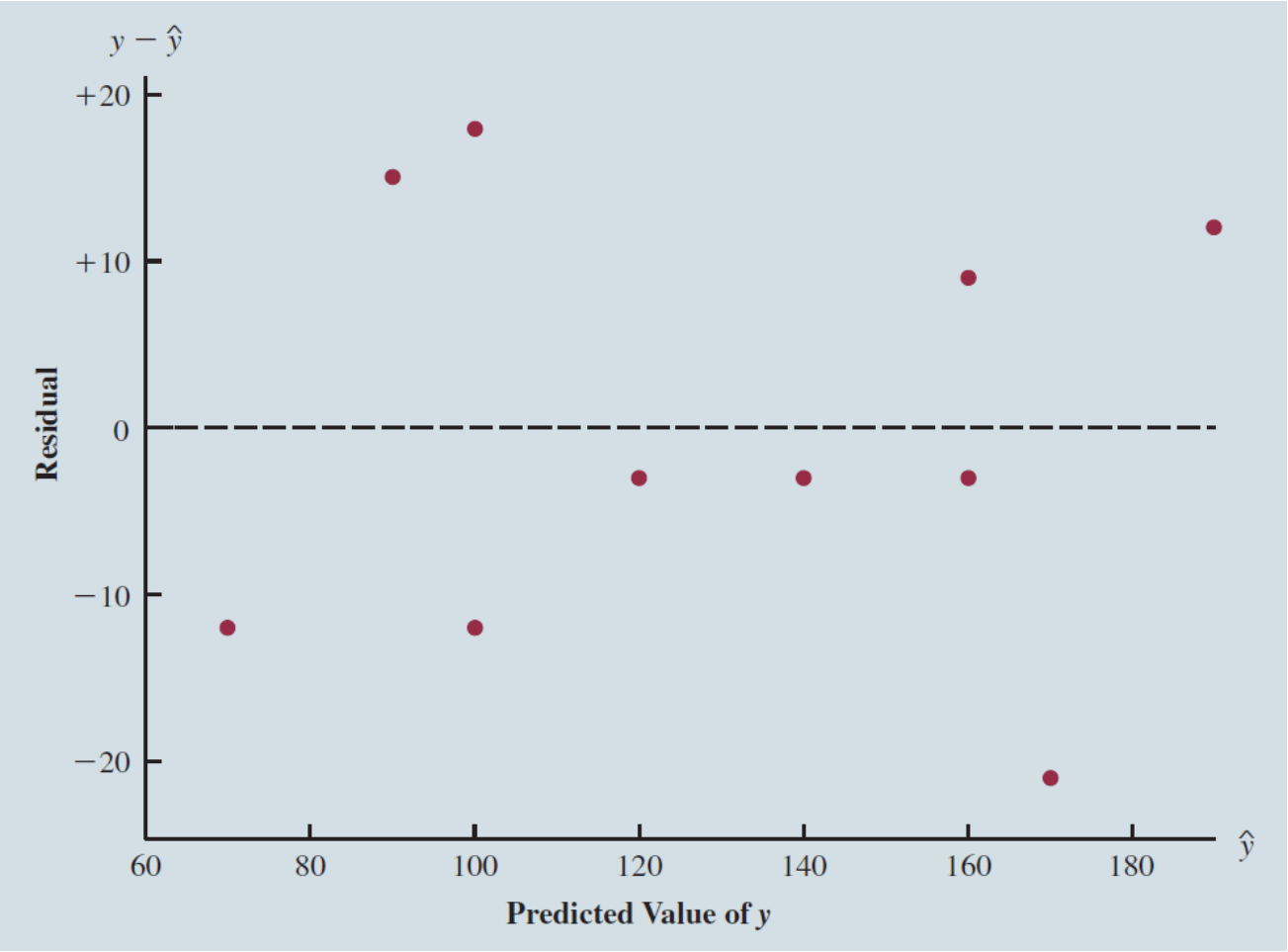
Armand's Pizza Parlors

| Student Population $(x_i)$ | Sales $(y_i)$ | Predicted sales $y_i = 60 + 5(x_i)$ | Residuals $(y_i - \hat{y}_i)$ |
|---|---|---|---|
| 2 | 58 | 70 | −12 |
| 6 | 105 | 90 | 15 |
| 8 | 88 | 100 | −12 |
| 8 | 118 | 100 | 18 |
| 12 | 117 | 120 | −3 |
| 16 | 137 | 140 | −3 |
| 20 | 157 | 160 | −3 |
| 20 | 169 | 160 | 9 |
| 22 | 149 | 170 | −21 |
| 26 | 202 | 190 | 12 |

# Residual Plot Against *x* (6 of 7)

**Example:** Armand's Pizza Parlors

- Plot of the residuals against independent variable *x*.



# Residual Plot Against *x* (7 of 7)

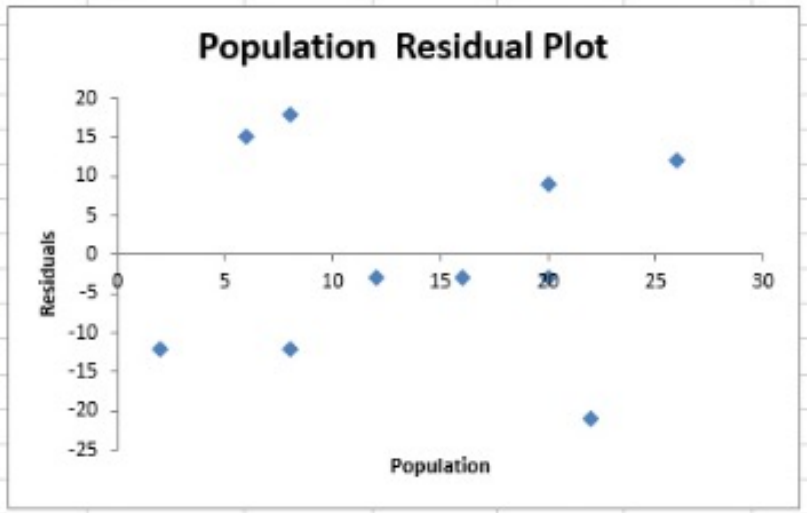
- Using Excel to Produce a Residual Plot
  - The steps outlined earlier to obtain the regression output are performed with one change.
  - When the Regression dialog box appears, we must also select the **Residual Plot** option.
  - The output will include two new items:
    - A plot of the residuals against the independent variable, and
    - A list of predicted values of *y* and the corresponding residual values.

# Using Excel to Produce a Residual Plot

**Example:**

Armand's Pizza Parlors

| | A | B | C |
|---|---|---|---|
| 1 | Restaurant | Population | Sales |
| 2 | 1 | 2 | 58 |
| 3 | 2 | 6 | 105 |
| 4 | 3 | 8 | 88 |
| 5 | 4 | 8 | 118 |
| 6 | 5 | 12 | 117 |
| 7 | 6 | 16 | 137 |
| 8 | 7 | 20 | 157 |
| 9 | 8 | 20 | 169 |
| 10 | 9 | 22 | 149 |
| 11 | 10 | 26 | 202 |
| 33 | | | |
| 34 | RESIDUAL OUTPUT | | |
| 35 | | | |
| 36 | *Observation* | *Predicted Sales* | *Residuals* |
| 37 | 1 | 70 | -12 |
| 38 | 2 | 90 | 15 |
| 39 | 3 | 100 | -12 |
| 40 | 4 | 100 | 18 |
| 41 | 5 | 120 | -3 |
| 42 | 6 | 140 | -3 |
| 43 | 7 | 160 | -3 |
| 44 | 8 | 160 | 9 |
| 45 | 9 | 170 | -21 |
| 46 | 10 | 190 | 12 |
| 47 | | | |



# Standardized Residuals

- Standardized Residual for Observation *I*

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

where

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

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

# Standardized Residual Plot (1 of 4)

- The standardized residual plot can provide insight about the assumption that the error term $\varepsilon$ has a normal distribution.
- If this assumption is satisfied, the distribution of the standardized residuals should appear to come from a standard normal probability distribution.

# Standardized Residual Plot (2 of 4)

**Example:**

Armand's Pizza Parlors

| Observation | Predicted sales $y_i = 60 + 5(x_i)$ | Residuals $(y_i - \hat{y}_i)$ | Standardized Residual |
|---|---|---|---|
| 1 | 70 | −12 | −1.0792 |
| 2 | 90 | 15 | 1.2224 |
| 3 | 100 | −12 | −.9487 |
| 4 | 100 | 18 | 1.4230 |
| 5 | 120 | −3 | −.2296 |
| 6 | 140 | −3 | −.2296 |
| 7 | 160 | −3 | −.2372 |
| 8 | 160 | 9 | .7115 |
| 9 | 170 | −21 | −1.7114 |
| 10 | 190 | 12 | 1.0792 |

# Standardized Residual Plot (3 of 4)

**Example:**

Armand's Pizza Parlors

# Standardized Residual Plot (4 of 4)

- All of the standardized residuals are between −2 and +2 indicating that there is no reason to question the assumption that $\varepsilon$ has a normal distribution.

# Using Excel to construct a Standardized Residual Plot

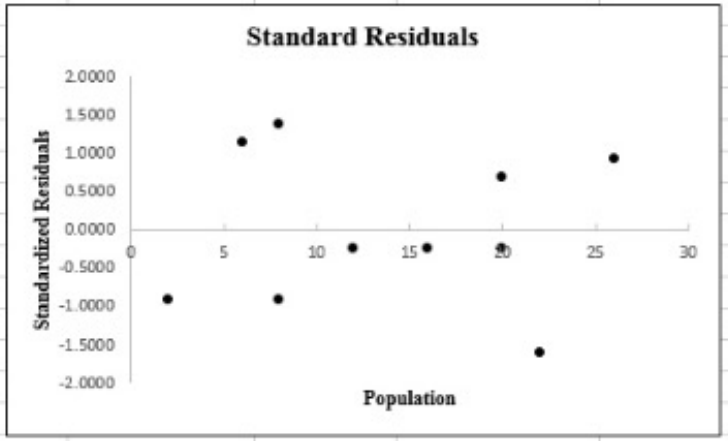
**Example:**

Armand's Pizza Parlors

RESIDUAL OUTPUT

| Observation | Predicted Sales | Residuals | Standard Residuals |
|---|---|---|---|
| 1 | 70 | -12 | -0.9204 |
| 2 | 90 | 15 | 1.1504 |
| 3 | 100 | -12 | -0.9204 |
| 4 | 100 | 18 | 1.3805 |
| 5 | 120 | -3 | -0.2301 |
| 6 | 140 | -3 | -0.2301 |
| 7 | 160 | -3 | -0.2301 |
| 8 | 160 | 9 | 0.6903 |
| 9 | 170 | -21 | -1.6106 |
| 10 | 190 | 12 | 0.9204 |

| Population | Standard Residuals |
|---|---|
| 2 | -0.9204 |
| 6 | 1.1504 |
| 8 | -0.9204 |
| 8 | 1.3805 |
| 12 | -0.2301 |
| 16 | -0.2301 |
| 20 | -0.2301 |
| 20 | 0.6903 |
| 22 | -1.6106 |
| 26 | 0.9204 |



# Outliers and Influential Observations

- Detecting Outliers
  - An outlier is an observation that is unusual in comparison with the other data.
  - Minitab classifies an observation as an outlier if its standardized residual value is $< -2$ or $> +2$.
  - This standardized residual rule sometimes fails to identify an unusually large observation as being an outlier.


# RWTH BUSINESS SCHOOL

Mathematics & Statistics
M.Sc. Data Analytics and Decision Science

Prof. Dr. Thomas S. Lontzek

CENGAGE

© 2021 Cengage Learning. All Rights Reserved. May not be scanned, copied or duplicated, or posted to a publicly accessible website, in whole or in part.