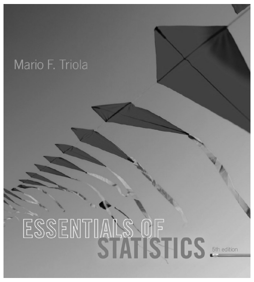
Lecture Slides



Essentials of Statistics
5th Edition

and the Triola Statistics Series

by Mario F. Triola

Chapter 10
Correlation and Regression

10-1 Review and Preview

10-2 Correlation

10-3 Regression

10-4 Rank Correlation

Review

In Chapter 9 we presented methods for making inferences from two samples.

In Section 9-4 we considered two dependent samples, with each value of one sample somehow paired with a value from the other sample, and we illustrated the use of hypothesis tests for claims about the population of differences.

We also illustrated the construction of confidence-interval estimates of the mean of all such differences.

In this chapter we again consider paired sample data, but the objective is fundamentally different from that of Section 9-4.

11

Preview

In this chapter we introduce methods for determining whether a correlation, or association, between two variables exists and whether the correlation is linear.

For linear correlations, we can identify an equation that best fits the data and we can use that equation to predict the value of one variable given the value of the other variable.

In addition, we consider methods for identifying linear equations for correlations among two variables.

Chapter 10
Correlation and Regression

10-1 Review and Preview

10-2 Correlation

10-3 Regression

10-4 Rank Correlation

Key Concept

Part 1 of this section introduces the linear correlation coefficient, r, which is a number that measures how well paired sample data fit a straight-line pattern when graphed.

Using paired sample data (sometimes called bivariate data), we find the value of r (usually using technology), then we use that value to conclude that there is (or is not) a linear correlation between the two variables.

Key Concept

In this section we consider only linear relationships, which means that when graphed, the points approximate a straight-line pattern.

In Part 2, we discuss methods of hypothesis testing for correlation.

Part 1: Basic Concepts of Correlation

Definition

A correlation exists between two variables when the values of one are somehow associated with the values of the other in some way.

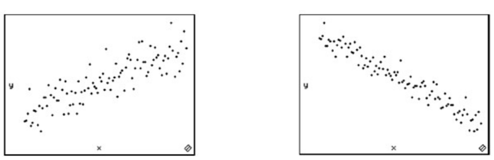
A linear correlation exists between two variables when there is a correlation and the plotted points of paired data result in a pattern that can be approximated by a straight line.

Exploring the Data

We can often see a relationship between two variables by constructing a scatterplot.

The following slides show scatterplots with different characteristics.

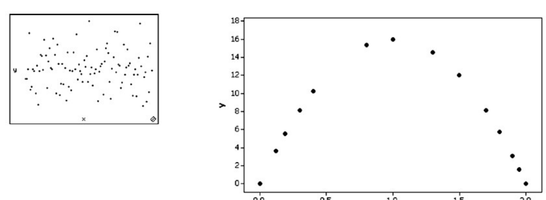
Scatterplots of Paired Data



(a) **Positive correlation: $r = 0.851$**

(b) **Negative correlation: $r = -0.965$**

Scatterplots of Paired Data



(c) **No correlation: $r = 0$**

(d) **Nonlinear relationship: $r = -0.087$**

Requirements for Linear Correlation

1. The sample of paired (*x, y*) data is a simple random sample of quantitative data.
2. Visual examination of the scatterplot must confirm that the points approximate a straight-line pattern.
3. The outliers must be removed if they are known to be errors. The effects of any other outliers should be considered by calculating *r* with and without the outliers included.

Notation for the Linear Correlation Coefficient

n	number of pairs of sample data
$\sum$	denotes the addition of the items indicated
$\sum x$	sum of all x-values
$\sum x^2$	indicates that each x-value should be squared and then those squares added
$\left(\sum x\right)^2$	indicates that each x-value should be added and the total then squared
$\sum xy$	indicates each x-value is multiplied by its corresponding y-value. Then sum those up.
r	linear correlation coefficient for sample data
ρ	linear correlation coefficient for a population of paired data

Formula

The linear correlation coefficient *r* measures the strength of a linear relationship between the paired values in a sample. Here are two formulas:

$$r = \frac{n\left(\Sigma xy\right) - (\Sigma x)(\Sigma y)}{\sqrt{n(\Sigma x^2) - (\Sigma x^2)}\sqrt{n(\Sigma y^2) - (\Sigma y^2)}}$$

$$r = \frac{\sum\left(z_x z_y\right)}{n-1}$$

Technology can (and should) compute this

Interpreting *r*

Using Table A-6: If the absolute value of the computed value of *r*, exceeds the value in Table A-6, conclude that there is a linear correlation. Otherwise, there is not sufficient evidence to support the conclusion of a linear correlation.

Using Software: If the computed *P*-value is less than or equal to the significance level, conclude that there is a linear correlation. Otherwise, there is not sufficient evidence to support the conclusion of a linear correlation.

Caution

Know that the methods of this section apply to a *linear* correlation.

If you conclude that there does not appear to be linear correlation, know that it is possible that there might be some other association that is not linear.

Properties of the Linear Correlation Coefficient *r*

1. $-1 \le r \le 1$
2. If all values of either variable are converted to a different scale, the value of *r* does not change.
3. The value of *r* is not affected by the choice of *x* and *y*. Interchange all *x*- and *y*-values and the value of *r* will not change.
4. *r* measures strength of a linear relationship.
5. *r* is very sensitive to outliers, which can dramatically affect the value of *r*.

Example

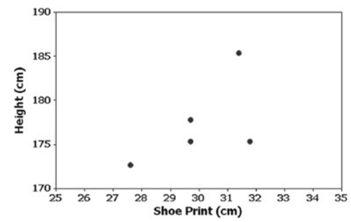
The paired shoe / height data from five males are listed below. Use a computer or a calculator to find the value of the correlation coefficient *r*.

Table 10-1 Shoe Print Lengths and Heights of Males

Shoe Print (cm)	29.7	29.7	31.4	31.8	27.6
Height (cm)	175.3	177.8	185.4	175.3	172.7

Example - Continued

Requirement Check: The data are a simple random sample of quantitative data, the plotted points appear to roughly approximate a straight-line pattern, and there are no outliers.



Example - Continued

A few technologies are displayed below, used to calculate the value of *r*.

MINITAB

```
Correlations: Shoe Print, Height

Pearson correlation of Shoe Print and Height = 0.591
P-Value = 0.294
```

TI-83/84 PLUS

```
LinRegTTest
 y=a+bx
 β≠0 and ρ≠0
↑b=1.72745239
 s=4.538762333
 r²=.3495991312
 r=.5912690853
```

STATDISK

```
Correlation Results:
Correlation coeff, r:  0.5912691
Critical r:           ±0.8783393
P-value (two-tailed):  0.29369
```


Using the Formulas to Calculate Correlation

Technology is highly recommended, and as such, we refer you to the textbook, pages 501 and 502 for the manual calculations using the formulas.

Is There a Linear Correlation?

We found previously for the shoe and height example that *r* = 0.591.

We now proceed to interpret its meaning.

Our goal is to decide whether or not there appears to be a linear correlation between shoe print lengths and heights of people.

We can base our interpretation on a *P*-value or a critical value from Table A-6.

Interpreting the Linear Correlation Coefficient *r*

Using computer software:

If the *P*-value is less than the level of significance, conclude there is a linear correlation.

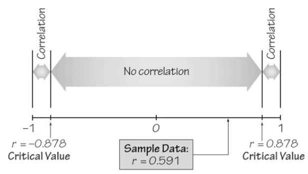
Our example with technologies provided a *P*-value of 0.294.

Because that *P*-value is not less than the significance level of 0.05, we conclude there is not sufficient evidence to support the conclusion that there is a linear correlation between shoe print length and heights of people.

Interpreting the Linear Correlation Coefficient *r*

Using Table A-6:

Table A-6 yields $r = 0.878$ for five pairs of data and a 0.05 level of significance. Since our correlation was $r = 0.591$, we conclude there is not sufficient evidence to support the claim of a linear correlation.



Interpreting *r*: Explained Variation

The value of r^2 is the proportion of the variation in y that is explained by the linear relationship between x and y.

Example

We found previously for the shoe and height example that $r = 0.591$.

With $r = 0.591$, we get $r^2 = 0.349$.

We conclude that about 34.9% of the variation in height can be explained by the linear relationship between lengths of shoe prints and heights.

Common Errors Involving Correlation

1. Causation: It is wrong to conclude that correlation implies causality.
2. Averages: Averages suppress individual variation and may inflate the correlation coefficient.
3. Linearity: There may be some relationship between x and y even when there is no linear correlation.

Caution

Know that correlation does not imply causality.

Part 2: Formal Hypothesis Test

Formal Hypothesis Test

We wish to determine whether there is a significant linear correlation between two variables.

Notation:

n = number of pairs of sample data

r = linear correlation coefficient for a sample of paired data

ρ = linear correlation coefficient for a population of paired data

Hypothesis Test for Correlation Requirements

1. The sample of paired (x, y) data is a simple random sample of quantitative data.
2. Visual examination of the scatterplot must confirm that the points approximate a straight-line pattern.
3. The outliers must be removed if they are known to be errors. The effects of any other outliers should be considered by calculating r with and without the outliers included.

Hypothesis Test for Correlation Hypotheses

$H_0: \rho = 0$ (There is no linear correlation.)

$H_1: \rho \neq 0$ (There is a linear correlation.)

Test Statistic: r

Critical Values: Refer to Table A-6.

P-values: Refer to technology.

Hypothesis Test for Correlation

If $|r|$ > critical value from Table A-6, reject the null hypothesis and conclude that there is sufficient evidence to support the claim of a linear correlation.

If $|r| \leq$ critical value from Table A-6, fail to reject the null hypothesis and conclude that there is not sufficient evidence to support the claim of a linear correlation.

Example

We found previously for the shoe and height example that $r = 0.591$.

Conduct a formal hypothesis test of the claim that there is a linear correlation between the two variables.

Use a 0.05 significance level.

Example - Continued

We test the claim:

$H_0: \rho = 0$ (There is no linear correlation)

$H_1: \rho \neq 0$ (There is a linear correlation)

With the test statistic $r = 0.591$ from the earlier example. The critical values of $r = \pm 0.878$ are found in Table A-6 with $n = 5$ and $\alpha = 0.05$.

We fail to reject the null and conclude there is not sufficient evidence to support the claim of a linear correlation.

P-Value Method for a Hypothesis Test for Linear Correlation

The test statistic is below, use *n* – 2 degrees of freedom.

$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}}$$

P-values can be found using software or Table A-3.

Example

Continuing the same example, we calculate the test statistic:

$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} = \frac{0.591}{\sqrt{\frac{1-0.591^2}{5-2}}} = 1.269$$

Table A-3 shows this test statistic yields a *P*-value that is greater than 0.20. Technology provides the *P*-value as 0.2937.

Example - Continued

Because the *P*-value of 0.2937 is greater than the significance level of 0.05, we fail to reject the null hypothesis.

We conclude there is not sufficient evidence to support the claim of a linear correlation between shoe print length and heights.

One-Tailed Tests

One-tailed tests can occur with a claim of a positive linear correlation or a claim of a negative linear correlation. In such cases, the hypotheses will be as shown here.

Claim of *Negative* Correlation (Left-tailed test)	**Claim of *Positive* Correlation (Right-tailed test)**
H_0: $\rho = 0$ H_1: $\rho < 0$	H_0: $\rho = 0$ H_1: $\rho > 0$

For these one-tailed tests, the *P*-value method can be used as in earlier chapters.

Chapter 10 Correlation and Regression

Key Concept

In Part 1 of this section we find the equation of the straight line that best fits the paired sample data. That equation algebraically describes the relationship between two variables.

The best-fitting straight line is called a regression line and its equation is called the regression equation.

In Part 2, we discuss marginal change, influential points, and residual plots as tools for analyzing correlation and regression results.

Part 1: Basic Concepts of Regression

Regression

The regression equation expresses a relationship between x (called the explanatory variable, predictor variable or independent variable), and $\hat{y}$ (called the response variable or dependent variable).

The typical equation of a straight line $y = mx + b$ is expressed in the form $\hat{y} = b_0 + b_1x$, where b_0 is the y-intercept and b_1 is the slope.

Definitions

Regression Equation:

Given a collection of paired sample data, the regression line (or line of best fit, or least-squares line) is the straight line that "best" fits the scatterplot of data.

The regression equation $\hat{y} = b_0 + b_1x$ algebraically describes the regression line.

Notation for Regression Equation

	Population Parameter	Sample Statistic
y-Intercept of regression equation	β_0	b_0
Slope of regression equation	β_1	b_1
Equation of the regression line	$y = \beta_0 + \beta_1x$	$\hat{y} = b_0 + b_1x$

Requirements

1. The sample of paired (x, y) data is a random sample of quantitative data.
2. Visual examination of the scatterplot shows that the points approximate a straight-line pattern.
3. Any outliers must be removed if they are known to be errors. Consider the effects of any outliers that are not known errors.

Formulas for b_1 and b_0

Slope: $$b_1 = r\frac{s_y}{s_x}$$

y-intercept: $$b_0 = \overline{y} - b_1\overline{x}$$

Technology will compute these values.

Example

Let us return to the example from Section 10.2. We would like to use the explanatory variable, *x*, shoe print length, to predict the response variable, *y*, height.

The data are listed below:

Table 10-1 Shoe Print Lengths and Heights of Males					
Shoe Print (cm)	29.7	29.7	31.4	31.8	27.6
Height (cm)	175.3	177.8	185.4	175.3	172.7

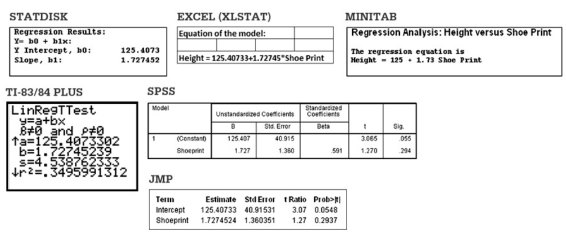
Example - Continued

Requirement Check:

1. The data are assumed to be a simple random sample.
2. The scatterplot showed a roughly straight-line pattern.
3. There are no outliers.

The use of technology is recommended for finding the equation of a regression line.

Example – Continued



STATDISK

```
Regression Results:
Y= b0 + b1x:
Y Intercept, b0:      125.4073
Slope, b1:            1.727452
```

EXCEL (XLSTAT)

Equation of the model:

Height = 125.40733+1.72745*Shoe Print

MINITAB

Regression Analysis: Height versus Shoe Print

```
The regression equation is
Height = 125 + 1.73 Shoe Print
```

TI-83/84 PLUS

```
LinRegTTest
 y=a+bx
 β≠0 and ρ≠0
↑a=125.4073302
 b=1.72745239
 s=4.538762333
↓r²=.3495991312
```

SPSS

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	125.407	40.915		3.065	.055
	Shoeprint	1.727	1.360	.591	1.270	.294

JMP

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	125.40733	40.91531	3.07	0.0548
Shoeprint	1.7274524	1.360351	1.27	0.2937

Example - Continued

All these technologies show that the regression equation can be expressed as:

$$\hat{y} = 125 + 1.73x$$

Now we use the formulas to determine the regression equation (technology is recommended).

Example

Recall from the previous section that *r* = 0.591269.

Technology can be used to find the values of the sample means and sample standard deviations used below.

$$b_1 = r\frac{s_y}{s_x} = 0.591269 \cdot \frac{4.87391}{1.66823} = 1.72745$$

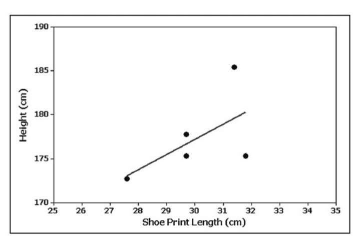
$$b_0 = y - b_1 x = 177.3 - (1.72745)(30.04) = 125.40740$$

(These are the same coefficients found using technology)

Example

Graph the regression equation on a scatterplot:

$$\hat{y} = 125 + 1.73x$$



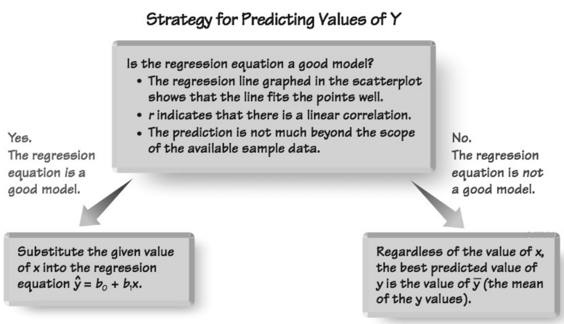
Using the Regression Equation for Predictions

1. Use the regression equation for predictions only if the graph of the regression line on the scatterplot confirms that the regression line fits the points reasonably well.
2. Use the regression equation for predictions only if the linear correlation coefficient *r* indicates that there is a linear correlation between the two variables (as described in Section 10-2).

Using the Regression Equation for Predictions

3. Use the regression line for predictions only if the data do not go much beyond the scope of the available sample data. (Predicting too far beyond the scope of the available sample data is called extrapolation, and it could result in bad predictions.)
4. If the regression equation does not appear to be useful for making predictions, the best predicted value of a variable is its sample mean.

Strategy for Predicting Values of *y*



Using the Regression Equation for Predictions

If the regression equation is not a good model, the best predicted value of *y* is simply $\bar{y}$, the mean of the *y* values.

Remember, this strategy applies to linear patterns of points in a scatterplot.

If the scatterplot shows a pattern that is not a straight-line pattern, other methods apply, as described in Section 10-6.

Example

Use the 5 pairs of shoe print lengths and heights to predict the height of a person with a shoe print length of 29 cm.

The regression line does not fit the points well. The correlation is $r = 0.591$, which suggests there is not a linear correlation (the *P*-value was 0.294).

The best predicted height is simply the mean of the sample heights:

$$\bar{y} = 177.3 \text{ cm}$$

Example

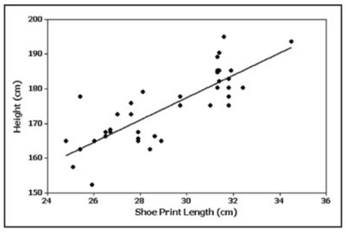
Use the 40 pairs of shoe print lengths from Data Set 2 in Appendix B to predict the height of a person with a shoe print length of 29 cm.

Now, the regression line does fit the points well, and the correlation of $r = 0.813$ suggests that there is a linear correlation (the *P*-value is 0.000).

Example - Continued

Using technology we obtain the regression equation and scatterplot:

$$\hat{y} = 80.9 + 3.22x$$



Example - Continued

The given shoe length of 29 cm is not beyond the scope of the available data, so substitute in 29 cm into the regression model:

$$\begin{aligned}\hat{y} &= 80.9 + 3.22x \\ &= 80.9 + 3.22(29) \\ &= 174.3 \text{ cm}\end{aligned}$$

A person with a shoe length of 29 cm is predicted to be 174.3 cm tall.

Part 2: Beyond the Basics of Regression

Definition

In working with two variables related by a regression equation, the marginal change in a variable is the amount that it changes when the other variable changes by exactly one unit.

The slope b_1 in the regression equation represents the marginal change in *y* that occurs when *x* changes by one unit.

Example

For the 40 pairs of shoe print lengths and heights, the regression equation was:

$$\hat{y} = 80.9 + 3.22x$$

The slope of 3.22 tells us that if we increase shoe print length by 1 cm, the predicted height of a person increases by 3.22 cm.

Definition

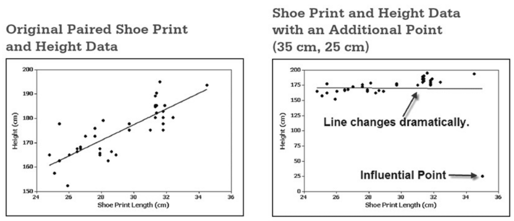
In a scatterplot, an outlier is a point lying far away from the other data points.

Paired sample data may include one or more influential points, which are points that strongly affect the graph of the regression line.

Example

For the 40 pairs of shoe prints and heights, observe what happens if we include this additional data point:

x = 35 cm and y = 25 cm



Example - Continued

The additional point is an influential point because the graph of the regression line because the graph of the regression line did change considerably.

The additional point is also an outlier because it is far from the other points.

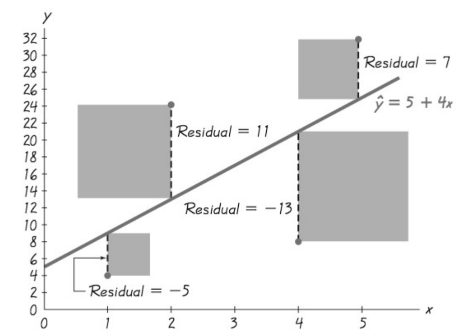
Definition

For a pair of sample *x* and *y* values, the residual is the difference between the observed sample value of *y* and the *y*-value that is predicted by using the regression equation.

That is:

$$\text{residual} = \text{observed } y - \text{predicted } y = y - \hat{y}$$

Residuals



Definition

A straight line satisfies the least-squares property if the sum of the squares of the residuals is the smallest sum possible.

Definition

A residual plot is a scatterplot of the (x, y) values after each of the y-coordinate values has been replaced by the residual value $y - \hat{y}$ (where $\hat{y}$ denotes the predicted value of y).

That is, a residual plot is a graph of the points (x, $y - \hat{y}$).

Residual Plot Analysis

When analyzing a residual plot, look for a pattern in the way the points are configured, and use these criteria:

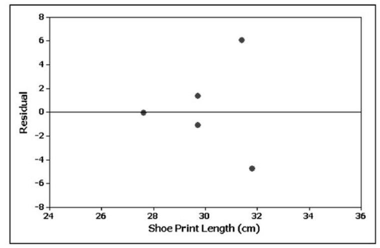
The residual plot should not have any obvious patterns (not even a straight line pattern). This confirms that the scatterplot of the sample data is a straight-line pattern.

The residual plot should not become thicker (or thinner) when viewed from left to right. This confirms the requirement that for different fixed values of *x*, the distributions of the corresponding *y* values all have the same standard deviation.

Example

The shoe print and height data are used to generate the following residual plot:

MINITAB



Example - Continued

The residual plot becomes thicker, which suggests that the requirement of equal standard deviations is violated.

Example - Continued

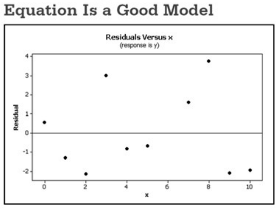
On the following slides are three residual plots.

Observe what is good or bad about the individual regression models.

Example - Continued

Regression model is a good model:

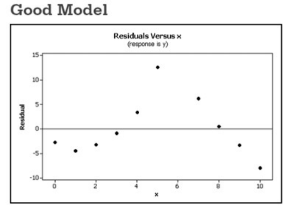
Residual Plot Suggesting That the Regression Equation Is a Good Model



Example - Continued

Distinct pattern: sample data may not follow a straight-line pattern.

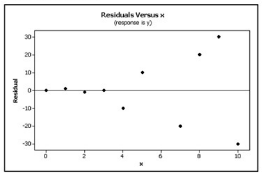
Residual Plot with an Obvious Pattern, Suggesting That the Regression Equation Is Not a Good Model



Example - Continued

Residual plot becoming thicker: equal standard deviations violated.

Residual Plot That Becomes Thicker, Suggesting That the Regression Equation Is Not a Good Model



Complete Regression Analysis

1. Construct a scatterplot and verify that the pattern of the points is approximately a straight-line pattern without outliers. (If there are outliers, consider their effects by comparing results that include the outliers to results that exclude the outliers.)
2. Construct a residual plot and verify that there is no pattern (other than a straight-line pattern) and also verify that the residual plot does not become thicker (or thinner).

Complete Regression Analysis

3. Use a histogram and/or normal quantile plot to confirm that the values of the residuals have a distribution that is approximately normal.

4. Consider any effects of a pattern over time.