Math Q114 **Name____________________________________**

Regressions Lines **Complete and turn in Monday 3/29**

In section 2.8, we informally sketched linear functions to approximate a linear trend in data. To do this formally, we use what's called a *regression line*.

You can think of a regression line as the line that comes as close as possible to all of the data values. (In fact, it's calculated by adding up the squares of the vertical distances between the line and the data points, and making that sum as small as possible. But we will have Excel calculate it for us.)

Go to the Explorations folder and open the software called "Linear Regression". Go to *R7: Regression and Correlation Examples*. You'll see 6 scatter plots. Try to guess what the regression lines will look like. Click on *Linear Regression Lines* when you're ready to see the regression lines. You can click on *New Data Points* if you want to try another set of examples.

Correlation coefficient or cc is a numerical measure of the reliability or “fit” of the regression line to the actual data. If the data points of a scatterplot fit perfectly on a line then the $|cc| = 1$. If the data points are randomly distributed and do not fit a linear pattern then $|cc| = 0$. The correlation coefficient can have values between –1 and +1 or we say that $-1 \leq CC \leq 1$. When a regression line has a <u>negative slope</u>, then it’s <u>cc is also negative</u>, when the line has a <u>positive slope</u>, then it’s <u>cc is positive</u>.

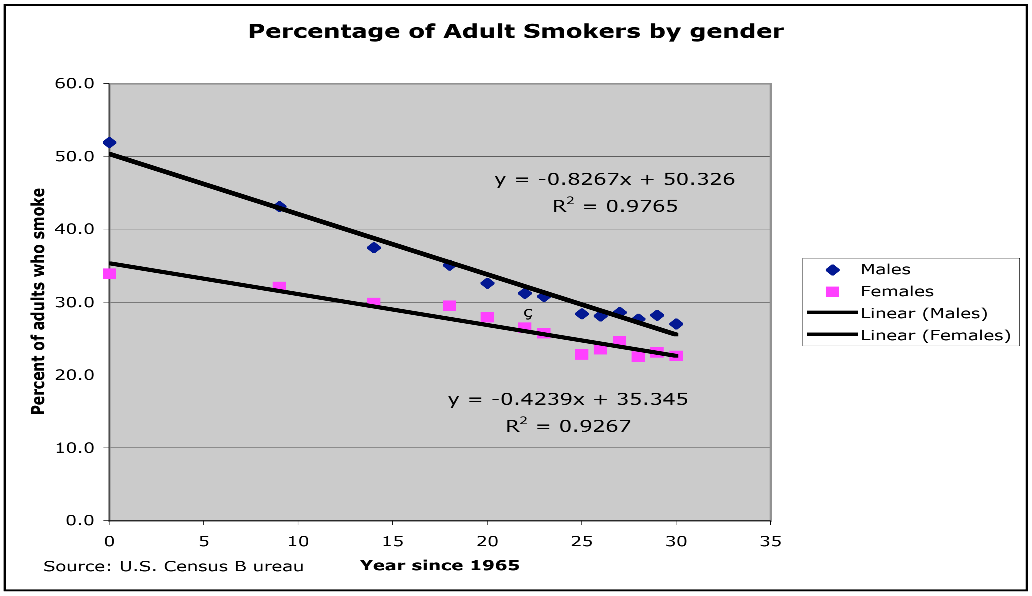
R^2 value and correlation coefficient.

Excel is a powerful toll that can calculate the precise regression line for a set of data, but does not give the cc for the line instead it give the R^2 value. The two are related, since the square root of the *R^2 value = correlation coefficient*. However be careful in choosing the correct sign for cc, since when you take the square root of a number the answer can be either + or -. If a line has a positive slop then the cc is positive, if the slope is negative so is the cc; in other words: $cc = +/- \sqrt{R^2}$

Positive slope of line then $cc = + \sqrt{R^2}$

Negative slope of line then $cc = - \sqrt{R^2}$

Exercise 1: The graph below shows the percentage of adults who smoke from 1965 to 1995 by gender. The linear equations and R-squared values are given for the regression lines for male and female smokers.



Percentage of Adult Smokers by gender

Source: U.S. Census B ureau

a. Write below the average rate of change for adult male smokers (including units) and explain what it means in a complete sentence.

b. Write below the average rate of change for adult female smokers (including units) and explain what it means in a complete sentence.

c. What is the main observation you make in comparing adult male to adult female smokers. IN a brief paragraph describe the similarities and differences in the trend for each. Cite the average rate of change for each regression line to back up your observations.

d. The equations for adult smokers by gender rounded to one decimal place are:

Male: $y = -0.8x + 50.3$ Female: $y = -0.4x +35.3$

Use these two equations to estimate the year when a higher percentage of adult women will be exceed the percentage of men smokers. (Show your calculations below).

Exercise 2: We will now put to practice what we've learned about regression lines, R^2 value and correlation coefficient.

Open the Excel data file **LONGJUMP.XLS** in the course folder on your desktop.

a. Create a scatter plot of women's long jump records from 1954 to 1988. But be sure before making the plot to change the year column to one that measures years from 1955 (1954 = year 0). Be sure also to include your name on the graph and a suitable title and labels for both of the graph's axes.

b. Use Excel to create a best fit lines through the scatter plot. Have Excel calculate the equation of this line and include the R^2 value. Record the equation and R^2 value below.

c. Record below the average rate of change and vertical intercept for women's long jump records. INCLUDE UNITS FOR EACH.

d. In complete sentences explain what the average rate of change tells you about women's long jump records between 1954 and 1988.

e. Predict what the women's long jump record will be in the year in 2010 using the equation of the best fit line.

f. What is the correlation coefficient for this best fit line? ___________. Explain what this tells you about the reliability of your prediction in question e above?

Format your Excel spreadsheetso that the data and graph appear on a single page. Print and attach to this paper.

(Continued on next page)

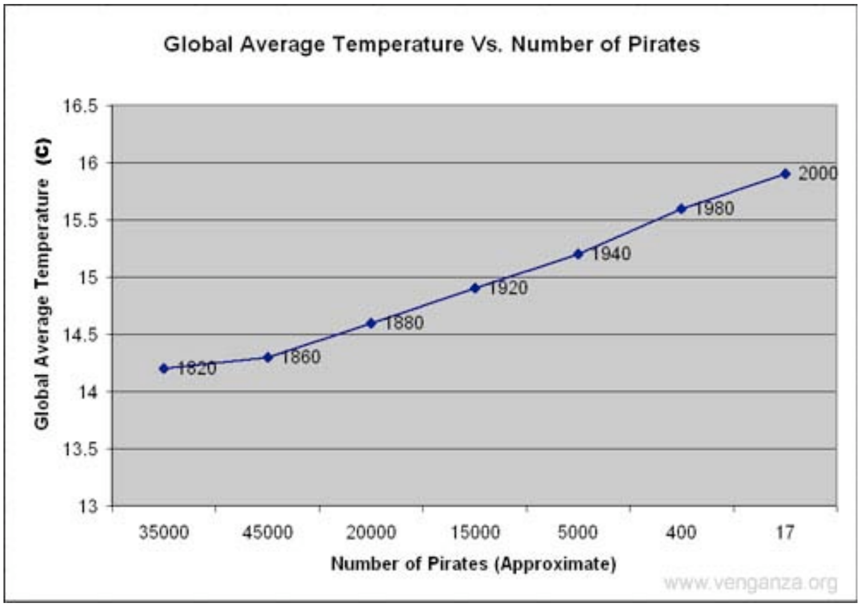
To interpret the correlation coefficient
Be careful with correlation. It describes the relationship between the variables, but it does not give the entire story. It's quite possible that there are other variables that are influencing the situation, and you should certainly try to think about those before drawing any conclusions.

Another important point to keep in mind is that correlation is not the same as causation. This is so important, so it's worth repeating:

CORRELATION IS NOT THE SAME AS CAUSATION!

What this means is that there is only an *association* between the two variables – we cannot say that a change in one variable *causes* a change in the other. Here is an example to think about.

Exercise 3. The following graph shows the number of pirates over the past 180 years along with the average global temperature (in degrees Celsius).



The author of this chart noted that, "global warming, earthquakes, hurricanes, and other natural disasters are a direct result of the shrinking numbers of pirates since the 1800s." Is it the case that because there are fewer pirates, the average temperature has increased? Explain why this graph shows correlation but not causation. (For the original graph, see www.venganza.org)