

Overview

- Scatterplots
- Explanatory and Response Variables
- Describing Association
- The Regression Equation and Interpreting the Slope and Intercept
- Correlation, r , Lurking Variables and Causation
- Explained Variance, R^2
- Residuals and Residual Plots
- Making Predictions: Interpolation vs. Extrapolation

Exploring relationships between two quantitative variables

Example 1. A survey was conducted in the United States and 10 countries of Western Europe to determine the percentage of teenagers who had used marijuana and other drugs. The results are summarized in the following table.

Country	% who have used Marijuana	% who have used other drugs
Czech Republic	22	4
Denmark	17	3
England	40	21
Finland	5	1
Ireland	37	16
Italy	19	8
No. Ireland	23	14
Norway	6	3
Portugal	7	3
Scotland	53	31
United States	34	24

- Do you think there might be a relationship between Marijuana use and other drug use?
- Make a scatterplot of the data using Excel.

To make a scatterplot in Excel:

1. Enter the data in two columns
2. Select both columns with headers
3. Select Insert -> Scatter. Select the option with markers only

To make a scatterplot in GeoGebra:

1. Select View -> Spreadsheet and enter the data in two columns
2. Select both columns
3. Click on the histogram icon and select Two Variable Regression Analysis. Click Analyze and then click on $\sum x$ to show statistics.

Explanatory and Response Variables

The **response variable** is the dependent variable (y). In our example:

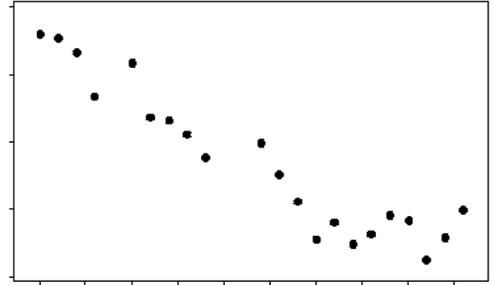
The **explanatory variable** is the independent variable (x). In our example:

We think that changes in the explanatory variable might explain changes in the response variable.

A Framework to Describe Association

Describe four features of the association between two variables:

- Direction
- Form
- Strength
- Unusual Features (subgroups or outliers)



Direction:

positive

An empty coordinate system with a vertical y-axis and a horizontal x-axis, intended for describing a positive association.

negative

An empty coordinate system with a vertical y-axis and a horizontal x-axis, intended for describing a negative association.

neither

An empty coordinate system with a vertical y-axis and a horizontal x-axis, intended for describing an association that is neither positive nor negative.

Form:

linear

An empty coordinate system with a vertical y-axis and a horizontal x-axis, intended for describing a linear association.

curved

An empty coordinate system with a vertical y-axis and a horizontal x-axis, intended for describing a curved association.

no pattern

An empty coordinate system with a vertical y-axis and a horizontal x-axis, intended for describing an association with no discernible pattern.

Strength:

strong

An empty coordinate system with a vertical y-axis and a horizontal x-axis, intended for describing a strong association.

moderate

An empty coordinate system with a vertical y-axis and a horizontal x-axis, intended for describing a moderate association.

weak

An empty coordinate system with a vertical y-axis and a horizontal x-axis, intended for describing a weak association.

Unusual Features:

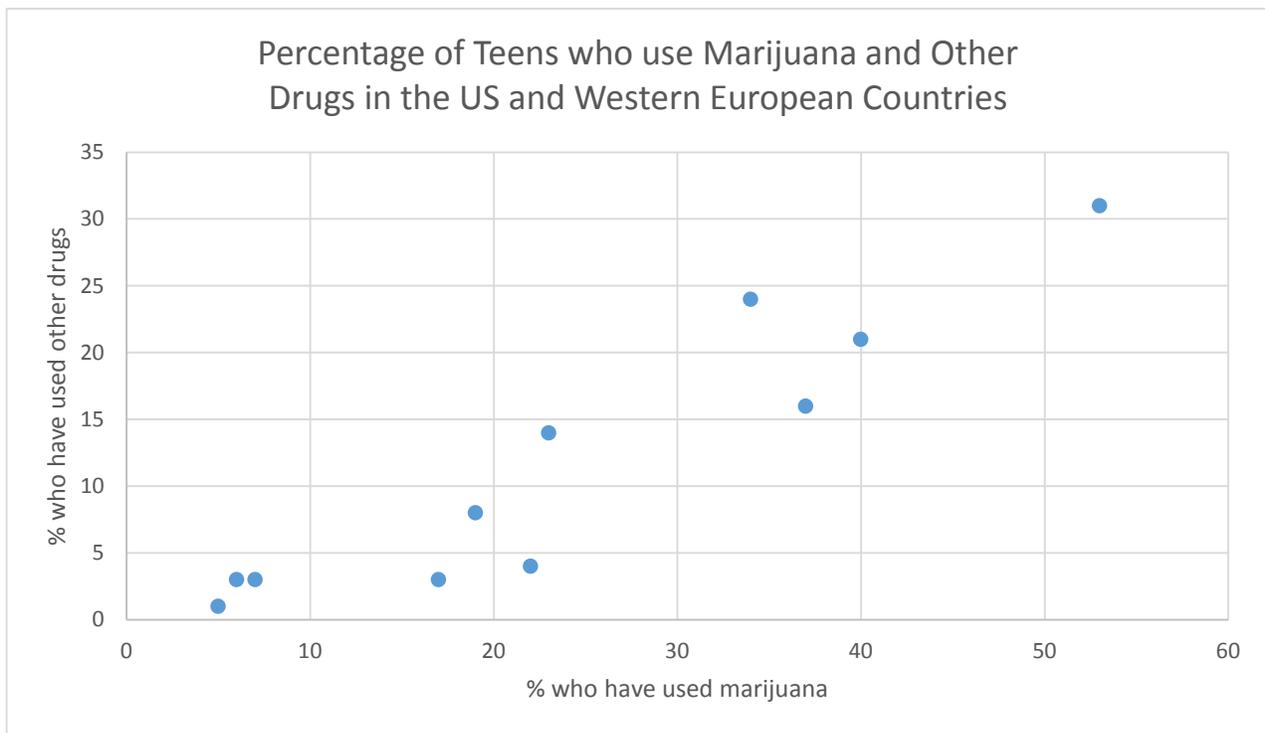
groupings

An empty coordinate system with a vertical y-axis and a horizontal x-axis, intended for describing an association with groupings or subgroups.

outliers

An empty coordinate system with a vertical y-axis and a horizontal x-axis, intended for describing an association with outliers.

Example 1 Continued. Here is the scatterplot made in Excel:



c. Write a description of the association including all four characteristics.

The Line of Best Fit

d. On the graph above, use a ruler or straightedge to draw a line that models this relationship.

e. Draw the vertical distance between each point and the line. These are the residuals.

$$\text{Residual} = \text{observed value} - \text{predicted value}$$

The least squares regression line is the line that minimizes the sum of the squared residuals (deviations of y). We will use technology to calculate this for us.

The Least Squares Regression Line, $\hat{y} = mx + b$

Recall the equation of a line from algebra: $y = mx + b$

m represents the _____ b represents the _____

\hat{y} is read "y-hat," and represents the predicted value of y . The values of m and b are the parameters of the linear model.

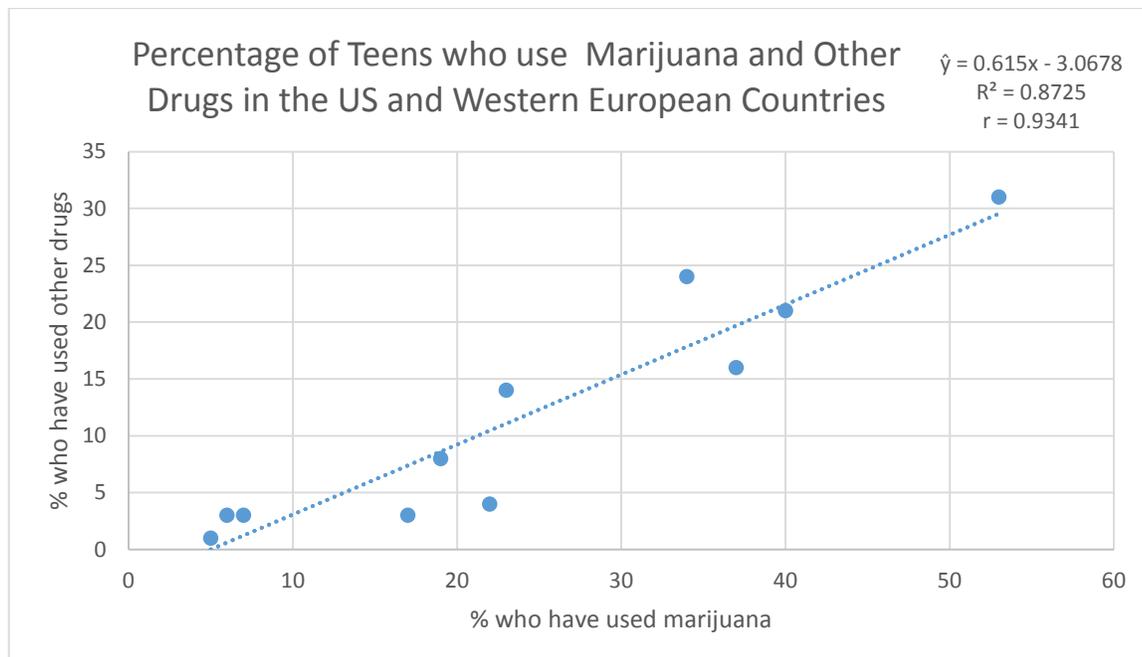
Calculate the Regression Line with Excel

1. Click on your scatterplot
2. Depending on your version of Excel:
 - If you have symbols on the right side of your scatterplot, Click on the + sign for chart elements. Check the box next to Trendline and click on the right arrow and select more options. Check "Display Equation on Chart" and "Display R-squared value on chart."
 - If you do not have symbols on the right, select Layout -> Trendline -> Linear Trendline from the menu bar at the top. Select "More Trendline Options" and check "Display Equation on Chart" and "Display R-squared value on chart."

Calculate the Regression Line with GeoGebra

1. Under Regression Model in the lower left corner select Linear.
2. Back in the spreadsheet screen select Options -> Rounding -> 4 decimal places

f. Calculate the regression line for our example (Excel shown):



g. Write the equation of the least squares regression line from Excel, using \hat{y} :

h. Give the slope and interpret it in the context of the data:

i. Give the y-intercept and interpret it in the context of the data:

Correlation, r

Caution: Do not say correlation when you mean association.

Association is a vague term describing a relationship between two variables.

Correlation is a precise value describing the strength and direction of the linear relationship.



Characteristics of r:

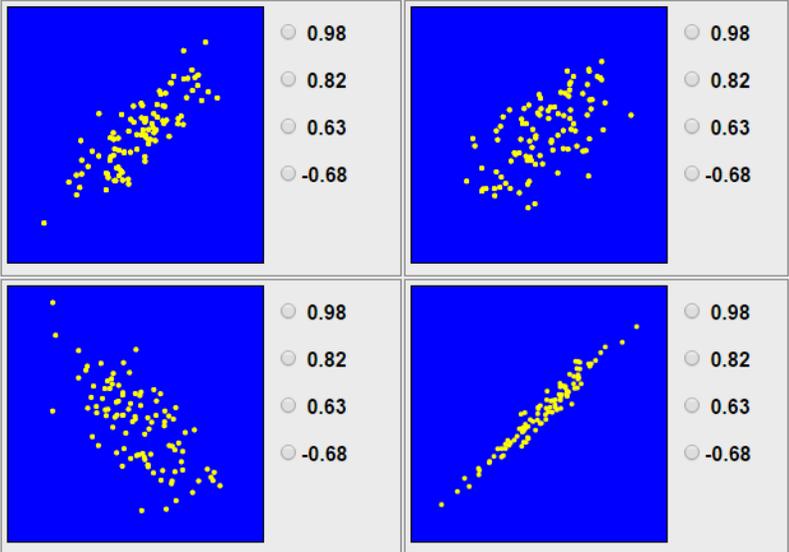
- The value of r only makes sense for linear relationships
- r has no units
- The sign of r is the direction of the association.

We will not calculate r by hand, but we can use the formula to understand what it means:

$$r = \frac{\sum(x-\bar{x})(y-\bar{y})}{(n-1)s_x s_y} \quad \text{or} \quad r = \frac{\sum z_x z_y}{n-1}$$

Activity: Matching Correlations Applet. Follow the link and try matching the correlations.
<http://www.istics.net/Correlations/>

Guessing Correlations



The applet displays four scatter plots on a blue background with yellow data points. Each plot is accompanied by four radio button options: 0.98, 0.82, 0.63, and -0.68. The top-left plot shows a moderate positive correlation. The top-right plot shows a moderate negative correlation. The bottom-left plot shows a strong positive correlation. The bottom-right plot shows a strong negative correlation.

Match the correlations with the scatter plots. Check answers

j. What is the correlation between the percent of teens who have used marijuana and the percent who have used other drugs? What does this mean?

k. Do these results show that marijuana is a “gateway drug,” that is, that marijuana use leads to the use of other drugs? Explain.



Correlation does not imply causation!

Lurking variables

A lurking variable is a hidden variable that is responsible for the apparent association.

For example, internet use is on the rise and people have more pets. Can we say that using the internet causes one to have more pets?

l. What lurking variables could cause the relationship between marijuana use and other drug use?

The Coefficient of Determination, r^2 or R^2

$$R^2 = 1 - \frac{\text{Sum of the squares of the residuals}}{\text{Sum of the squares of the } y - \text{values}}$$

$$\text{In GeoGebra, } R^2 = 1 - \frac{SSE}{S_{yy}}$$

The coefficient of determination, R^2 , tells us the percentage of the variation in the y-values that can be explained by the least squares regression line of y on x.

Approximately _____% of the variation in _____

can be explained by the variation in _____.

Therefore, _____ % of the variation in _____ is due to other variables.

m. In our example, R^2 means:

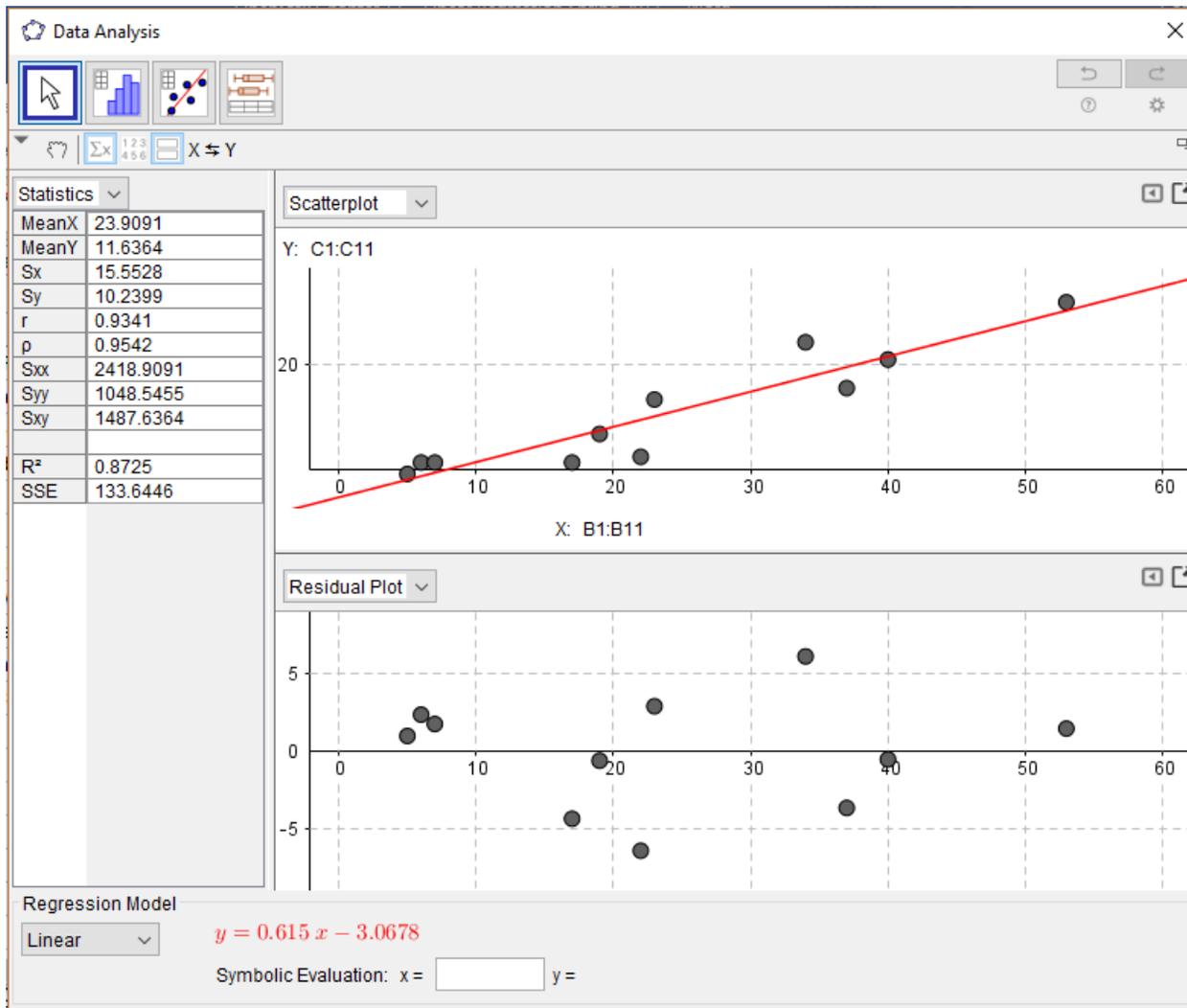
Residual Plots: $(x, y - \hat{y})$

The residual is the vertical distance (error) between the data point and the predicted value (point on the regression line). It is important to examine the residual plot to see if there are any patterns that we could not see by looking at the scatterplot. We can also look at residuals to identify outliers.

n. Looking at the scatterplot with the regression line, which country has the largest residual?

Making a Residual Plot in GeoGebra

1. Click on the icon for "show second plot," that looks like an equal sign.
2. Select Residual Plot from the dropdown menu if it isn't already selected.



Calculating Residuals

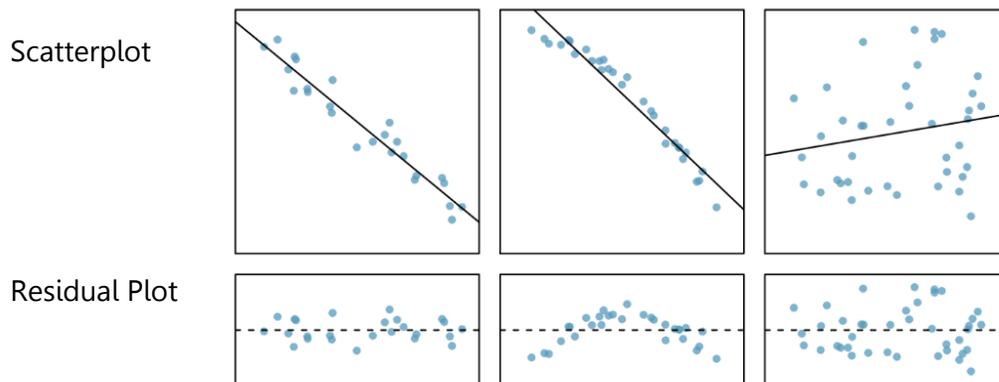
$$\text{Residual} = \text{observed value} - \text{predicted value}$$

o. Calculate the residual for the United States in Example 1. Does the model underestimate or overestimate the use of other drugs for the US?

Check the residual plot for:

- Bends or other patterns in the residuals
- Outliers that weren't clear before
- Any change in spread of the residuals from one part of the plot to another

Which of these residual plots are ok?



Source: [OpenIntro Advanced High School Statistics, page 382](#)

p. Use the residual plot for Example 1 to explain whether the linear model is appropriate for the data.

Making Predictions

q. If a country had 20% of its teenagers using marijuana, how many would use other drugs according to the model? Do you think this is a reasonable prediction?

r. If a country had 75% of its teenagers using marijuana, how many would use other drugs according to the model? Do you think this is a reasonable prediction?

s. If a country had 105% of its teenagers using marijuana, how many would use other drugs according to the model? Do you think this is a reasonable prediction?

Interpolation vs. Extrapolation

Interpolation is making a prediction within the x-values of your data set.

Extrapolation is making a prediction beyond the data set – be careful!! How do you know that the trend will continue?

t. In questions q, r, and s above, which are interpolation and which are extrapolation?

Practice 1. The data below shows the cost of the airfare and the distance traveled to each destination from Baltimore, Maryland.

a. Which is the explanatory and which is the response variable?

b. Make a scatterplot using GeoGebra.

c. Write a description of the association, mentioning all four characteristics and include the correlation value.

Destination	Distance	Airfare
Atlanta	576	178
Boston	370	138
Chicago	612	94
Dallas	1216	278
Detroit	409	158
Denver	1502	258
Miami	946	350
New Orleans	998	188
New York	189	98
Orlando	787	179
Pittsburgh	210	138
St. Louis	737	98

d. If the relationship is linear, calculate the regression equation and write it using proper notation.

e. Give the slope and interpret it in the context of the data.

f. Give the y-intercept and interpret it in the context of the data.

g. Give the value of R^2 and its interpretation.

h. Make a residual plot using GeoGebra and use it to explain whether the linear model is appropriate.

i. Calculate the residual for Chicago and explain what it means.

Practice 2. Here is a least squares regression line for the relationship between gas mileage (mpg) and engine size (in liters). The model used data from 35 different models of 2014 vehicles.

$$\widehat{mpg} = 36.25 - 3.867 \cdot Engine\ Size$$

a. If the car you are thinking of buying has a 4-liter engine, what does this model suggest your gas mileage would be?

b. What does a positive residual mean in this context?

c. What is the slope and what does it mean in this context?

d. What is the y-intercept and what does it mean in this context?

e. The correlation for the model is $r = -0.8476$. What does that mean?

f. How much of the variation in fuel economy is accounted for by the engine size?