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.

	X	y
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
	explanatory	response

a. Do you think there might be a relationship between Marijuana use and other drug use?

b. 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

y

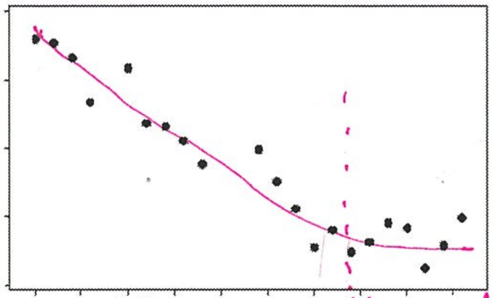
The **response variable** is the dependent variable (y). In our example:

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

X

We think that changes in the explanatory variable might <u>explain</u> changes in the response variable.

A Framework to Describe Association



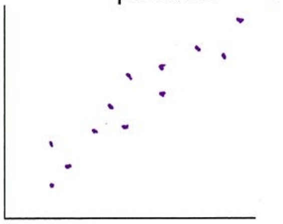
negative, curved, moderate association. There may be 2 different patterns.

Describe four features of the association between two variables:

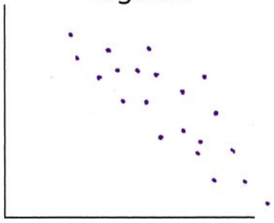
- Direction
- Form
- Strength
- Unusual Features (subgroups or outliers)

Direction:

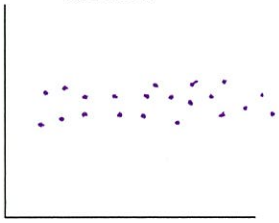
positive

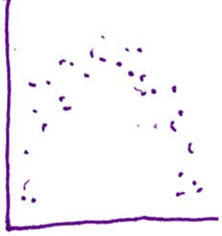


negative



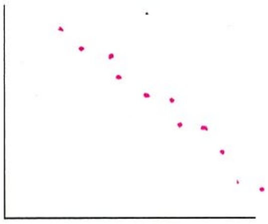
neither



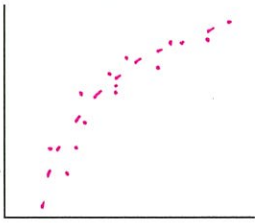


Form:

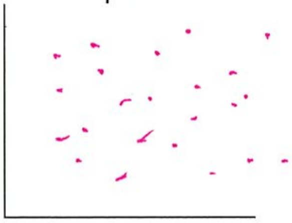
linear



curved

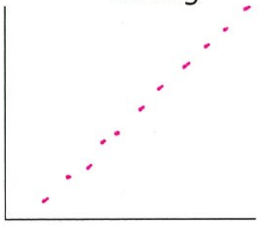


no pattern

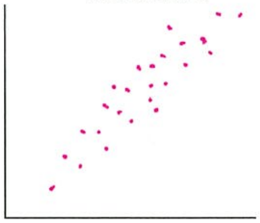


Strength: – measure strength with correlation, r

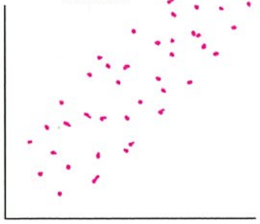
strong



moderate

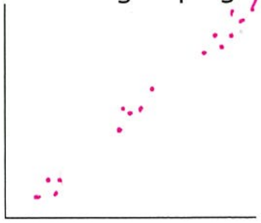


weak

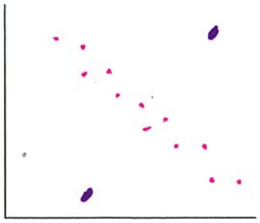


Unusual Features:

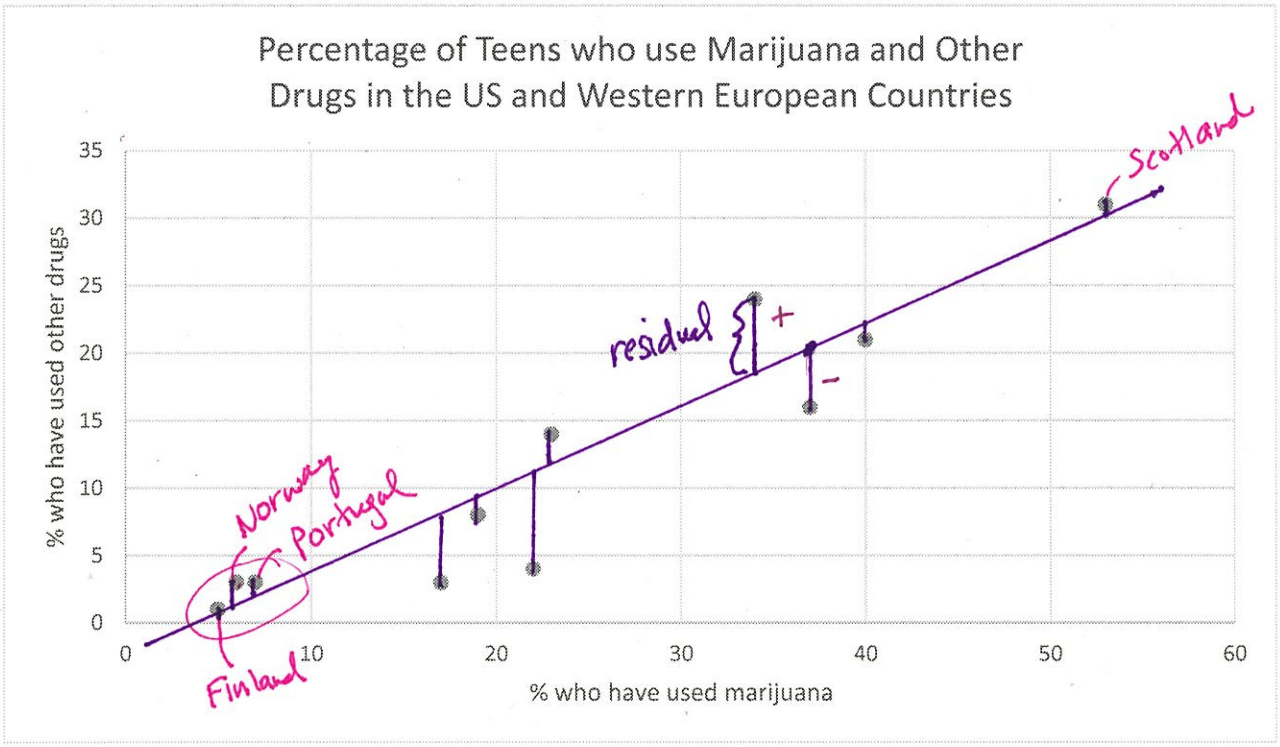
groupings/gaps



outliers



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



c. Write a description of the association including all four characteristics. strong, r = .934!

There is a positive, ~~moderate~~, linear association between the percentage of teens who have used marijuane and other drugs. There is a grouping of three countries with very low percentages (Finland, Norway and Portugal). There also seem to be gaps from 7-17%, 24-34% and 41-53%. Scotland is higher on the percentages, but might fit the pattern.

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.

Residual = $\overset{\text{data}}{\text{observed value}}$ – predicted value

y $\quad$ y-value on the line : $\hat{y}$

$y - \hat{y}$

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.

$\frac{\Delta y}{\Delta x}$ $\frac{\text{change in } y}{\text{change in } x}$

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

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

m represents the slope $\frac{\text{rise}}{\text{run}}$ $\frac{\Delta y}{\Delta x}$ b represents the y-intercept

$\hat{y}$ 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

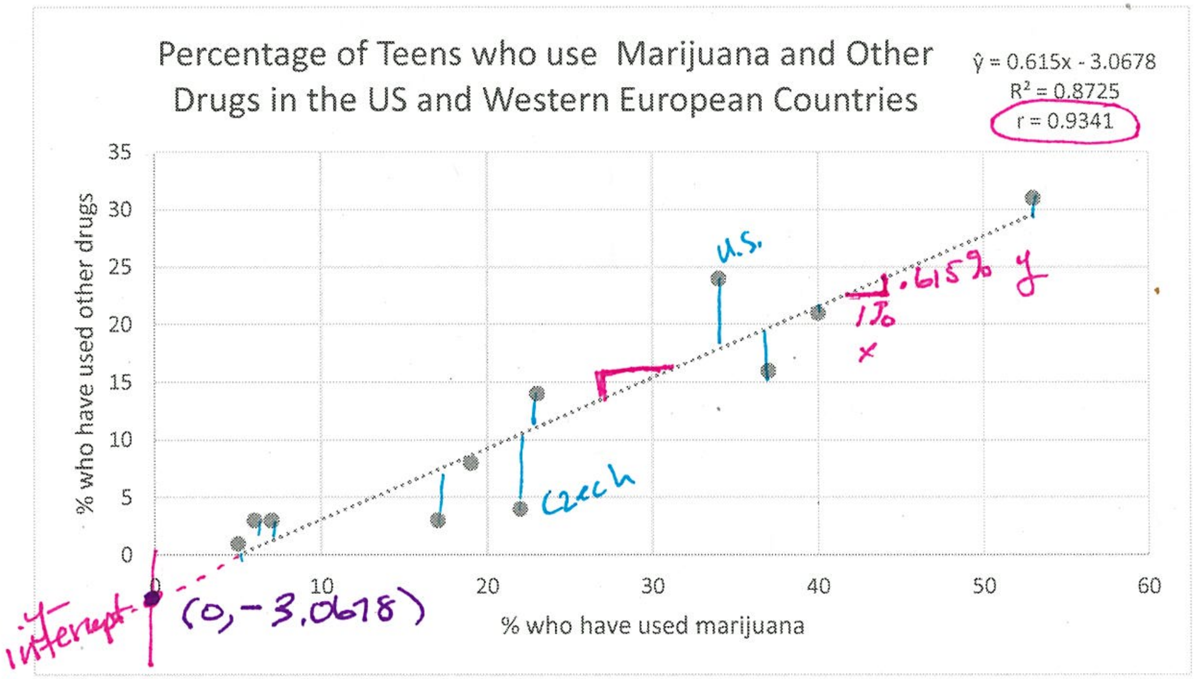
2. Depending on your version of Excel:
   - If you have symbols on the right side of your scatterplot, Click on the + sign for chart elements. Check the box next to Trendline and click on the right arrow and select more options. Check "Display Equation on Chart" and "Display R-squared value on chart."
   - 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 y-hat:

$$\hat{y} = \underbrace{0.615}_{m}x \underbrace{-3.0678}_{\text{y-interept}}$$

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

$\frac{\text{rise}}{\text{run}}$ $m = \frac{0.615 \text{ \% used other drugs}}{1 \text{ \% used marijuana}}$

generic: a 1 unit increase in the x-variable is associated with a \$____ unit increase in the y-variable

A 1% increase in marijuana use is associated with a .615% increase in other drug use according to the model.

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

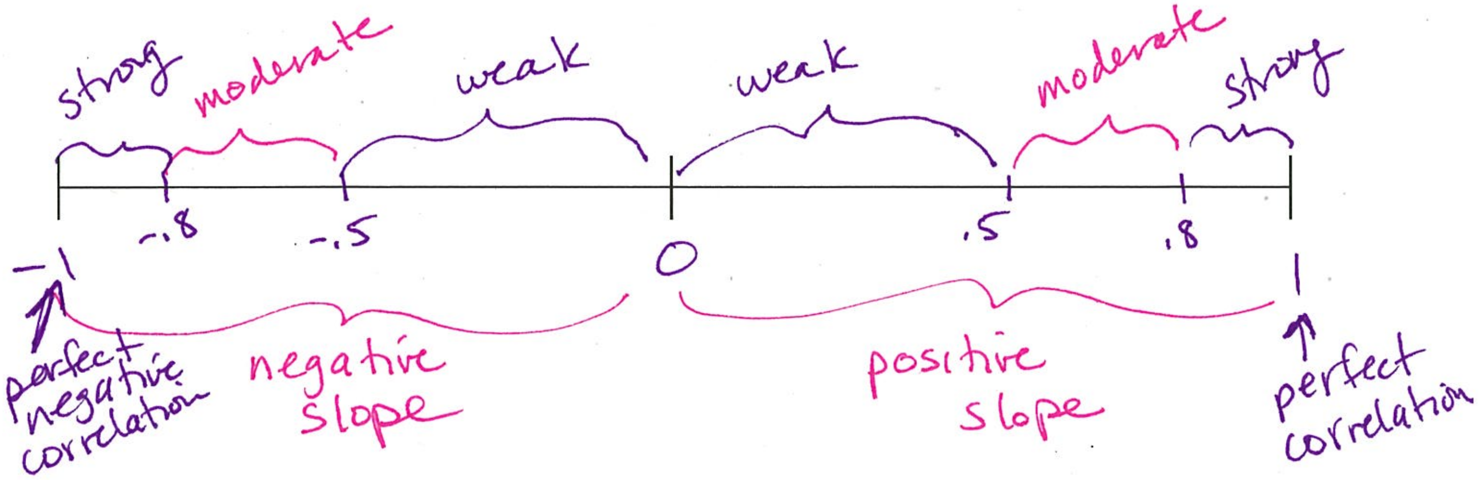
The y-interept is (0, -3.0678). This means when there is 0% marijuana use the other drug use would be -3% according to the model. This point is not meaningful in our context because we can't have a negative percentage.

generic: The starting value or where the line crosses the y-axis. (0, ____)

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.



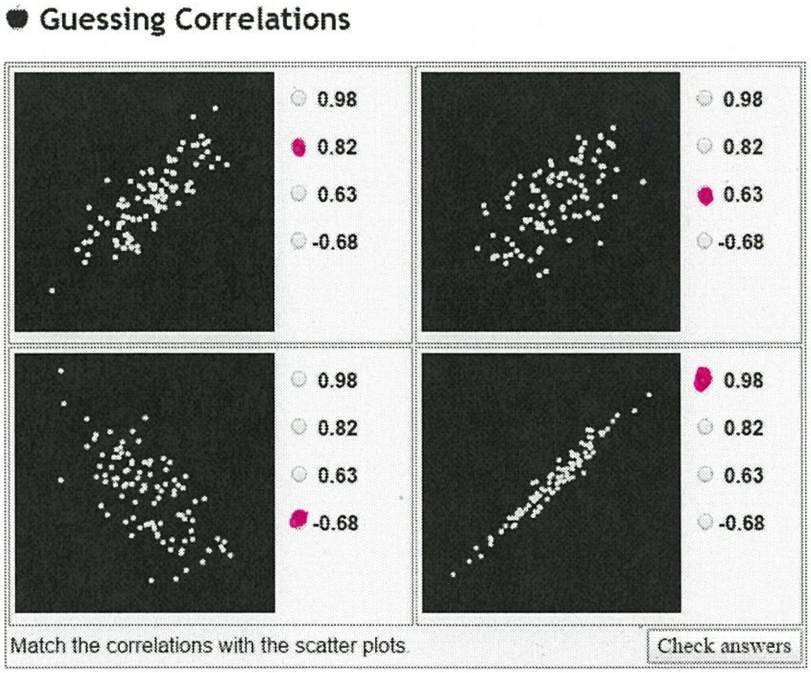
Characteristics of r:

- 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

Top left: ○ 0.98 ● 0.82 ○ 0.63 ○ -0.68

Top right: ○ 0.98 ○ 0.82 ● 0.63 ○ -0.68

Bottom left: ○ 0.98 ○ 0.82 ○ 0.63 ● -0.68

Bottom right: ● 0.98 ○ 0.82 ○ 0.63 ○ -0.68

Match the correlations with the scatter plots.

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?

The correlation between the percentage of teens who use marijuana and other drugs is .9341 so this is a strong association.

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

Even though there is a strong correlation we cannot infer causation. We cannot conclude that marijuana use causes other drug use.

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?

social settings, age, location, peer pressure, laws, grades, education, genetics, stress + trauma, medical issues, household size.

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

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

In GeoGebra, $\boldsymbol{R^2} = 1 - \frac{SSE}{Syy}$

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 ____insert y-value context____

can be explained by the variation in ____insert x-value context____.

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

$(.9341)^2 = .8725$

m. In our example, R^2 means:

Approximately 87% of the variation in other drug use can be explained by the variation in marijuana use. Therefore, 13% of the variation is due to other variables.

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?

Czech Republic : Residual = observed − Predicted

$= y - \hat{y}$

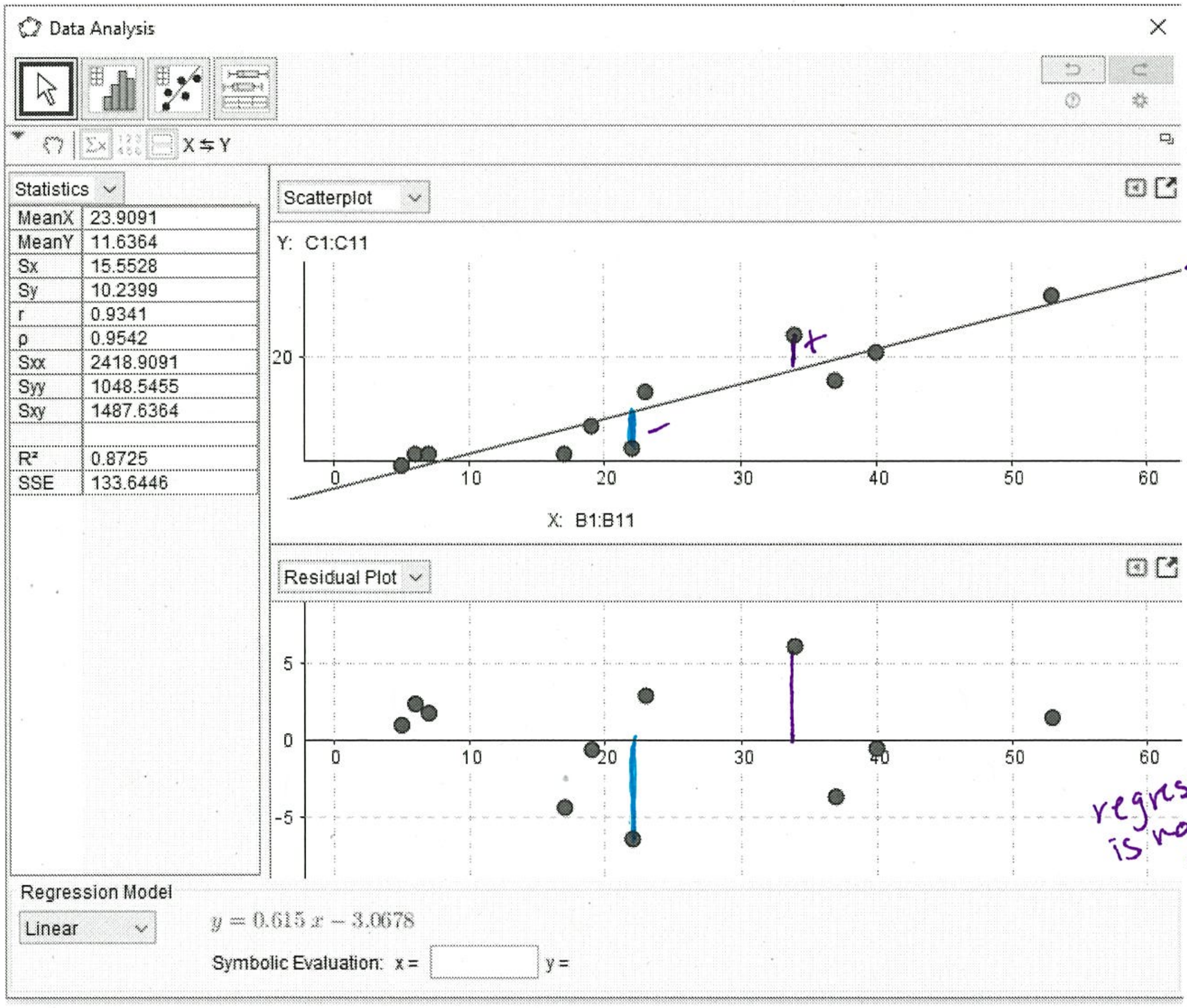
$= 4 - 10.4623$ – from the model

$= -6.4623$

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.



Statistics	
MeanX	23.9091
MeanY	11.6364
Sx	15.5528
Sy	10.2399
r	0.9341
ρ	0.9542
Sxx	2418.9091
Syy	1048.5455
Sxy	1487.6364
R²	0.8725
SSE	133.6446

Regression Model: Linear

$y = 0.615\,x - 3.0678$

Symbolic Evaluation: x = y =

regression line is now the x-axis

Calculating Residuals

Residual = observed value – 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?

US: observed – predicted

$y - \hat{y}$

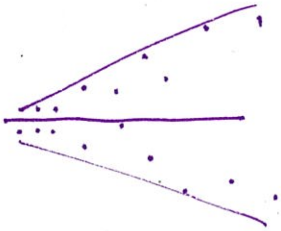
$= 24 - 17.8423$

$= 6.1577$

The model underestimates the other drug use for the U.S.

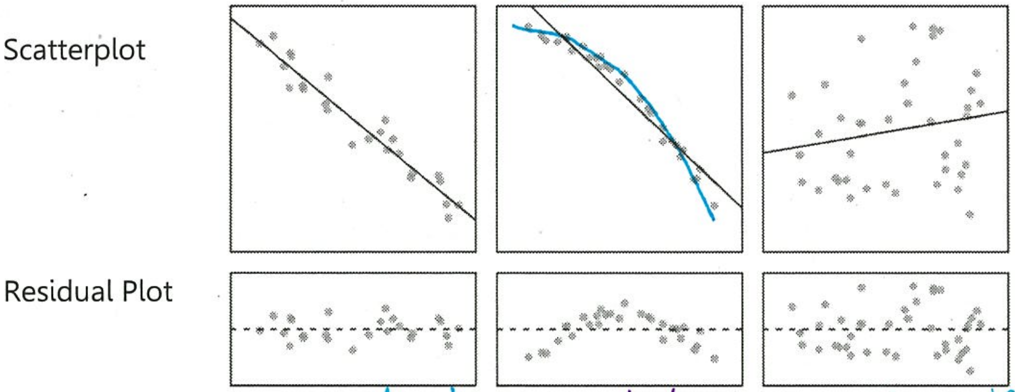
Check the residual plot for:

- 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?

curved



OK - randomly scattered

not ok - there is a pattern

OK - evenly scattered

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.

The linear model is appropriate because the residuals are evenly scattered in the residual plot. (There is the grouping of 3 countries that we noted earlier)

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?

$\hat{y} = 9.2323$

$\hat{y} = .615x - 3.0678$

$= .615(20) - 3.0678$

The predicted value for other drug use would be 9.2% (interpolation)

This prediction is reasonable because 20% is within our x-values.

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?

$\hat{y} = 43.0574$

This prediction is an extrapolation.
we don't want to predict outside of our x-values.

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?

a percentage of 105% is not possible
This is a silly example - be careful what you plug into a model.

$\hat{y} = 61.5075$ (also extrapolation)

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?

q is interpolation
r and s are extrapolation

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

Destination	Distance X	Airfare y
Atlanta	576 mi	$ 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

a. Which is the explanatory and which is the response variable? Distance is the explanatory variable and the cost of airfare 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.

There is a moderate, positive, linear association. The correlation is .70. Miami may be an outlier.

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

$\hat{y} = .1373x + 81.7636$

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

$m = \frac{\$.1373}{1 \text{ mile}}$

The slope is about $.14/mile which means an increase of 1 mile in the flight distance corresponds to a $.14 increase in the price.

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

The y-intercept is $81.76 which means a 0-mile flight would cost $81.76. This could be interprete as the fixed cost of a flight.

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

R^2 is .48, so about 48% of the variation in airfare price can be explained by the distance of the flight. 52% is due to other factors

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

The residuels look evenly scattered so this linear model is appropriate. Miami looks like an outlier on the residual plot.

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

Residual $= y - \hat{y}$ – plug 612 into the model

$= 94 - 165.766$

$= -71.766$

The model overestimates the airfare to Chicago by $72. or ($71.77)

Practice 2

~~**Example 3.**~~ 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.

$$\underset{y}{\widehat{mpg}} = 36.25 - 3.867\underset{x}{Engine\ Size}$$

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

$$\widehat{mpg} = 36.25 - 3.867(4)$$
$$= 20.782$$

The gas mileage for a 4-liter engine is predicted to be 21 mpg.

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

The predicted value would be above the line so the model underestimates the mpg for that engine size.

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

$\frac{rise}{run}$

The slope is $\frac{-3.867 \text{ mpg}}{1 \text{ Liter}}$.

For an increase of 1 liter in engine size, the gas mileage goes down by 3.9 miles per gallon.

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

The y-intercept is 36.25 miles per gallon.

A 0-Liter engine would get 36mpg – does this make sense?

This could be interpreted as a max mpg.

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

There is a strong negative correlation between engine size and gas mileage.

f. What fraction of the variability in fuel economy is accounted for by the engine size?

$R^2 = (-.8476)^2 = .718$

About 72% of the variation in fuel economy can be accounted for by the variation in engine size.