Chapter 9: Topics Relating to Straight Lines

This week we consider some important ideas that are relevant to straight lines and to linear growth. We define linear growth as things that increase at a *constant or steady* rate. Let's first look at some definitions.

Interpolation (inter = "between")

If you know values at two points and want to find an *intermediate* value, you use interpolation.

Example 1: Look at the following data for the U.S. Population in certain years. From this information, **estimate** the U.S. Population in 1904.

Year	**U.S. Pop. (Millions)**
1900	76.0
1910	92.0

If we want an estimate of the population in 1904, we are interpolating, i.e. finding a value *between* two values.

Between 1900 and 1910 the population increased by 16 million. So in these 10 years that's an average of 16÷10 or 1.6 million per year. We assume a *steady growth* per year. This is linear growth.

So, since there are 4 years from 1900, we EXPECT an increase of 4 times 1.6 million added on to the population in 1900:

1900 population + 4 years growth = Estimated Population in 1904

76.0 + 4 x 1.6 =
76.0 + 6.4 = 82.4 million

The actual population was 81.8 million! So we are fairly confident of our prediction.

NOTICE: The 1.6 is a *slope*. Remember, slope $= \dfrac{\text{the change in y}}{\text{the change in x}}$

$= \dfrac{\text{the change in the population}}{\text{the change in years}}$

Extrapolation (extra = "outside")

If you want to calculate a value that is *outside* all known data points, use extrapolation. Here's an example of extrapolation.

Example 2. Let's estimate the U.S. population in 1976 from the data in the table below assuming a steady growth once again.

Year	**U.S. Pop. (Millions)**
1950	151.3
1960	179.3

Between 1950 and 1960 the population grew by 2.8 million per year.
If it continued to grow at this rate, the estimated population in 1976, 16 more years =

179.3 + 16x2.8 = 224.1 million.

However in 1976, the U.S. Census gives a population of 214.3 million. We are off by about 10 million folks. The rate of growth turned down quite sharply. Here we are ***extrapolating***, or *estimating beyond the known data values*. In general, extrapolation should be used with caution. We can usually be more confident of interpolation or of estimates close to known data.

Example 3. Let's extrapolate from 1960 and see how close we come to the population in the year 2000. Assume we do not know the actual population.

Again between 1950 and 1960 the population grew by 2.8 million per year.
If it continued to grow at this rate, the estimated population in 2000, 40 more years =

179.3 + 40x2.8 = 291.3 million.

This is off by about 7 million. (The population given in the 2000 census was 284.1 million.*)

This just reinforces our belief that extrapolation is not quite as reliable as interpolation. **However it may be the best estimate we have at the time.**

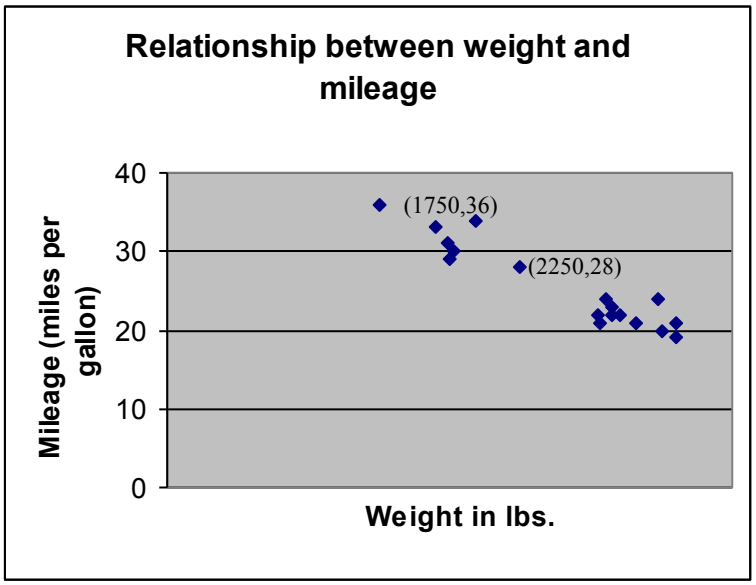
* From www.census.gov/main/www/cen 2000

LINEAR REGRESSION - An Introduction

Regression analysis is used to find an approximate *linear relationship* when the data do not lie exactly on a straight line. We use regression when we have many data values. It is interpolation or extrapolation, but **using all the data values** rather than just two at a time.

Scatterplot

As we learned in the first few lectures, a scatterplot is useful for demonstrating the relationship between *paired* data. By graphing values, a pattern can be detected about the relationship between the two variables. Notice the graph below, which shows the relationship between the weight of a car in pounds and the miles per gallon for each type of automobile.



Weight in lbs	1750	1995	2090	2000	1950	2015	2250	2550	2740	2530	2660	2525	2600	2576	2750	2800	2571	2800
Mileage Miles/gal	36	31	34	29	33	30	28	24	24	21	21	22	22	23	20	19	22	21

Notice, the data value labeled, (1750,36). This means that a car that weighs 1,750 pounds gets 36 miles to a gallon. Another data value is (2250,28), also labeled on the graph.

As you can see, the data is clustered rather closely together. This demonstrates that there is a very close *correlation* between these two variables, weight and mileage. We will discuss *correlation* a bit later.

You can also notice that the heavier cars seem to get fewer miles to a gallon. Thus, we can see a definite relationship here.

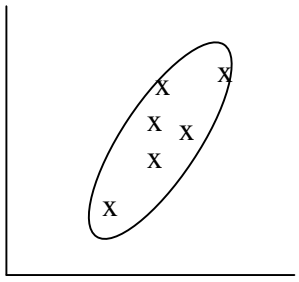
Correlation

This is a *measurement* of the *relationship* between two types of data, e.g. between smoking and lung cancer etc. This does not necessarily imply that one causes the other.

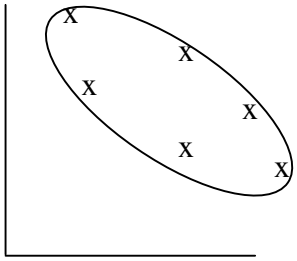
Correlation is a measure of fit between two variables. If the data are close to an imaginary line, then the variables may be closely correlated. However, if the data are more scattered, then the variables may not have much correlation or relationship to each other.

Pearson's 'r'

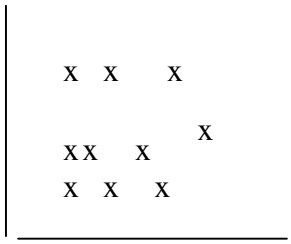
The correlation or strength of a linear relationship between two variables is measured by a numerical value **called Pearson's 'r' value**. 'r' is a number between – 1 and + 1. The sign tells the 'direction' of the correlation. The number **1** tells us that the data are perfectly correlated. As you get closer to '0' the less relationship there is between the variables. Here are some examples:



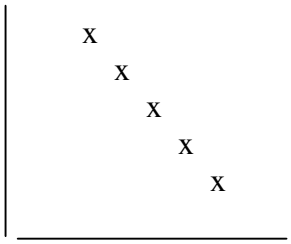
strong positive correlation
r ≅ .6 or .7



moderately weak negative correlation
r ≅ –.3 or – .4



no apparent correlation
r = 0



perfect negative correlation
r = – 1

Causality

This is relevant if one variable is not only *correlated* with the other variable, but can be shown to actually *cause* its behavior. It is very difficult to prove causality. We will not consider it in this class. **Therefore, Correlation** does not necessarily mean **Causality**.

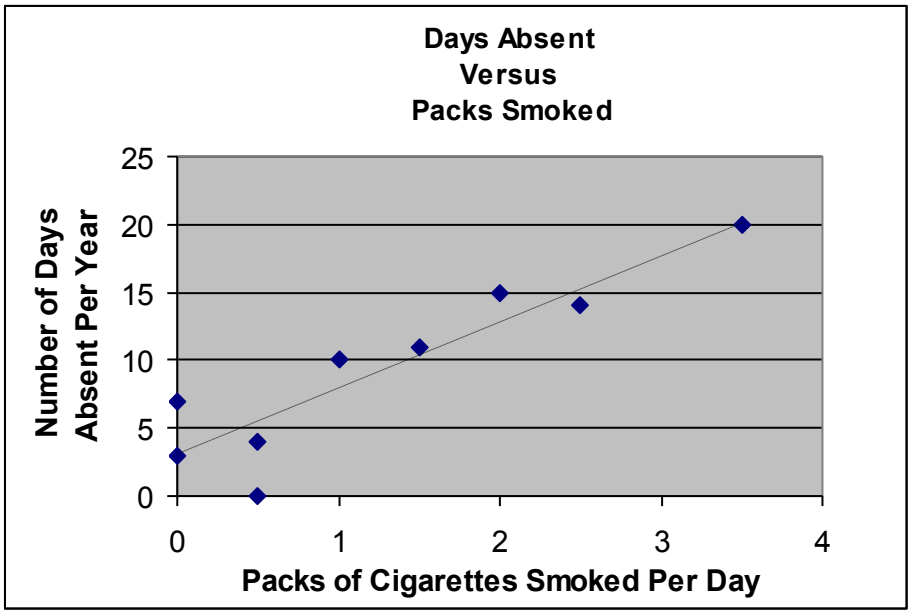
This example and the next one show how to do a regression problem from the beginning to the end. It will do you well to study the process in these examples. The solutions are given below each part.

Example 4: As part of an analysis of the relationship between smoking and absenteeism, the following data were obtained relating an individual's number of packs smoked per day with the same individual's number of days absent per year from his or her job. We would like to determine if there is a correlation between Absenteeism and Packs smoked?

Individual	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**
Number of Packs Smoked per Day	0.5	1.5	2	0.5	0	1	3.5	0	2.5
Number of Days Absent per Year	4	11	15	0	3	10	20	7	14

a) Using a sheet of graph paper, draw a scatterplot using these data. Make sure Packs Smoked per Day is on the horizontal axis. WHY NOT THE OTHER WAY AROUND?

Solution: It's not the other way around – that is Days Absent on the Horizontal- because the number of Packs Smoked per Day is the 'independent variable.' The Days Absent is the 'dependent' variable because the Days Absent 'depends' on how many packs are Smoked.



b) Describe in appropriate statistical language the trend you see AND choose (circle) which of the following numbers is most likely to be the *correlation coefficient*. Briefly explain your choice.

Trend: *"From the graph, it appears that the more packs smoked, the more days absent there are. (Also, the less packs smoked, the less days absent.)*

Correlation Coefficient: –1 –.7 –.3 0 .3 (.7) 1

Explanation: *I chose .7 because the data values are fairly close to the line. Remember, 'r' is a number between –1 and 1. In this case, from the choices given above, .7 seems the best as the other choices do not fit. For example, .3 indicates the data will be further from the*

regression line. The '1' indicates the data values would be perfectly aligned. In addition, we have a positive slope, so none of the negative 'r' values will apply.

c) Assuming that the relationship is linear, draw a regression line that best represents the data. FIND THE EQUATION OF YOUR REGRESSION LINE using the techniques learned in class.

The regression line is drawn on the above graph. It is the dotted line.

Equation:

Find the slope and y – intercept. First choose two points. As we learned in our study of lines, the easiest point to use is the **y-intercept.** Why? Because it is where the line crosses the
y-axis. I drew my line through (0, 3). The other point I chose was (3.5, 20).

I chose (3.5, 20) because my regression line goes through this point. Make sure your second point is ON YOUR LINE. You can choose any point on the graph through which the line runs. It need NOT be a data value. (More on that later.) Here, the data values were easy to see as my line went through them nicely.

Slope: (0, 3) and (3.5, 20)

$$m = \frac{\Delta y}{\Delta x} = \frac{(20)-(3)}{(3.5)-(0)} = \frac{17}{3.5} = 4.857... = 4.86$$ (round to two or three decimal places)

Equation: $y = mx + b$

$y = 4.86x + 3$

d) Explain in words the meaning of YOUR particular slope and y-intercept.

i) Slope: The slope = 4.86 This means that (on average) there are 4.86 additional days absent per year (y) for each one pack of cigarettes smoked per day (x).

ii) y-intercept: The y-intercept = 3. This means that there are 3 days absent per year (on average) even if no cigarettes are smoked per day.

e) Use your equation to predict the number of days absent if a person smoked 3.2 packs of cigarettes per day. What type of estimation is this? Explain. How confident are you of your prediction?

$y = 4.86 (3.2) + 3 = 15.552 + 3 = 18.552 = 18.55$ days on average

This is interpolation. Because we are predicting in-between known data values, we are fairly confident of our prediction.

f) Use your equation to predict the number of days absent if a person smoked 5 packs of cigarettes per day. What type of estimation is this? Explain. How confident are you of your prediction?

$y = 4.86 (5) + 3 = 24.3 + 3 = 27.3$ days on average

This type of estimation is extrapolation. We are predicting outside of our given data so we are not as confident of our prediction.

g) Can you conclude that there smoking more packs per day causes absenteeism?

Answer: No. We cannot conclude that smoking causes absenteeism as there are so many other factors that should be considered. For example, perhaps a person is absent because he/she needs to care for an ill relative. A person may be absent because he/she had an appointment with his/her child's teacher, etc.

Example 5: You may have heard that people in the US tend to save proportionately less of their income than do people in other countries, such as Japan. (I never seem to be able to save much of anything!)

However, there is another factor that needs to be considered in saving part of one's income, namely the tax rate that is charged on the interest earned from your savings. In other words, the money that you save or invest, will earn interest, and you must pay tax on that interest. The tax rate on the interest varies greatly from one country to another and one wonders whether there is a connection between the following two figures:

i) the percentage of one's income saved and **ii)** the tax rate on the interest.

Let's explore this relationship. The table below compares data for eight countries.

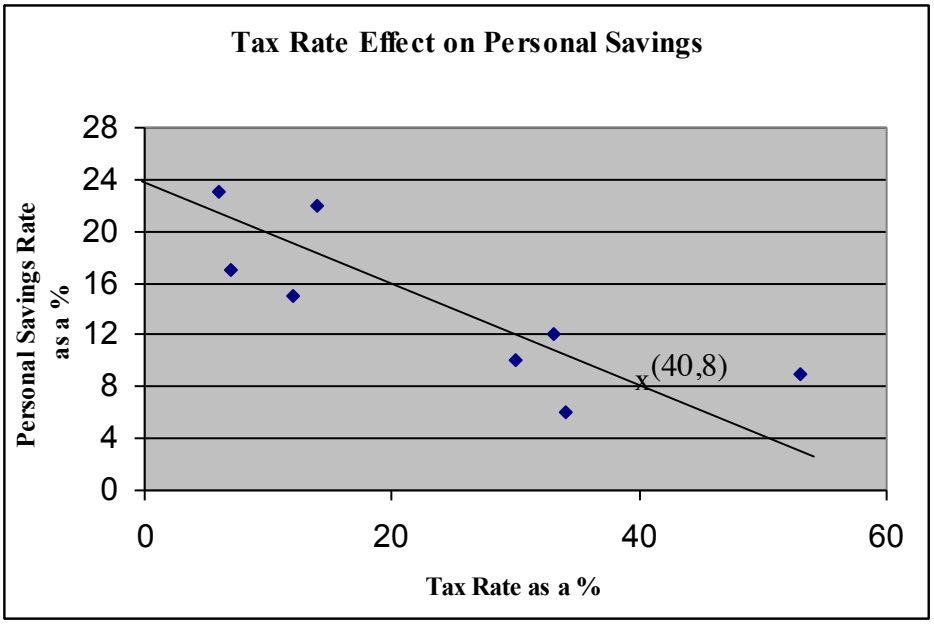
Country	Personal Savings Rate as a %	Tax Rate on Investments(%)
Italy	23	6
Japan	22	14
France	17	7
W. Germany	15	12
United Kingdom	12	33
Canada	10	30
Sweden	9	53
United States	6	34

a) Graph the data on a scatterplot. What goes on the horizontal axes? BE CAREFUL HERE! Notice that the columns in the table MAY NOT be in the order (X,Y)!

b) By eye, draw a regression line, i.e. the line which best fits (represents) the trend of the data.

Solution:

a) and b)Tax Rate on Investments goes on the horizontal axes because it is the independent variable.



c) Describe in words any trend or pattern that you see. What kind of correlation is there between the two variables? Estimate an 'r' value.

Answer: It appears that for low Tax Rates there is a tendency towards higher Personal Savings Rate and for high values of the Tax Rate there seems to be low Personal Savings Rates. There is a moderate negative correlation between Personal Savings Rate and Tax Rate on Investments. I might estimate that **r = – 0.6.** Notice the negative sign.

d) Find the equation of YOUR regression line. Describe in words the meaning of the slope and y-intercept.

Answer: Use two points on the line you drew: I chose the y-intercept because it is the easiest to use and another point ON my line, not necessarily one of the data points. So I picked: (0,24) and (40, 8). (You may choose a different point, but you must make sure it is ON YOUR line.)

$$\textbf{Slope} = \frac{24-8}{0-40} = \frac{16}{-40} = -0.4 \quad \text{this can be written } \frac{-.4}{1} \text{ or } \frac{-4}{10}$$

• This slope means that with there is an additional drop of 0.4 percentage points in the Personal Savings rate for each 1 percentage point in the Tax Rate. (Don't say .4 percent because we are working with percentage *points*. We are not taking the percent of anything.)

More simply, if we use the second form, -4/10, we can say there is an additional drop of 4 percentage points in the Personal Savings Rate for each additional 10 percentage points in the Tax Rate.

Equation of regression line is: **y = – 0.4x + 24**

y-intercept = 24 (since b = 24 and b is the y-intercept in the equation of a line.)

• The **y-intercept means** that if a country had no Tax Rate on Investments, that country's Personal Savings Rate would still be 24%. This of course is a theoretical value and may not be entirely true. If there were no taxes on investments, the Savings Rate might even be higher. Here we are also 'extrapolating' beyond known data!

e) Use your equation to predict the Personal Savings Rates that correspond to the following tax rates, and comment on the reliability of each answer:

i) 20% **ii)** 60% **iii)** 1%

Answers:

i) Tax Rate = 20%
y = - 0.4(20) + 24
y = 16

ii) Tax Rate = 60%
y = - 0.4(60) + 24
y = 0

iii) Tax Rate = 1%
y = - 0.4(1) + 24
y = 23.6

Very Reliable – Interpolation

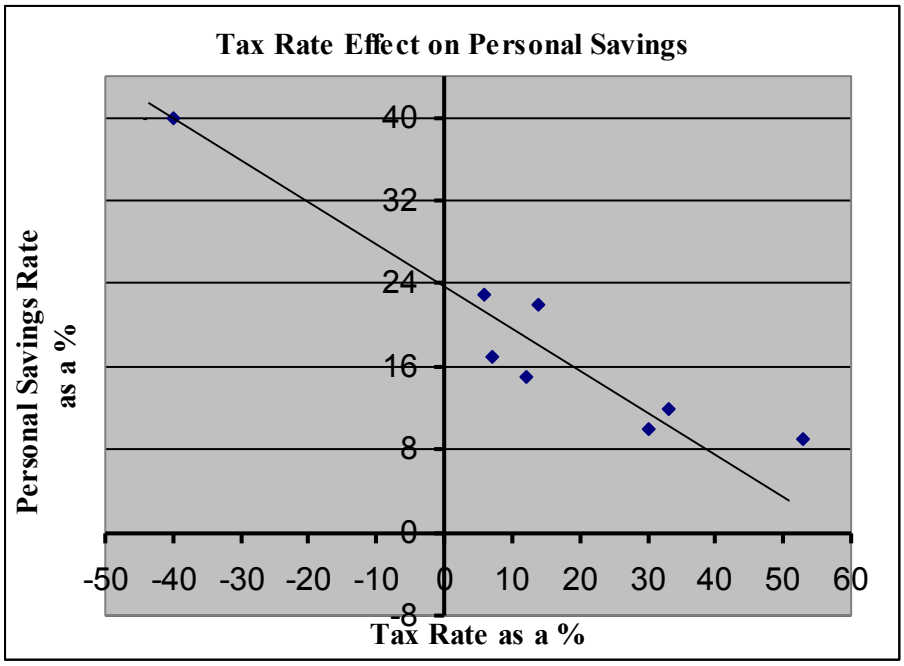
Not reliable – Extrapolation

Not quite as reliable- Extrapolation

Note: Your graph should estimate these answers quite closely. That is, your regression line should go through i) (20,16) ii) (60, 0) iii) (1, 23.6)

f) According to your equation would it be possible for a country to achieve a Personal Savings Rate of 40%? Explain why or why not.

Answer: From the graph it can be seen that in order to have a *Personal Savings Rate* of 40% our regression line would have to extend into the second quadrant indicating a negative Tax Rate on Investments. See graph below. Since it is not possible to have a negative Tax Rate, we must assume that something is wrong with our prediction. When extrapolating, the assumption of linearity may not hold. Here, perhaps at very low values of the Tax Rate, the line may curve.



g) Can you conclude that there is a cause-and-effect relationship between a country's Personal Savings Rate and its Tax Rate on Investments? Be clear.

Answer: No, you cannot assume a cause and effect relationship here. A country's Tax Rate on Investments may certainly contribute to the Rate of Personal Savings, but there are many other factors that may be involved in personal savings. Some examples may be: Financial Status, Employment Status, Inflation Rate, etc.