Another area of inferential statistics involves determining whether a \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ exists between two, or more, numerical, that is, quantitative, variables.

For instance, is there a \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ between a person’s blood pressure and his or her age or is there a \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ between the size of a house and its appraised value.

We will investigate whether there is a \_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_between the variables; if so, what is the \_\_\_\_\_\_\_\_\_\_\_\_\_\_ of the \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_; what type of \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ does exist; and what \_\_\_\_\_\_\_\_\_\_\_\_\_\_ might be made from the relationship?

# 10 - 1. Scatter Plots and Correlation

Objective 1. Draw a Scatter Plot for a Set of Ordered Pairs.

In simple correlation and regression studies, data from two \_\_\_\_\_\_\_\_\_\_\_\_\_ or \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ variables are collected to study whether a relationship exists between the two variables. The data are collected as \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ pairs.

One variable is the \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ variable and the other is the \_\_\_\_\_\_\_\_\_\_\_\_\_\_ variable.

The independent variable can be \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_, is known as the \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ variable and is called *x.*

The dependent variable, which \_\_\_\_\_\_\_\_\_\_\_\_\_ be controlled or manipulated, is known as the \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ variable and is called *y*.

The researcher \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ which variable will be considered the independent variable and which will be the dependent variable. Sometimes, one variable can be \_\_\_\_\_\_\_\_\_\_\_\_\_ to affect the result of the other. For instance, the hours a student studies for an exam is likely to have an impact on the percentage of correct answers on the exam. In this case, the hours studied would be the explanatory variable and the percentage of correct answers would be the response variable. However, when considering the relationship between one’s income and one’s height, it is not clear which should be independent and which should be dependent, so the determination is \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_.

The independent and dependent variables can be plotted on a graph called a \_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_. The horizontal axis is labeled with the name of the \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ variable and vertical axis is labeled with the name of the \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ variable. The coordinates of the axes are determined by the \_\_\_\_\_\_\_\_\_\_\_ to the \_\_\_\_\_\_\_\_\_\_\_\_\_data values of the variables.

### Definition: Scatter Plot

A scatter plot is a graph of the \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ pairs (*x, y*) of numbers consisting of the independent variable *x* and the dependent variable *y*.

The scatter plot is a \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ way to describe the nature of the \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ between the independent and dependent variables, even if the scales are not the same.

Researchers look for \_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_ relationships when the data points increase from left to right; \_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_ relationships when the data points decrease from left to right; curvilinear relationships or other \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ relationships, or \_\_\_\_\_\_\_\_\_\_\_ relationship at all when the scatter shows no pattern, line or curve.

Four scatter plots illustrating the scatter of pairs of data showing a positive linear relationship, a negative linear relationship, a curvilinear relationship and no relationship. 

### Procedure for Drawing a Scatter Plot

**Step 1** Draw and label the *x* and *y* axes.

**Step 2** Plot each point on the graph.

**Step 3** Determine the type of relationship (if any) that exists for the variables.

### Example 10-1. Absences and Final Grades

Construct a scatter plot for the data obtained in a study on the number of absences and the final grades of seven randomly selected students from a statistics class.

| **Student** | **Number of absences, *x*** | **Final grade, *y* (%)** |
| --- | --- | --- |
| A | 6 | 82 |
| B | 2 | 86 |
| C | 15 | 43 |
| D | 9 | 74 |
| E | 12 | 58 |
| F | 5 | 90 |
| G | 8 | 78 |

*Solution*

**Step 1** Draw and label the *x* and *y* axes. Label in relation to size of variables.

**Step 2** Plot each point on the graph.

Scatter plot of the points showing number of absences x and final grade y. The ordered pairs of (x, y) are (6 82), (2, 86), (15, 43), (9, 74), 912, 58), (5, 90), and (8,78).

**Step 3** Determine the type of relationship, if any, that exists.

In this example, there appears to be a negative linear relationship between the number of absences and the final grade, as a percentage, of the students.

### Example 10-2. Number of Teachers (in 1000s) and Number of Pupils per Teacher

A researcher wants to check for a relationship between the number of pupils per teacher and the number of teachers, in thousands, employed by the school district. The data are from a randomly selected sample.

| **School**  **District** | **Number of Teachers (in thousands)** | **Pupils per**  **Teacher** |
| --- | --- | --- |
| 1 | 7 | 12.4 |
| 2 | 34 | 14.3 |
| 3 | 9 | 14.3 |
| 4 | 8 | 9.2 |
| 5 | 16 | 18.3 |
| 6 | 15 | 12.1 |
| 7 | 6 | 12.3 |
| 8 | 14 | 12.4 |
| 9 | 32 | 15.2 |
| 10 | 10 | 13.4 |

*Solution:*

**Step 1** Draw and label the *x* and *y* axes. Label in relation to size of variables.

**Step 2** Plot each point on the graph.

**Step 3** Determine the type of relationship, if any, that exists.

In this example, there appears to be \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_.

## Objective 2. Compute the Correlation Coefficient.

The correlation coefficient measures the \_\_\_\_\_\_\_\_\_\_\_\_ of a relationship between two variables.

### Definition: Population Correlation Coefficient

The population correlation coefficient, denoted by *ρ*, is the correlation computed using \_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_ of data values, (*x, y*), taken from a population.

### Definition: Linear Correlation Coefficient

The linear correlation coefficient, denoted by *r*, computed from the sample measures the \_\_\_\_\_\_\_\_\_\_\_\_\_\_ and \_\_\_\_\_\_\_\_\_\_\_\_\_ of a \_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ between two quantitative variables. (We use the Pearson product moment correlation coefficient, PPMC.)

### Properties of the Linear Correlation Coefficient.

1. The correlation coefficient is a \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ measure.
2. The value of *r* will always be \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_, inclusive.  
   That is, .
3. If the values of *x* and *y* are \_\_\_\_\_\_\_\_\_\_\_\_\_\_, the value of *r* is unchanged.
4. If the values of *x* and/or *y* are converted to a \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_, the value of *r* is unchanged.
5. The value of *r* is sensitive to \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ and can change dramatically if they are present in the data.

The graphs below show the relationship between the value of the correlation coefficient and their respective scatter plots:

Six scatter plots are shown.  The first three, a through c, show positive linear relationships.  (a) shows a group of data points with a relatively positive increase from left to right but are not closely following a line with a correlation coefficient of 0.50; (b) shows a group of data points clustered together indicating a line with positive slope and a correlation of 0.90; (c) shows a set of data points following a line with a positive slope and a correlation coefficient of 1.00.  The last three, d thorugh f, show negative linear relationships.  (a) shows a group of data points with a relatively negative decrease from left to right but are not closely following a line with a correlation coefficient of -0.50; (b) shows a group of data points clustered together indicating a line with negative slope and a correlation of -0.90; (c) shows a set of data points following a line with a negative slope and a correlation coefficient of -1.00.

### Assumptions for the Correlation Coefficient

1. The sample is a \_\_\_\_\_\_\_\_\_\_\_\_\_ sample.
2. The data pairs fall approximately on a \_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_ and are measured at the interval or ratio level.
3. The variables have a \_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_ distribution. That is, given any specific value of *x*, the *y* values are \_\_\_\_\_\_\_\_\_\_\_\_ distributed; and given any specific value of *y*, the *x* values are \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ distributed.

Be sure to check that the assumptions are met when the exercise does not state that they hold.

### A Formula for the Linear Correlation Coefficient, *r*.

where *n* is the number of data pairs. Round *r* to three decimal places.

### Procedure for Finding the Value of the Linear Correlation Coefficient.

**Step 1** Make a table:

| ***x*** | ***y*** | ***xy*** |  |  |
| --- | --- | --- | --- | --- |

**Step 2** Fill out the table.

Place the values of *x* in the *x* column and the values of *y* in the *y* column.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ each *x* value by the corresponding *y* value, and place the products in the *xy* column.

\_\_\_\_\_\_\_\_\_\_\_\_ each *x* value and place the squares in the column.

\_\_\_\_\_\_\_\_\_\_\_\_ each *y* value and place the squares in the column.

\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_ of each column.

**Step 3** Substitute the values into the formula and find the value of *r*:

where *n* is the number of data pairs. Round *r* to three decimal places.

### Example 10-3. Absences and Final Grades

For the data in example 10-1, compute the linear correlation coefficient.

We already have this data and have noticed that the scatterplot has a negative slope.

*Solution:*

**Steps 1 & 2**

| ***x*** | ***y*** | ***xy*** |  |  |
| --- | --- | --- | --- | --- |
| 6 | 82 | 492 | 36 | 6724 |
| 2 | 86 | 172 | 4 | 7396 |
| 15 | 43 | 645 | 225 | 1849 |
| 9 | 74 | 666 | 81 | 5476 |
| 12 | 58 | 696 | 144 | 3364 |
| 5 | 90 | 450 | 25 | 8100 |
| 8 | 78 | 624 | 64 | 6084 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  |  |  |  |
| *57* | *511* | *3745* | *579* | *38995* |

**Step 3**

The linear correlation coefficient suggests a strong negative linear relationship between the number of absences students have in class and their final grade. That is, the more absences a student has, the lower the final grade will be.

### Example 10-4. Number of Teachers (in 1000s) and Number of Pupils per Teacher

For the data in example 10-2, compute the linear correlation coefficient.

*Solution:*

**Steps 1 & 2**

| ***x*** | ***y*** | ***xy*** |  |  |
| --- | --- | --- | --- | --- |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  |  |  |  |

**Step 3**

## Objective 3. Test the Hypothesis

The correlation coefficient is between and . When *r* is close to either or , the linear relationship is \_\_\_\_\_\_\_\_\_\_\_. When *r* is near , the linear relationship is \_\_\_\_\_\_\_\_\_ or \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ . Because the value of *r* is computed from sample data, there are two possibilities when *r* is not equal to zero: either the value of *r* is high enough to conclude that there is a \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ linear relationship between the two variables, or the value of *r* is due to \_\_\_\_\_\_\_\_\_\_\_\_\_\_.

### Assumptions for Testing the Significance of the Linear Correlation Coefficient

1. The data are \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ and are obtained from a simple \_\_\_\_\_\_\_\_\_\_\_ sample.
2. The scatter plot shows that the data are approximately \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ related.
3. There are no \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ in the data.
4. The variables *x* and *y* must come from \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ distributed populations.

The null and alternative hypotheses are:

There is \_\_\_\_\_\_\_\_\_\_\_\_\_\_ linear correlation between *x* and *y*.

There is \_\_\_\_\_\_\_\_\_\_\_\_\_\_ linear correlation between *x* and *y*.

We will discuss \_\_\_\_\_\_\_\_\_\_\_\_ methods to test the significance of the correlation coefficient. The first is the \_\_\_\_\_\_\_\_\_\_\_\_\_\_ method and uses the *t* test; the second uses the *\_\_\_\_\_\_\_\_\_\_*; and the third uses \_\_\_\_\_\_\_\_\_\_\_\_ to find the \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ for a specific degrees of freedom and a given significance level.

### Formula for the *t* Test for the Correlation Coefficient

for degrees of freedom equal to ,

where *n* is the number of ordered pairs .

### The Traditional Method of Hypothesis Testing Significance of *r*

**Step 1** State the hypotheses.

**Step 2** Find the critical values of *t*.

**Step 3** Compute the test value.

**Step 4** Make the decision.

**Step 5** Summarize the results.

### The *P*-Value Method of Hypothesis Testing Significance of *r*

**Step 1** State the hypotheses.

**Step 2** Compute the test value of *t*.

**Step 3** Find the *P*-value.

**Step 4** Make the decision.

**Step 5** Summarize the results.

### The *r* Method of Hypothesis Testing for Significance of *r* (Two-Tailed Tests only)

**Step 1** State the hypotheses.

**Step 2** Find the critical values of *r* from Table I.

**Step 3** Compute the correlation coefficient, *r*.

**Step 4** Make the decision.

**Step 5** Summarize the results.

### Example 10-5. Test the Significance of the Correlation Coefficient in Example 10-3

Test the significance of the claim that there is a linear correlation between absences and final grade. Use .

In Example 10-3, we found .

*Solution:*

**Step 1** State the hypotheses.

There is no linear correlation between number of absences and final grade.

There is a linear correlation between number of absences and final grade.

**Method 1**

**Step 2** Find the critical values of *t*.

For degrees of freedom = , .

**Step 3** Compute the test value.

**Step 4** Make the decision.

The test value = = The critical value. Therefore, reject the null hypothesis.

**Step 5** Summarize the results.

There is sufficient evidence to support a significant linear relationship between the number of absences and final grade.

**Method 2**

**Step 2** Compute the test value.

**Step 3** Find the *P*-value.

Using Table F, the *P*-value .

Using technology, *P*-value

**Step 4** Make the decision.

or

Therefore, reject the null hypothesis.

**Step 5** Summarize the results.

There is sufficient evidence to support a significant linear relationship between the number of absences and final grade.

**Method 3**

**Step 2** Find the critical values of *r* from Table I.

In Table I, locate the row for the degrees of freedom.

The critical value of *r* is listed for or .

For d.f.= 5 and , the critical value of *r* is .

**Step 3** Compute the correlation coefficient, *r*.

**Step 4** Make the decision.

Reject the null hypothesis.

**Step 5** Summarize the results.

There is sufficient evidence to support a significant linear relationship between the number of absences and final grade.

### Example 10-6. Test the Significance of the Correlation Coefficient in Example 10-4

For the number of teachers and number of pupils per teacher data, test the significance of the correlation coefficient. Use

### Possible Relationships Between Variables

When the null hypothesis has been rejected for a specific α value, any of the following possibilities can exist:

1. *There is a \_\_\_\_\_\_\_\_\_\_\_\_\_ cause-and-effect relationship between the variables.* That is, *x* causes *y*. For example, water causes plants to grow, poison causes death, and heat causes ice to melt.
2. *There is a \_\_\_\_\_\_\_\_\_\_\_ cause-and-effect relationship between the variables.* That is, *y* causes *x*. For example, a researcher may believe that excessive coffee consumption causes nervousness, but it may be that a nervous person craves coffee to calm his or her nerves.
3. *The relationship between the variables may be \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_ variable.* For example, if a statistician correlated the number of deaths due to drowning and the number of cans of soft drink consumed daily during the summer, he or she would probably find a significant relationship. However, the soft drink is not necessarily responsible for the deaths, since both variables may be related to heat and humidity.  
   This third variable is a variable that the researcher does not know about or that is not accounted for in the study and is called a \_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_.
4. *There may be a \_\_\_\_\_\_\_\_\_\_\_\_\_\_ of \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ among many variables.* For example, a researcher may find a significant relationship between students’ high school grades and college grades. But there probably many other variables involved, such as IQ, hours of study, influence of parents, motivation, age, and instructors.
5. *The relationship may be \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_.* For example, a researcher may be able to find a significant relationship between the increase in the number of people who are exercising and increase in the number of people who are committing crimes. But common sense dictates that any relationship between these two values must be due to coincidence.

**Caution:** When averages are used, instead of individual data, the results may be a higher correlation that actually exists. Results of correlation studies using averages cannot be generalized to individuals because averaging tends to smooth out variability among individual data values.

Even if the **correlation** between the two variables is significant, it does not imply \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_.

# 10 – 2. Regression

## Objective 4. Compute the Equation of the Regression Line.

If the value of the correlation coefficient is \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_, then determine the equation of the \_\_\_\_\_\_\_\_\_\_\_\_\_ line, the line that fits the data best, i.e., the line of best fit. The purpose of the regression line is to make it possible to see the trend and make \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ on the basis of data.

If the value of *r* is not significant, the predictions are \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_.

### Line of Best Fit

The line of best fit is the line for which the sum of the squares of the vertical distance from each point to the line is a \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_.

The difference between the actual value of *y* and the predicted value of *y*, also called , is the \_\_\_\_\_\_\_\_\_\_\_\_ or the predicted error. The \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ are used to calculate the line of best fit.

The line of best fit is used to predict values of *y* for specific values of *x*. The closer the points are to the line, the better the fit and the better the predictions are. When *r* is positive, the slope of the line of best fit is \_\_\_\_\_\_\_\_\_\_\_\_\_. When *r* is negative, the slope of the line of best fit is \_\_\_\_\_\_\_\_\_\_\_\_\_\_.

### Line of Best Fit for a Set of Data Points, Showing the Differences or Residuals

The graph shows the resifuals, or differences between the observed value of y and the predicated value of y.

### Formulas for the Regression Line

the intercept

the slope of the line

Round the values of *a* and *b* to three decimal places.

### Procedure for Finding the Regression Line Equation

**Step 1** Make a table

**Step 2** Fill in the values of the table for each data point, and find the sum of each

column.

| ***x*** | ***y*** | ***xy*** | ***x2*** | ***y2*** |
| --- | --- | --- | --- | --- |
| **.**  **.**  **.** | **.**  **.**  **.** | **.**  **.**  **.** | **.**  **.**  **.** | **.**  **.**  **.** |
|  |  |  |  |  |

**Step 3** When *r* is significant, substitute in the formulas to find the values of *a* and *b* for the regression line equation .

### Example 10-7. Find the Regression Equation for the Data in 10-1, Absences and Final Grade.

**Step 1 & 2:**

| ***x*** | ***y*** | ***xy*** |  |  |
| --- | --- | --- | --- | --- |
| 6 | 82 | 492 | 36 | 6724 |
| 2 | 86 | 172 | 4 | 7396 |
| 15 | 43 | 645 | 225 | 1849 |
| 9 | 74 | 666 | 81 | 5476 |
| 12 | 58 | 696 | 144 | 3364 |
| 5 | 90 | 450 | 25 | 8100 |
| 8 | 78 | 624 | 64 | 6084 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| *57* | *511* | *3745* | *579* | *38995* |

**Step 3:**

(See Example 10-4.)

Using Technology: Correlation coefficient, r: 0.9442152

Significance of r:

For , and, the critical value of *r* is .

, so the value of *r* is significant.

Using Technology, the critical value and the *P*-value are calculated:

Critical r: ±0.7544919

*P*-value (two-tailed): 0.00137

.

Technology results concur:

Regression Results:

Y= b0 + b1x

Y Intercept, b0: 102.492537

Slope, b1: 3.621891

### Guidelines for Making Predictions

1. The points of the scatter plot \_\_\_\_\_the linear regression line reasonably well.
2. The value of *r* is \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_.
3. The value of a specific *x* is not much \_\_\_\_\_\_\_\_\_\_ the observed values (*x* values) in the original data.
4. If *r* is not significant, then the best predicted value for a specific *x* value is the \_\_\_\_\_\_\_\_\_\_\_\_ of the *y* values in the original data.

### Assumptions for Valid Predictions in Regression

* + 1. The sample is a \_\_\_\_\_\_\_\_\_\_\_\_\_ sample.
    2. For any specific value of the independent variable *x*, the value of the dependent variable *y* must be \_\_\_\_\_\_\_\_\_\_\_\_distributed about the regression line. (See Figures Illustrating the Assumptions for Predictions *a*).
    3. The standard deviation of the dependent variable must be the \_\_\_\_\_\_\_\_\_ for each value of the independent variable. (See Figures Illustrating the Assumptions for Predictions *b*)

### Figures Illustrating the Assumptions for Predictions

The first, part a, figure shows the dependent variable, y, is normally distributed for a given value of the independent variable, x.  The second, part b, shows that the standard deviations for the dependent variable is the same for each value of the independent variable.

### Example 10-8. Forest Fires and Acres Burned

The number of fires and the number of acres burned are as follows:

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Fires *x*** | 72 | 69 | 58 | 47 | 84 | 62 | 57 | 45 |
| **Acres *y*** | 62 | 42 | 19 | 26 | 51 | 15 | 30 | 15 |

Is the number of fires related to the acres burned?

Draw the scatter plot. Find the correlation coefficient. Determine its significance when . If significant, find ** and predict number of acres burned when fires occur.

*Solution:*

Scatter Plot:

| ***x*** | ***y*** | ***xy*** | ***x2*** | ***y2*** |
| --- | --- | --- | --- | --- |
| **.**  **.**  **.** | **.**  **.**  **.** | **.**  **.**  **.** | **.**  **.**  **.** | **.**  **.**  **.** |
|  |  |  |  |  |

Significance of r:

.

Prediction for fires.

### Definition: Marginal Change

The magnitude of the change in one variable when the other variable changes exactly 1 unit is the \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ change. The value of the slope *b* of the regression line equation represents the \_\_\_\_\_\_\_\_\_\_\_\_ change.

### Definition: Extrapolation

Making predictions beyond the bounds of the data is extrapolation. Be cautious. When predictions are made, they are made based on present conditions and on the premise that the present trends \_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_. That assumption may or may not be accurate.

Check the scatter plot for outliers. An outlier is a point that seems out of place when compared to the other points. When an outlier affects the equation of the regression line, the points are called **influential points** or **influential observations.**

# 10 – 3. Coefficient of Determination and Standard Error of Estimate.

### Types of Variation for Regression Model

In a regression model, it is possible to calculate the \_\_\_\_\_\_ variation, the \_\_\_\_\_\_\_\_\_\_\_\_ variation, and the \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ variation.

These are called deviations.

### Total Variation

The sum of the squares of the vertical distances (*y*) each point is from the mean) is the \_\_\_\_\_\_\_\_\_\_\_ variation. The total variation is made up of two parts: (1) that which is attributed to the \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ of *x* and *y* and (2) that which is due to \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_.

Total Variation

Explained variation + Unexplained variation

### Explained Variation

The variation obtained from the \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ between *x* and *y*, is the sum of the vertical distances between the predicted values of *y*, , and the mean of the values of *y*, .

Explained variation

### Unexplained Variation

The variation due to \_\_\_\_\_\_\_\_\_\_\_\_\_\_ is the vertical distance from the observed values of *y* and the predicted values of *y*, .

Unexplained variation

Illustration showing the total deviation, observed value of y minus the mean of y, made up of the sum of the unexplained deviation, the  observed value of y minus the predicted value of y, and the explained deviation, the predicted value of y minus the mean of y.

### Example 10-9. Find the Total, Explained, and Unexplained Variation for Number of Absences and Final Grades

For the data given in 10-1, find the total, explained, and unexplained variation.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Number of absences, *x*** | 6 | 2 | 15 | 9 | 12 | 5 | 8 |
| **Final grade, *y* (%)** | 82 | 86 | 43 | 74 | 58 | 90 | 78 |

| ***x*** | | ***y*** | | ***y’*** | |  |  | **Residual** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 6 | | 82 | | 80.761 | | 9 | 7.761 |  |
| 2 | | 86 | | 95.249 | | 13 | 22.249 |  |
| 15 | | 43 | | 48.164 | |  |  |  |
| 9 | | 74 | | 69.896 | | 1 |  |  |
| 12 | | 58 | | 59.030 | |  |  |  |
| 5 | | 90 | | 84.383 | | 17 | 11.383 |  |
| 8 | | 78 | | 73.517 | | 5 | 0.517 |  |
|  | |  | |  | |  |  |  |
|  |  | |  | |
| 81 | 60.233 | |  | |
| 169 | 495.018 | |  | |
|  |  | |  | |
| 1 |  | |  | |
|  |  | |  | |
| 289 | 129.573 | |  | |
| 25 | 0.267 | |  | |

Total Variation

Technology calculates the variation as well:

Total Variation: 1690

Explained Variation: 1506.706

Unexplained Variation: 183.2935

### Residual Plots

Residuals, , are also called \_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_. When the values are plotted with the *x* values, the plot, called a residual plot, can be used to determine how well the \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ line can be used to make predictions.

Use the plot to determine if the residuals form a pattern. If the residuals are more or less \_\_\_\_\_\_\_\_\_\_\_\_\_ distributed about the line, then the relationship between *x* and *y* is linear and the regression line can be used to make predictions.

### Example 10-10. Absences and Final Grade

The residuals, **,** calculated and shown in the table for Example 10-7 plotted with the respective values of *x*, show that the points are not evenly distributed about the line:

The points representing the residual for each value of x are plotted.  Three points are scattered below, but not close to the x-axis and four are scattered above it.  The points are not evenly distributed around the x-axis.

This plot shows that predictions are \_\_\_\_\_\_\_\_\_\_\_\_\_\_. However, the small sample size also makes the data questionable for making predictions.

## Objective 5. Compute the Coefficient of Determination.

### Coefficient of Determination

The coefficient of determination is a measure of variation of the \_\_\_\_\_\_\_\_\_\_\_\_\_\_ variable that is explained by the regression line and the independent variable. It is the \_\_\_\_\_\_\_\_\_\_\_ of explained variation and total variation.

When the decimal value is converted to a percentage, it gives the percentage of variation in the dependent variable, *y*, that is \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ for by the variation in the independent variable, *x*.

For instance, using the data from Example 10-1 for absences and final grades, the correlation coefficient is . Then . Therefore, \_\_\_\_\_\_\_\_% of the variation of final grades can be \_\_\_\_\_\_\_\_\_\_\_\_\_\_ by the variation of the number of absences.

### Coefficient of Nondetermination

The coefficient of nondetermination is the \_\_\_\_\_\_\_\_\_\_\_\_\_\_ between 1 and the coefficient of determination, , and represents the percentage of variation in the dependent variable , *y*, that \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ by the variation of the independent variable, *x*.

For instance, using the data from Example 10-1 for absences and final grades, the coefficient of nondetermination is or 10.9% of the variation in final grades \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ by the variation of the number of absences.

## Objective 6. Compute the Standard Error of the Estimate.

A value is a \_\_\_\_\_\_\_\_\_\_\_ prediction of the dependent variable for a given value of *x*. The disadvantage of a point prediction is that there is no information about how \_\_\_\_\_\_\_\_\_\_\_\_\_\_ the point prediction is. To overcome this disadvantage, we use a prediction, or interval, estimate. The calculation of the prediction interval uses the \_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_ \_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_.

### Standard Error of Estimate, ,

The standard deviation of the observed *y* values about the predicted values measures how the data points \_\_\_\_\_\_\_\_\_\_\_\_ from the regression line.

The formula for the standard error of estimate is

.

### Finding the Standard Error of Estimate

**Steps 1 – 5** Complete the table:

| ***x*** | ***y*** | ***y’*** |  |  |
| --- | --- | --- | --- | --- |
| **.**  **.**  **.** | **.**  **.**  **.** | **.**  **.**  **.** | **.**  **.**  **.** | **.**  **.**  **.** |
|  |  |  |  |  |

**Step 6** Substitute in the formula and find .

### Example 10-11 Standard Error of Estimate for Absences and Final Grade

Using the data from 10-1, find the standard error of estimate.

| ***x*** | ***y*** | ***y’*** |  |  |
| --- | --- | --- | --- | --- |
| 6 | 82 | 80.761 |  |  |
| 2 | 86 | 95.249 |  |  |
| 15 | 43 | 48.164 |  |  |
| 9 | 74 | 69.896 |  |  |
| 12 | 58 | 59.030 |  |  |
| 5 | 90 | 84.383 |  |  |
| 8 | 78 | 73.517 |  |  |
|  |  |  |  | **=** |

Using technology, the standard error of estimate is .

## Objective 7. Find a Prediction Interval.

### Formula for the Prediction Interval about a value of

with d.f. .

### Procedure Table

**Step 1** Find , , and .

**Step 2** Find for the specific *x* value.

**Step 3** Find .

**Step 4** Substitute in the formula and evaluate.

### Example 10-12. Prediction Interval for Absences and Final Grades for *x* = 10, .

**Step 1**

From Example 10-7 we know

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| *57* | *511* | *3745* | *579* | *38995* |

**Step 2**

We also know .

Let x = 10, .

**Step 3**

**Step 4**

, with d.f. = 5.

You can be 95% confident that the interval contains the actual value of *y*.

That is, for a person who has 10 absences, we can be 95% confident that their final grade will be between \_\_\_\_\_\_\_\_\_\_ and \_\_\_\_\_\_\_\_\_\_\_\_\_. The range is large because the sample size is small and the standard error of estimate is 6.055.

# 10 – 4.Multiple Regression (Optional)

## Objective 8. Be Familiar with the Concept of Multiple Regression.

In simple linear regression, the regression equation contains one independent variable *x* and one dependent variable , written as .

### General Form of the Multiple Regression Equation

In multiple regression, there are several independent variables and one dependent variable, the equation is

,

where , , , are the independent variables.

For example, a researcher may want to determine if there is a significant relationship between independent variables of the a person’s number of years of experience at a certain job, , the person’s number of workdays missed during the past month, , the person’s number of overtime hours during the past month, , and the dependent variable of the person’s salary, .

Another example is when a store manager wants to see if the amount spent on advertising and the amount of display space given a particular product affects the amount of sales of that product.

### Assumptions for Multiple Regression

The assumptions for multiple regression are similar to those for simple regression.

1. *Normality Assumption.* For any specific value of the independent variable, the values of the *y* variable are \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ distributed.
2. *Equal-variance Assumption.* The variances (or standard deviations) for the *y* variables are the \_\_\_\_\_\_\_\_\_\_ for each value of the independent variable.
3. *Linearity Assumption.* There is a \_\_\_\_\_\_\_\_\_\_\_\_ relationship between the dependent variable and the independent variables.
4. *Nonmulticollinearity Assumption.* The independent variables are not \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_.
5. *Independence Assumption.* The values for the *y* variables are \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_.

A multiple regression correlation *R* can be computed to determine if a significant relationship exists between the independent variables and the dependent variable. This process might be used to \_\_\_\_\_\_\_\_\_\_\_\_\_ the accuracy of predictions for the dependent variable over predictions using one independent variable.

The value of *R* can range from 0 and +1. The closer to 0, the \_\_\_\_\_\_\_\_\_\_\_\_ the relationship; the closer to +1, the \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ the relationship. The multiple correlation coefficient is always \_\_\_\_\_\_\_\_\_\_\_\_ than each of the individual correlation coefficients.

### Formula for the Multiple Correlation Coefficient with Two Independent Variables

The formula for *R* is

where is the value of the correlation coefficient for variables *y* and ;

is the value of the correlation coefficient for variables *y* and ; and

is the value of the correlation coefficient for variables and .

### Example 10-13a. Find *R* When Given Values for the Correlation Coefficients

Find *R* when , , and .

*Solution:*

is the coefficient of multiple determination and represents the amount of variation explained by the regression model. represents the amount of unexplained variation and is called the *error or residual variation.*

An *F* Test is used to test the significance of *R* for the hypotheses and where represents the population correlation coefficient for multiple correlation.

### F Test for Significance of *R*

The formula for the F Test for significance of *R* is

where *n* is the number of data groups (, ,, *y*) and *k* is the number of independent variables.   
The degrees of freedom are and .

### Example 10-13b. Find the Significance of *R* if

The value of is dependent on the number of data pairs, *n*, and the number of variables, *k*, so statisticians calculate an adjusted . The adjusted , , is smaller than , and takes into the account the fact that when *n* and *k* are approximately equal, the value of *R* may be artificially high. This happens because the chance variation of all the variables are used in conjunction with one another to derive the regression equation. Usually, both and are both reported in multiple regression analysis.

### Formula for Adjusted

The formula for the adjusted is

### Example 10-13c. Find the .