Chapter 4 - Video Script and Chapter 4 Popper

Additionally, let's get a bit more formal with a reminder:

Homework style sheet and rules:

Work on one side only; pdf it and upload it before the deadline on the calendar.

Work that is poorly scanned or illegible will be given a zero.
This includes sideways or upside down scans!

Do NOT crowd the work, leave at least 3" between problems.
Label the answers carefully so the grader can grade efficiently.

Chapter 4 – Data with Two Variables

Pages 85 – 117

We dealt with "univariate" data in Chapter 2…these are data sets with only one variable. Often, there are 2 variables to deal with – we've seen this in College Algebra. When there are 2 variables we call the data set "bivariate". This happens because the two variables are often related to one another and if they are related we call them "correlated". Think of time and technology prices…as time increases, the price for a particular laptop typically decreases. This is not easy to see if you study time separately from price! no causation necessary

Initially we will focus on linear relationships between two variables.

4.1 Scatter Plots and Correlation

With a scatter plot, we graph the data points just like we're in College Algebra on the Cartesian plane. Now two differences show up right away! Sometimes in Statistics, there is no "origin"…the intersection of the axes is at some other point values than (0,0) – check the numbers at the SW intersection (always) so you don't get caught by this fact. And the fact that real data is almost never perfect is the second difference – the points will be "off" the line that we use as a model. This is due to error, variation in experiments, the very "real" nature of what we're measuring.

Correlation is a number that tells how "perfect" your real data is. The formula is complex and we'll just use technology to come up with it.

As we are only working with linear data in the book, we'll use terminology from lines.

m is slope
b is the y-intercept

$y = mx + b$

We will also, and not from the book, look at some NON LINEAR data.

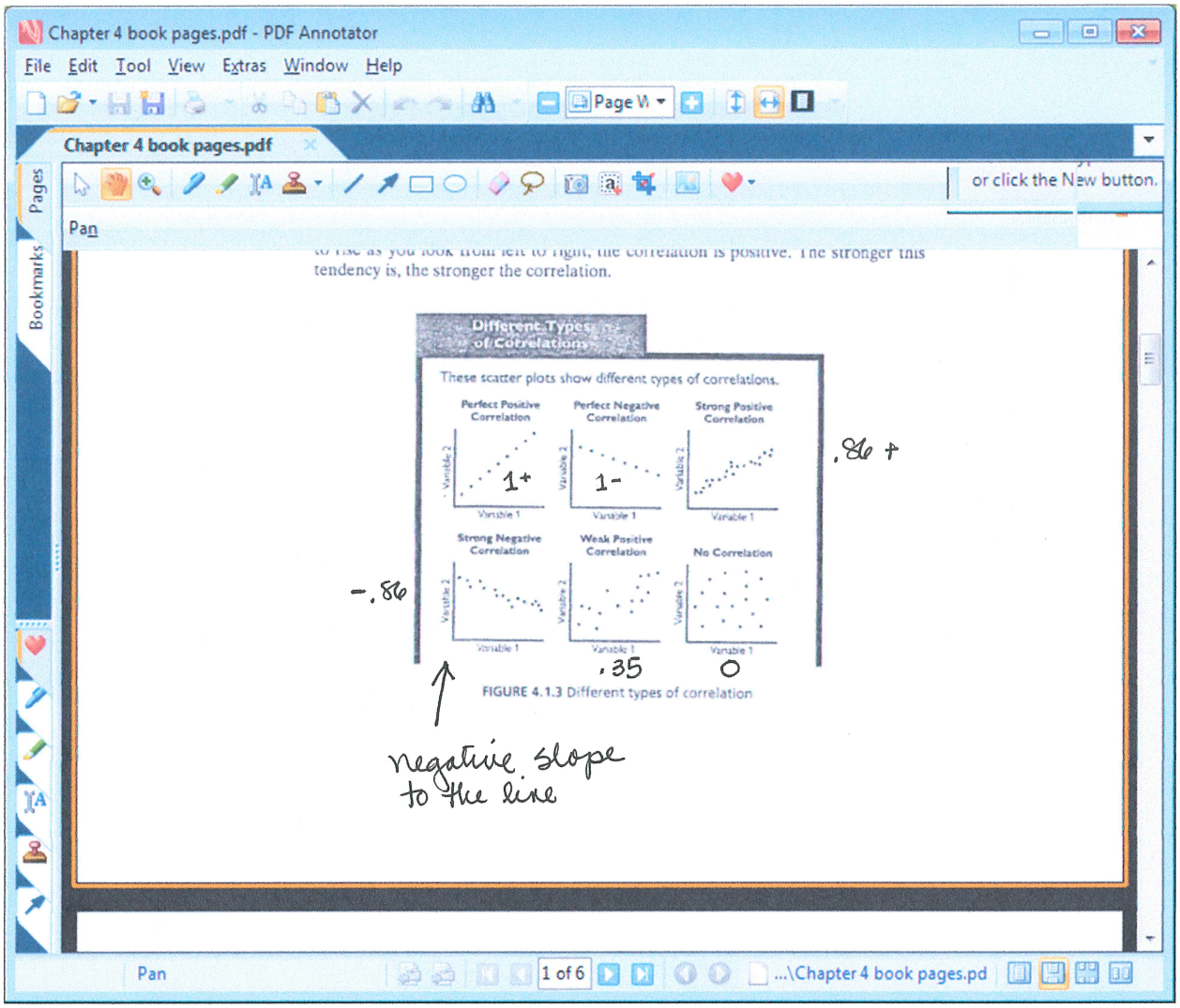
We can have a perfect positive correlation $R^2 = 1$

We can have a strong positive correlation $R^2 = .75ish$

We can have a weak positive correlation $R^2 = .35ish$

We can have NO correlation $R^2 = 0$

We can have each of these with NEGATIVE correlations, too: perfect, strong, and weak . Let's look at pictures of these, too. We use "R" w/o the square to show ± w/. linear regression



Positive, Negative, Strong, Weak…these are the descriptors to use

Think of the numbers as grades! 95% is an excellent grade and an excellent correlation, too. A small caution: you want $n = 30$ or bigger.

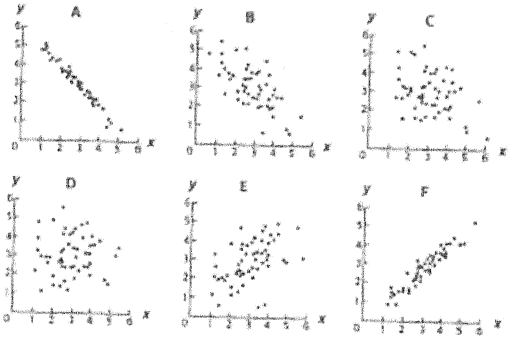
Now, one big caution:

There is a strong correlation between time and a laptop price…but time passing doesn't CAUSE the price decrease…people just stop paying so much because it's old technology now and they're looking forward to the next big thing. A good correlation coefficient doesn't imply that one thing is causing the other…just that there is a relationship.

Let's look at the diagrams on page 88 at the bottom and assign correlation coefficients to them along with describing them.

Scatter plots are often made to show the relationship between two variables. The points are a scatter plot, or scatter diagram, look like a "cloud" of points. When there is a strong relationship between the two variables, the graph resembles a line. If the points appear as a straight line, you can say there is a strong *correlation* between the two variables.

8. Describe the correlation in the following scatter plots. Indicate whether there is no correlation, a weak correlation, or a strong correlation.



9. For which of the above scatter plots can you best predict the value of y when $x = 4$? Explain your reasoning.

In the dingrams above, you can see that a correlation can be weak or strong. Correlations can also be positive or negative.

10. a. What do you think is meant by the phrase "negative correlation"?
b. Which of the above scatter plots show a negative correlation?

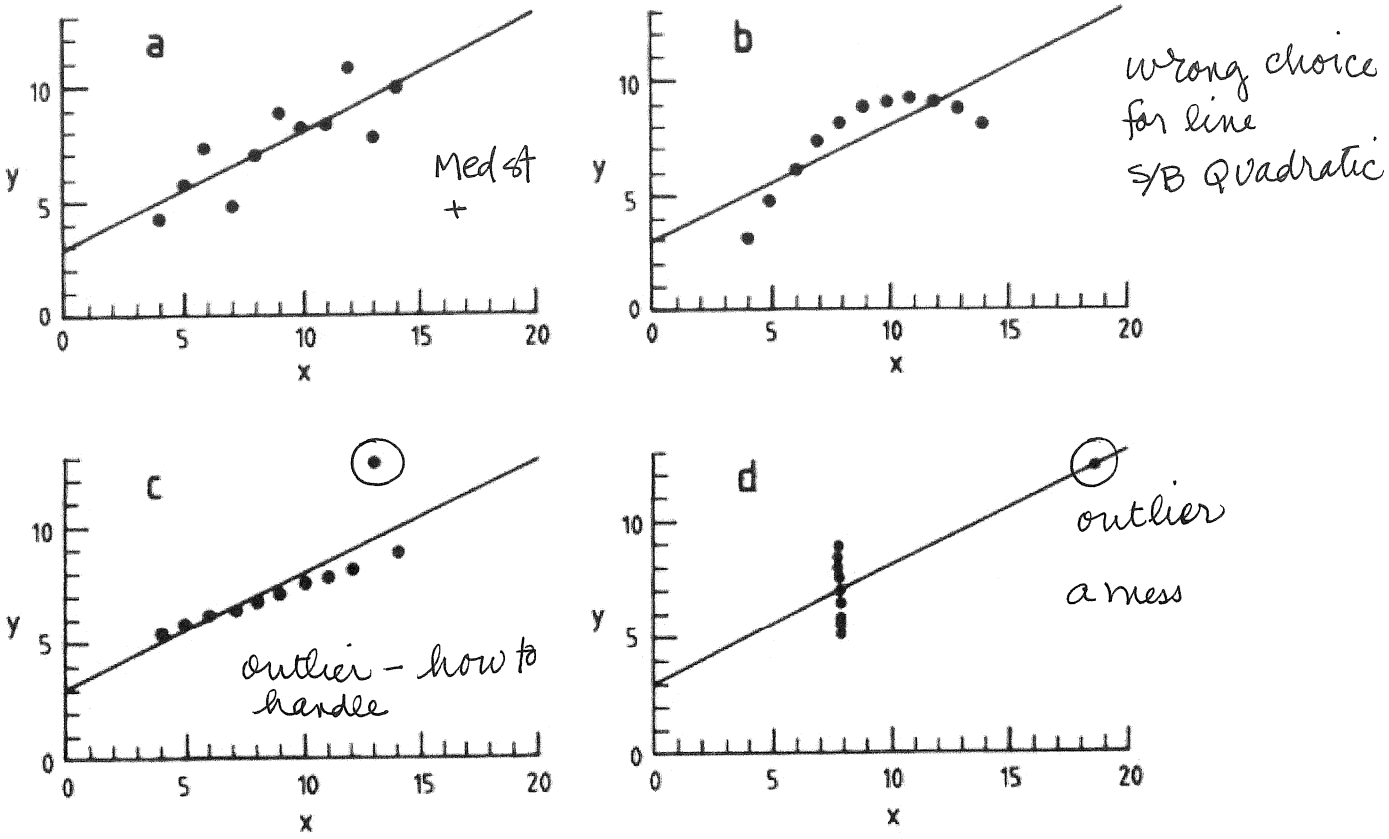
FIGURE 4.1.4 Describing correlation

A - 90 ish strong
B - 70 ish med
C - 30 ish - ? weak
D - 15 ish + ? weak
E + 65 ish med weak
F + 85 ish strongish

Positive, Negative, Strong, Weak

Also from the internet – more "regression lines"

Technology will do what you tell to



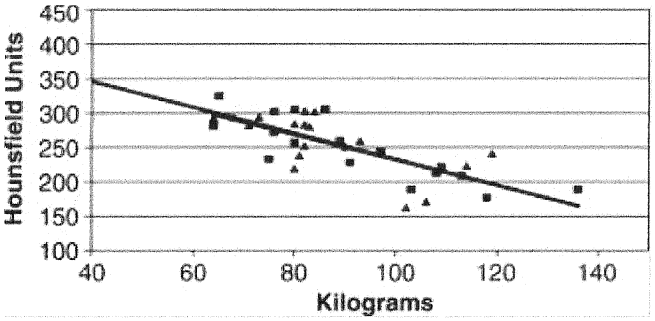
Let's discuss these, too. You want these to be "pattern free"

Chapter 4 Popper Question 1

Given this graph and regression line

Plus:

I is positive
II is negative
III is strong
IV is weak

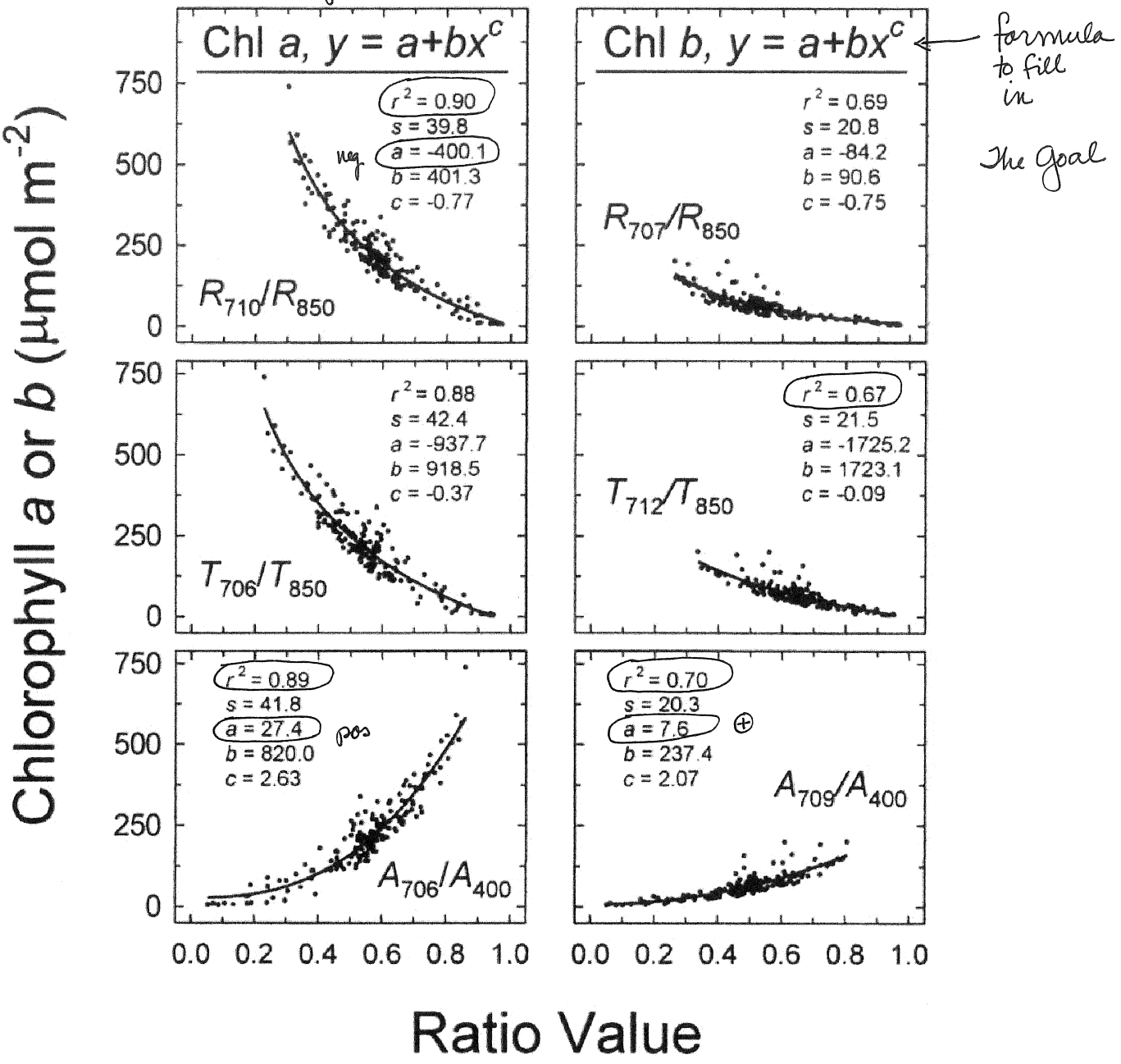


Which of the following describe this situation?

A. I and III
B. I and IV
C. II and III
D. II and IV

To show that not all regression curves are lines! Note the exponential formula !

$y = -400.1 + 401.3x^{-.077}$



formula to fill in

The goal

Reviewing the "cause and effect" notion, OYO at page 89 and check out the graph of manatee deaths and the stock market.

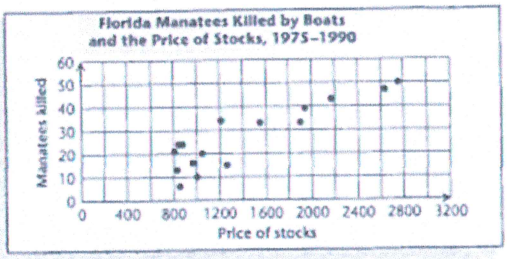
From an internet listing we find:

What is the difference between correlation and cause and effect?

Correlation means that two or more sets of data move in some consistent pattern. Perhaps during a 10-year period the number of cars sold in the U.S. moved in the same direction as the country's rate of inflation. Even with a 10-year correlation between the two sets of data, it is unlikely that more inflation caused an increase in the number of cars sold. In other words, *correlation does not assure that there is a cause and effect relationship.*
On the other hand, if there is a cause and effect relationship, there will have to be correlation.

Here's the page 89 chart.

13 The scatter plot below shows the number of Florida manatees killed by boats and the price of stocks as measured by the Dow Jones Industrial Average for the years 1975–1990.



a. Is there a *positive correlation*, a *negative correlation*, or *no correlation* between the price of stocks and the number of manatees killed by boats?

b. Would it be correct to say that a rise in stock prices tends to *cause* an increase in the number of manatees killed by boats?

c. **Discussion** If a correlation exists between two variables, does that necessarily mean there is a cause-and-effect relationship between the variables? Explain.

FIGURE 4.1.5 Stock prices and manatees killed

medium ⊕
stocks don't kill manatees; people kill manatees

Chapter 4 Popper Question 2

There is a strong causal relationship between events and things that have a strong correlation coefficient.

A. True
B. False

4.2 Pearson's Correlation Coefficient

Now let's look at the formula for the correlation coefficient:

$$r = \frac{S_{xy}}{\sqrt{S_{xx} \cdot S_{yy}}}$$ **this is the square root of Rsquared used above**

Let's take this apart:

$$S_{xx} = \sum x^2 - \frac{\left(\sum x\right)^2}{n} = \sum \left(x - \bar{x}\right)^2 \quad +$$

$$S_{yy} = \sum y^2 - \frac{\left(\sum y\right)^2}{n} = \sum \left(y - \bar{y}\right)^2 \quad +$$

$$S_{xy} = \sum xy - \frac{\left(\sum x\right)\left(\sum y\right)}{n} = \sum (x - \bar{x})(y - \bar{y})^2 \quad \pm$$

These are relatively similar!

Now, depending on WHERE the points are the product in S_{xy} is positive (Q1 or Q3) or negative (Q2 or Q4). And note that this product is the AREA between the data point and the two mean lines…even though it is sometimes negative. See the picture on page 91!

We will use technology to get our Correlations coefficents.

Let's check out the **properties** on page 93 for Pearson's Correlation Coefficient R and let's remember that we will, indeed, use technology to get that number!

1. It's always:

$$-1 \le r \le 1$$

the endpoints are perfect; the center: 0 is no correlation
The size of the sample matters…the more data points the better. Try to be over 30.

Remember: plus or minus 35 is weakish and plus or minus 75 is pretty good, And plus or minus 95 is excellent

2. Changing the units has no effect on it. Feet to inches; ounces to pounds; Dollars to renminbi…no problem. It is what it is.

3. The value is not affected by which variable is called x and which is called y. This is because there is NO cause and effect relationship involved.

***Pull out the TIs page 95 has the instructions for this!

Chapter 4 Popper Question 3

If we have a correlation coefficient of zero we are nicely in the middle and it's a good reading.

A. True
B. False

BREAK HERE FOR VIDEO 2, CHAPTER 4

4.3 Slopes and Equations of Fitted Lines

"Fitting" a line to the data is actually quite scientific these days and easy, too, with all the technology we have available to us.

For our uses, let's review two forms of the equation of a line:

The Point-Slope Equation $y - y_1 = m(x - x_1)$

This equation works very well with data sets....you can get the slope and use a point quite efficiently.

The Slope-Intercept Equation $y = mx + b$

Once we've used the preceding equation, we'll convert to this form for ease in reading!

Let's look at the data on page 97 – and at the lines for the two girls' work. NOTE that the data starts at (700, 50) and not the origin. See the squiggle on the graph that helps you see this?

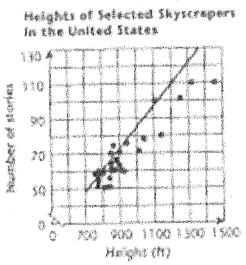
Now, which girl is right?

next page ↓

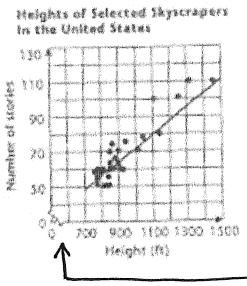
What about the difference in the slopes?

12. Choose the letter of the scatter plot that you think shows the better fitted line. Explain your choice.

A. Lorna's scatter plot



B. Andrea's scatter plot



too many above at end

too many below at end

squiggle! way out in Q1!

FIGURE 4.3.1 Comparing fitted lines

Actually neither is very good. The regression line should go through the center of the data from start to finish.

Now to uses: we will want to predict the number of stories from the height…whether it is a higher skyscraper (extrapolation – going beyond the data range above or below) or one in between the shortest and the tallest for which there is no data point (this is called interpolation – finding a point between the extremes).

What if we extrapolate back to zero? Or up to 1800? use the regression line formula

What about a building that is 1200 feet? We are interpolating here.

put 1200 in the formula for that line & calculate f(1200)… it'll be close-ish

Chapter 4 Popper Question 4

Extrapolation goes beyond or below the data we have.

A. True

B. False

Homework Question 1 – extrap and interp

Using the following data set:

X	Y
10	25
20	45
30	65
40	85

Extrapolation: What is the Y value for and X value of 0 and of 50.

Interpolation: What is the Y value for an X value of 25.

You will have to play with the values and see if you can come up with the formula. Hint: Look at the differences between the Y values. Subtract one from another and then check out the differences in the X values. Then graph it.

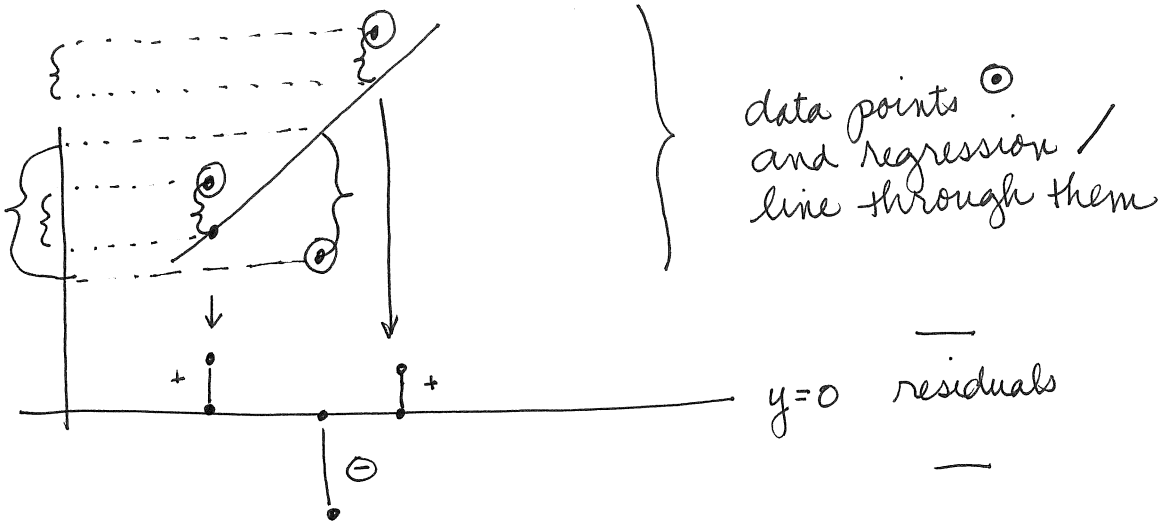
or use technology.

4.4 The Least Squares Line

First the **Least Squares Line Method**.

Let's define the perpendicular distance from a data point to the regression line as the "residual".

Illustration:

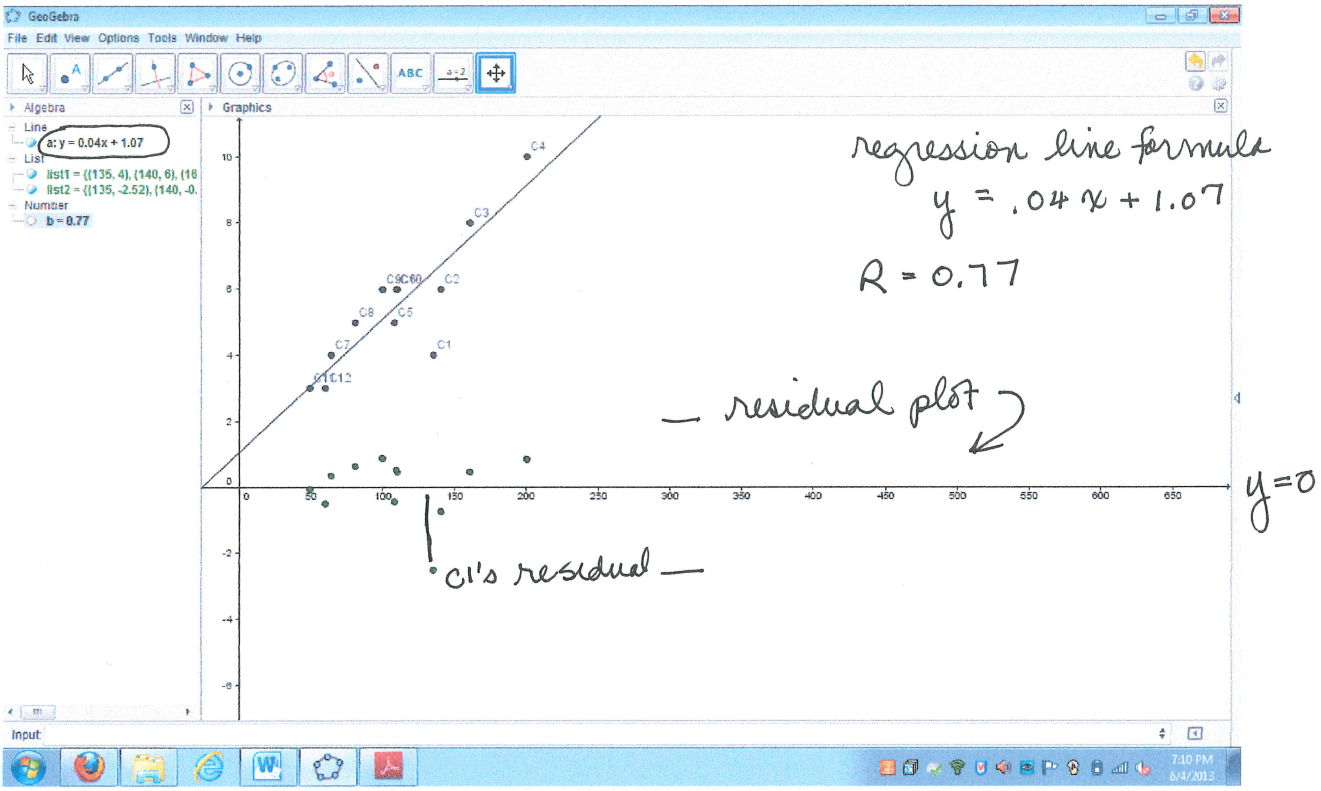


And our goal is to make the sum of the residuals as small as possible. Residuals can be positive (point above the line) or negative (point below the line).

Residuals Plot on a horizontal line: above ↑

Let's look at some residuals:

This is the "Tent and Sleeper" data on page 101.
I used Geogebra to create these visuals.



Note the regression line, the R^2, and the residual plot.

Let's look at the data points and the associated residuals plot. We want a pattern free collection of residual points, half above and half below and all pretty close to the horizontal plot line. This one is not bad; our R^2 is .77 here.

I recommend reading the book very carefully on the following pages and using your TI to do homework and other problems

OYO check out the analysis in the textbook on pages 102 and 103.

Note the area boxes on page 103. Note the discussion about the coefficients on page 103 and 104!

Let's look at our formulas for a linear regression line:

$$m = \frac{S_{xy}}{S_{xx}}$$ ← ± ; ← always ⊕

p.9 in script

$$b = \bar{y} - m\bar{x}$$

The instructions are on page 105 to do this AUTOMATICALLY! Let's do it the automatic way for all homework.

Chapter 4 Popper Question 4

If the residuals horizontal line plot shows about an equal number of residuals above and below the plot line and they are all pretty close, then the regression line is a pretty good one.

A. True
B. False

Homework Problem 2

Varnish Drying Time -

Amt	Hours to dry
1	7.2
2	6.7
3	4.7
4	3.7
5	4.7
6	4.2
7	5.2
8	5.7

Column one is the amount of varnish in grams; the second column is the number of hours it takes to dry. GRAPH this on graph paper. Use your TI to check out the Regression Coefficients for a linear best fit and then a Quadratic best fit. Which one is better and why?

Graph the data.

Find the best fit using technology. Be sure to support your conclusion by getting the Regression Coefficient R^2

Extension: How long will it take 10 grams of varnish to dry? Use the best formula to come up with this extrapolation.

Homework Problem 3

Let's look at a made up set of data and discuss it.

x	1	2	3	4	10	10
y	1	3	3	5	1	11

Make the scatter plot and find the regression line. Calculate the correlation coefficient.

What has happened to the data here? Is this a good experiment? Why or why not? What's happening with the two values for 10?

Would you consider any of the points to be anomalies or outliers? One of them is. Chose to eliminate the outlier, the anomaly, and redo the calculations for the Correlation Coeffient. What has happened?

Note that when you eliminate a data point as being an outlier it is very important to make a note of this in your report so the reader knows that you did this and which point it was. This leaves a better trail for people following your experiments in the future.

In addition to lines, you can curve fit to lots of natural behaviors. It's all called line fitting but the line just might be CURVED

Line fitting

Check down the menu on your TI and note all the different families of functions you can fit too….let's discuss some scenarios that are NOT linear and for which you'd need

Quadratic	a tossed ball
Exponential	interest in a CD
Logarithmic	earthquake data

4.5 We will skip this.

Wrapping up:

in the script

We have a 4 question popper and 3 homework problems that will need a TI or an Excel spreadsheet or some technology plus three problems from the book.

Chapter 4 Homework from the book 2, 4, and 6

SEE YOU in the Chapter 5 Video!