The Computational Basis of Measures of Central Tendency, Measures of Dispersion and Measures of Association

The following pages contain discussions, diagrams and computations useful to understanding some important univariate (one variable) and bivariate (two variables) statistical techniques commonly used in political science. We will start with univariate statistics. Univariate statistics are statistics that describe one variable (e.g., the mean, variance and standard deviation of X). Next we will work with bivariate statistics. Bivariate statistics describe the relationship between variables (e.g., the correlation between X and Y - see pages 31-34). Variable X, the independent or presumed "causal" variable, is the county unemployment rate (in percentage points) among young people. Variable Y, the dependent variable, is the county juvenile delinquency rate (also in percentage points). Thus, the county is the "unit of analysis." As we have data for the same 10 counties for both variables X and Y, our "N" (number of cases) is 10 (thus our sample size is 10). Our central hypothesis is that higher youth unemployment rates are associated with higher delinquency rates (i.e., that scores on variables X and Y are "positively" associated, see pages 2-3). Such a relationship is plausible because unemployed youth are more likely to be delinquent than employed youth. To make sure you understand what the data mean, look at page 65. County #1 has a score of 2 on variable Y. This means that the delinquency rate in county #1 is 2%. County #1 has a score of 10 on variable X. This means that the youth unemployment rate in county #1 is 10%. The data that will be analyzed are adapted from page 231 of Quantitative Methods for Public Administration, second edition, by Susan Welch and John Comer.

In order to understand the computations on page 65 let us first review some procedures that were discussed on pages 21-23. For example, in the following expression:

$$\Sigma(X - \bar{X})^2$$

the order of operations is: (1) subtract the mean of X (symbol $\bar{X}$) from the first score on variable X, (2) square the result, (3) repeat this process for as many scores as you have (in our case we would undertake this process 10 times) and (4) add (Σ is called a summation operator - i.e., "add") these squared values. Thus, we would square before adding. The aforementioned procedures are those appearing on page 65, columns 5 and 6. The entry for county #1 in column 6 (900) is the square of the entry for county #1 in column 5 (-30). Make sure you learn the symbols at the top of the columns on the next page as they are introduced over the following pages. The quizzes will require you to calculate some of the statistics prior to their being discussed in class. You do not need to memorize any formulas (except one simple formula so designated on next week's reading assignment). I will provide the data and formulas on the quiz. You will have to calculate the correct answer and interpret it. As the numbers on

Statistical Computations

→ Please Note: Go to page 66 and return to page 65 when told to over the next several pages.

	Column 1	Column 2	Column 3	Column 4	Column 5	Column 6	Column 7	Column 8	Column 9	Column 10	Column 11	Column 12
Units	Y	X	$(Y-\bar{Y})$	$(Y-\bar{Y})^2$	$(X-\bar{X})$	$(X-\bar{X})^2$	$(X-\bar{X})(Y-\bar{Y})$	$\hat{Y}$	$(Y-\hat{Y})$	$(Y-\hat{Y})^2$	$(\hat{Y}-\bar{Y})$	$(\hat{Y}-\bar{Y})^2$
County #1	2	10	−7	49	−30	900	210	3.07	−1.07	1.14	−5.93	35.16
County #2	4	20	−5	25	−20	400	100	5.05	−1.05	1.1	−3.95	15.60
County #3	5	20	−4	16	−20	400	80	5.05	−0.05	0.0025	−3.95	15.60
County #4	5	40	−4	16	0	0	0	9	−4	16	0.00	0.00
County #5	9	30	0	0	−10	100	0	7.02	1.98	3.92	−1.98	3.92
County #6	10	30	1	1	−10	100	−10	7.02	2.98	8.88	−1.98	3.92
County #7	11	50	2	4	10	100	20	10.98	0.02	0.0004	1.98	3.92
County #8	13	70	4	16	30	900	120	14.92	−1.92	3.68	5.92	35.05
County #9	14	70	5	25	30	900	150	14.92	−0.92	0.84	5.92	35.05
County #10	17	60	8	64	20	400	160	12.95	4.05	16.4	3.95	15.60
	$\bar{Y}=9$	$\bar{X}=40$										
	$\Sigma=90$	$\Sigma=400$	$\Sigma=0$	$\Sigma=216$	$\Sigma=0$	$\Sigma=4200$	$\Sigma=830$			$\Sigma=51.96$		$\Sigma=163.8$

a quiz will be easy to work with (e.g., 5-3), you will not need a calculator. You will not need to calculate square roots.

As noted in a previous reading (pages 15-16) it is often desirable to calculate a measure of central tendency (i.e., an "average"). The computations in this reading assume the data are measured at either the interval or ratio level ("percentages" are ratio level measures - see page 11). Thus, for a measure of central tendency we can utilize "the mean." The symbols used for the various statistics we will calculate (mean, variance, standard deviation, covariance, correlation and regression) refer to sample estimates of the population values for these statistics (just keep reading). Remember that the ten counties used in this example represent a sample. For example, you might think of these ten counties as ten randomly selected counties in California. The "population of interest" might be all the counties in California. We would be trying to estimate what the following statistics would be for all counties in California (e.g., the mean level of juvenile delinquency in California) from the ten counties which are used in this study. When you see Greek letters used in political science research, the researcher is referring to population values (the statistic in the "population of interest"), not those from a sample.

$$\textit{Sample Mean of Variable } X = \bar{X} = \frac{\Sigma X}{N} = \frac{400}{10} = 40$$

(400: page 65, bottom of column 2; 10: 10 counties)

$$\textit{Sample Mean of Variable } Y = \bar{Y} = \frac{\Sigma Y}{N} = \frac{90}{10} = 9$$

(90: page 65, bottom of column 1)

Previously we discussed the utility of a measure of dispersion (pages 17-24). The basic idea of a measure of dispersion is to see how representative our measure of central tendency (in this case the mean) is of all the scores in the distribution. The purpose of the variance (page 23) is to find the spread (variation) of the scores on a variable around the mean value. The purpose of the standard deviation is to return the variance to the original units of measurement. For example, the standard deviation of variable X is 21.6 percentage points of unemployment. Alternatively, the variance of variable X is 466.7 squared percentage points of unemployment. I think most people would probably find it easier to think in terms of unsquared units (i.e., the standard deviation) than squared units (i.e., the variance). For example, what would 16 squared units of unemployment mean to someone? Consequently,

when we discuss the distribution of scores on a variable we will use the standard deviation. Another advantage of the standard deviation is that we can make "percentage distribution" statements. For example, if we have a normal distribution, we know that approximately 68% of the scores are between one standard deviation below the mean and one standard deviation above the mean (see page 24). For example, if we have a normal distribution and the mean is 50 and the standard deviation is 5, then we know that approximately 68% of the scores are between 45 and 55 (50 - 5 = 45 and 50 + 5 = 55 - see page 24). Additionally, if we have a normal distribution, we know that approximately 95% of the scores are between two standard deviations below the mean and two standard deviations above the mean (page 24). If the distribution is non-normal we can use Tchebysheff's Theorem (page 25). The variance does not permit us to make percentage distribution statements. The variance will become important when we use regression.

page 65, bottom column 6

$$\text{Sample Variance of } X = S_x^2 = Var(X) = \frac{\text{Sample Variation of } X}{N-1} = \frac{\Sigma(X-\bar{X})^2}{10-1} = \frac{4200}{9} = 466.7$$

$$\text{Sample Standard Deviation of } X = S_x = \text{Positive Square Root of the Sample Variance of } X = \sqrt{466.7} = 21.6$$

page 65, bottom column 4

$$\text{Sample Variance of } Y = S_Y^2 = Var(Y) = \frac{\text{Sample Variation of } Y}{N-1} = \frac{\Sigma(Y-\bar{Y})^2}{10-1} = \frac{216}{9} = 24$$

$$\text{Sample Standard Deviation of } Y = S_Y = \text{Positive Square Root of the Sample Variance of } Y = \sqrt{24} = 4.9$$

As the computations on the previous page indicate, the standard deviation of X is greater than 50% of the mean of X (i.e., 21.6 is greater than 50% of 40). This tells us that the mean of X was probably the result of several (or many) scores being quite far from the mean rather than a series of scores being very close to the mean. Thus, variable X has a high degree of dispersion (review the first paragraph on page 23).

The standard deviation permits us to formulate a percentage distribution of the scores on a variable (see page 24). If variable X is normally distributed, approximately 68% of the scores should be within plus or minus 21.6 units of the mean score of 40 (see page 24). Is variable X normally distributed? Look down column 2 on page 65.

Why is "N" in the denominator of the mean and variance? "N" appears in the denominator of many statistics in order to "weight" the total score by the number of observations. Putting "N" in the denominator means that the answer will be on a "per observation" (or "per score") basis. Thus, a mean of 40 means an average "per score" of 40.

Why did we divide by N - 1 (rather than N) to get an estimate of the sample variance? An intuitive answer can be developed based upon the concept of degrees of freedom. Our sample is known to contain N (10) data points. However, in computing the sample variance a necessary first step was the computation of the sample mean. Computing the sample mean places one constraint upon the N data points. The constraint is that the N observations must sum to N times the computed mean ($\bar{X}$). Thus, if the mean on variable X is 40 and we have 10 observations, the total of the 10 scores on X must sum to 400. Once we know the first nine scores the tenth score is "constrained" (i.e., not free to assume any value). For example, on page 65, column 2, the first nine scores on variable X sum to 340. Given a mean of 40 and a total of 10 scores, the tenth score on variable X must be 60 (as it is for county #10). This leaves N - 1 unconstrained observations with which to estimate the sample variance given the mean (adapted from Econometric Models and Economic Forecasts, 2nd ed., by Robert S. Pindyck and Daniel L. Rubinfeld, page 25). While knowing the reason for dividing by "N" in the expression N - 1 is very important, the reason for -1 is much less important.

Thus far we have only been working with univariate (one variable) statistics. Since we want to test hypotheses (pages 2, 3 and 5), we must relate scores on one variable to scores on another variable. The next page will begin the process of hypothesis testing by introducing "covariance." The purpose of covariance is to determine if an individual scores above the mean on variable X, how likely will the same individual score above, or below, the mean on variable Y? Thus, we will be examining how scores on variables X and Y vary together (i.e., "covary"). If higher scores on X are associated with higher scores on Y, the relationship between X and Y is "positive" (review pages 2-3 on "positive" and "negative" relationships). Visualizing variable relationships is very helpful in understanding covariance. This is the purpose of pages 70-74.

Estimated (or sample) Covariance of X and Y

$$Cov(X,Y) = \frac{\Sigma(X - \bar{X})(Y - \bar{Y})}{N - 1} = \frac{830}{10-1} = \frac{830}{9} = 92.2$$

from page 65, bottom of column 7

on "N - 1" see page 68

we have data on 10 counties

The key to understanding the above formula is the first entry in column 7 on page 65: 210 (this is critical to your next quiz). To obtain 210 we first see how the score on variable X for county #1, 10 (see page 65, first entry in column 2) deviated (i.e., differed) from the mean score on variable X, 40 (see page 66). Since 10 - 40 = -30, county #1 "deviated" by -30 points from the mean score on variable X (i.e., was 30 points lower than the mean; -30 is the first entry in column 5). We now see how the score on variable Y for county #1, 2 (see page 65, first entry in column 1) deviated from the mean score on variable Y, 9 (see page 66). Since 2 - 9 = -7, county #1 deviated by -7 points from the mean score on variable Y (i.e., was 7 points lower than the mean; -7 is the first entry in column 3). We now multiply these deviations to obtain the first entry in column 7: 210 {i.e., (-30) (-7) = 210 }. We continue this process for the remaining 9 counties {i.e., each entry in column 7 is the number in column 5 multiplied times the corresponding entry in column 3: for county #2 this is (-20) (-5) = 100 }. Finally, we sum (add) the 10 entries in column 7 to obtain the total of 830 (830 is both the numerator of the formula above as well as the total at the bottom of column 7 on page 65). Performing the remaining calculations in the above formula yields a covariance of 92.2.

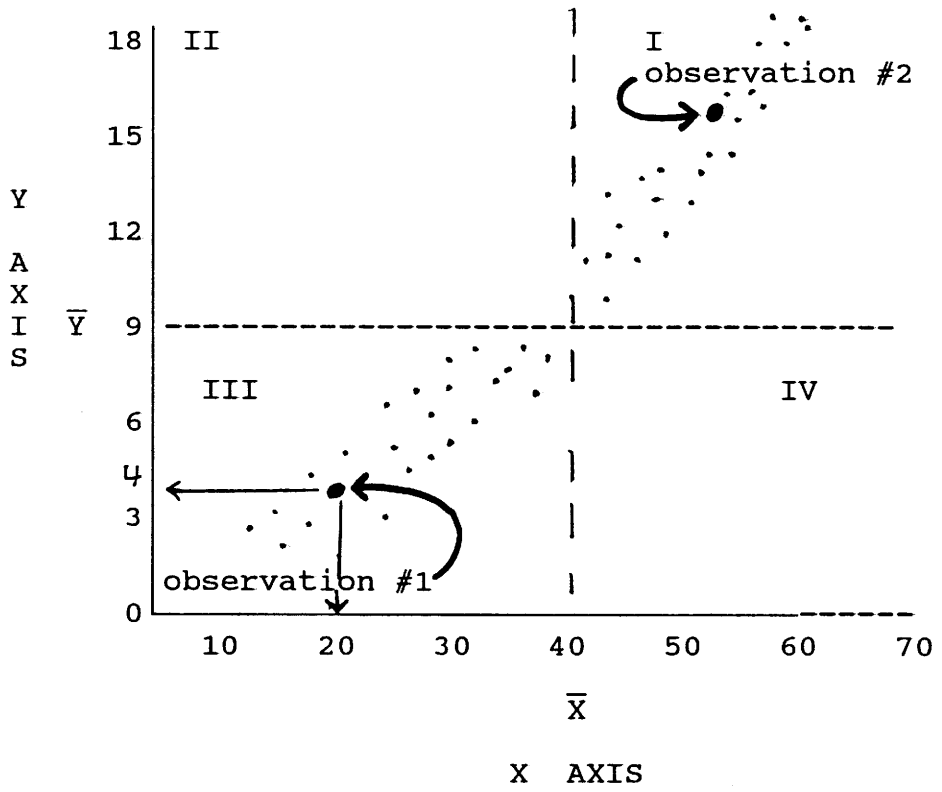
If we add the deviations from the mean on variable X (each score in column 5) before multiplying them by the corresponding deviations from the mean on Y (each score in column 3) the deviations on both X and Y total 0 (i.e., the total at the bottom of both columns 5 and 3 is 0) which when multiplied would yield a covariance of 0 { i.e., (0) (0) = 0 }. This would indicate no (i.e., 0) association between X and Y. Since there is a positive association between X and Y, 92.2, a covariance of 0 would have been incorrect. By multiplying before we add we take account of the fact that while county #1 is 30 points below the mean on variable X, it is also 7 points below the mean on variable Y. This suggests a "positive" association (i.e., counties below the mean on X are also below the mean on Y while counties above the mean on X are also above the mean on Y), which a covariance of 92.2 confirms.

Weaknesses of Covariance: What does a covariance of 92.2 tell us? We know that there is an association between X and Y (i.e., the covariance is 92.2, not 0) and that this association is "positive" (i.e., the covariance is 92.2, not -92.2). However, we do not know either the strength or magnitude of the association. To assess the strength of the association between X and Y we need a "benchmark" to compare 92.2 against and we do not have one (compared to 1,000, 92.2 is a "small" number but compared to 10, 92.2 is a "large" number). Correlation, on page 75, will tell us the strength while regression, page 76, will tell us the magnitude (how much, on average, Y increases with an increase in X).

The next quiz, coming the day this assignment is due, will ask you to calculate and interpret the covariance. You do not need a calculator. The covariance formula will appear on the quiz. Before introducing correlation and regression, pages 70-74 will visually show variable relationships (positive, negative, etc.).

The diagrams on pages 70-74 illustrate possible relationships between variables X and Y (i.e., positive, negative or no association - again, see pages 2-3). The X axis (range of possible scores on variable X) runs from left to right while the Y axis (range of possible scores on variable Y) runs from top to bottom. Each dot represents one county's (or observation's) score on both X and Y. Although I have retained the same mean (X = 40; Y = 9) and range of scores for both X and Y as appear on page 65, columns 1 and 2, the examples on pages 70-74 contain many more observations (many more than 10 counties, i.e., "dots") than in the data set on page 65. Please note that observations #1 and #2 below (as well as on page 71) are not the same scores on variables X and Y as counties #1 and #2 on page 65. I am just presenting diagrams illustrating possible relationships between variables X and Y. If you do not immediately understand the following diagram, do not "panic"! Just keep reading!!! Diagrams that show the relationship between scores on two, or more, variables are often called "scatter plots."

Strong Positive Covariance

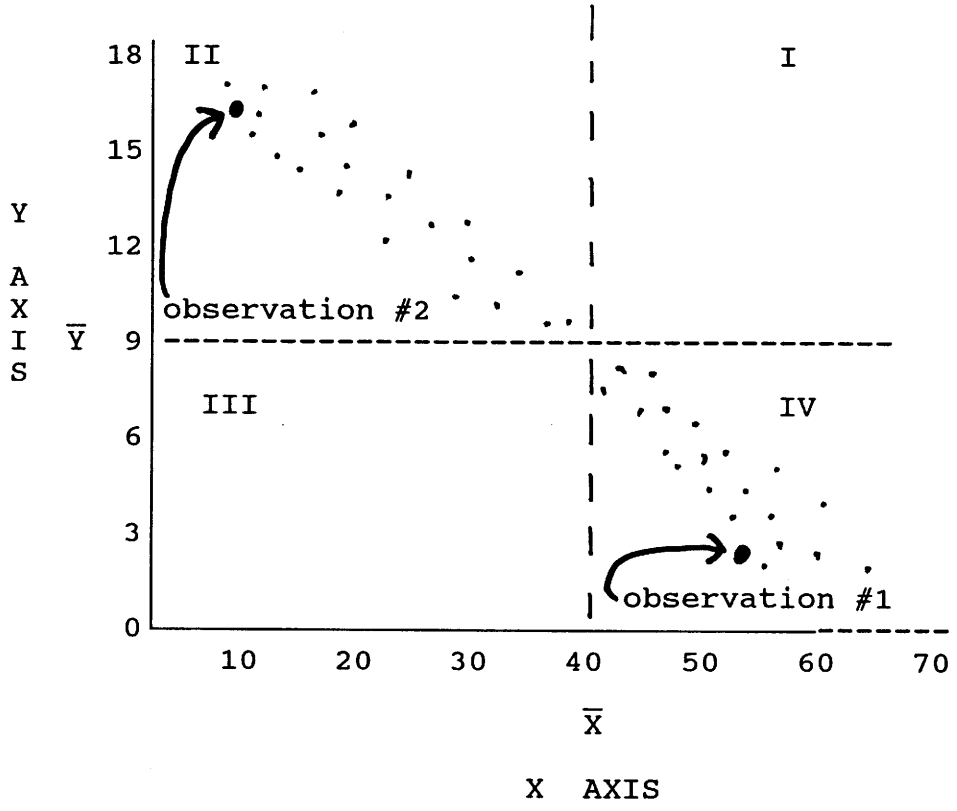


Observation #1 above represents a county with a score of approximately 20 on variable X (which is well below the mean score on X of 40) and approximately 4 on variable Y (which is well below the mean score on Y of 9). Dots in quadrant III ("III" above) indicate that when a county scores below the mean on variable X the same county scores below the mean on variable Y (as did observation #1). Observation #2 above represents a county which scored

approximately 53 on X (which is above the mean of 40) and approximately 16 on variable Y (which is above the mean of 9). Dots in quadrant I ("I" in the diagram on page 70) indicate that when a county scores above the mean on variable X the same county scores above the mean on variable Y (as did observation #2). If a county scores above the mean on variable X the same county scores above the mean on variable Y. Since lower scores on variable X are associated with lower scores on variable Y and higher scores on variable X are associated with higher scores on variable Y, the relationship between X and Y is "positive" (see pages 2-3). As few of the dots do not fit this pattern (i.e., are in quadrants II or IV), we can say this is a "strong" (or even "very strong") positive relationship (see the table on the bottom of page 34).

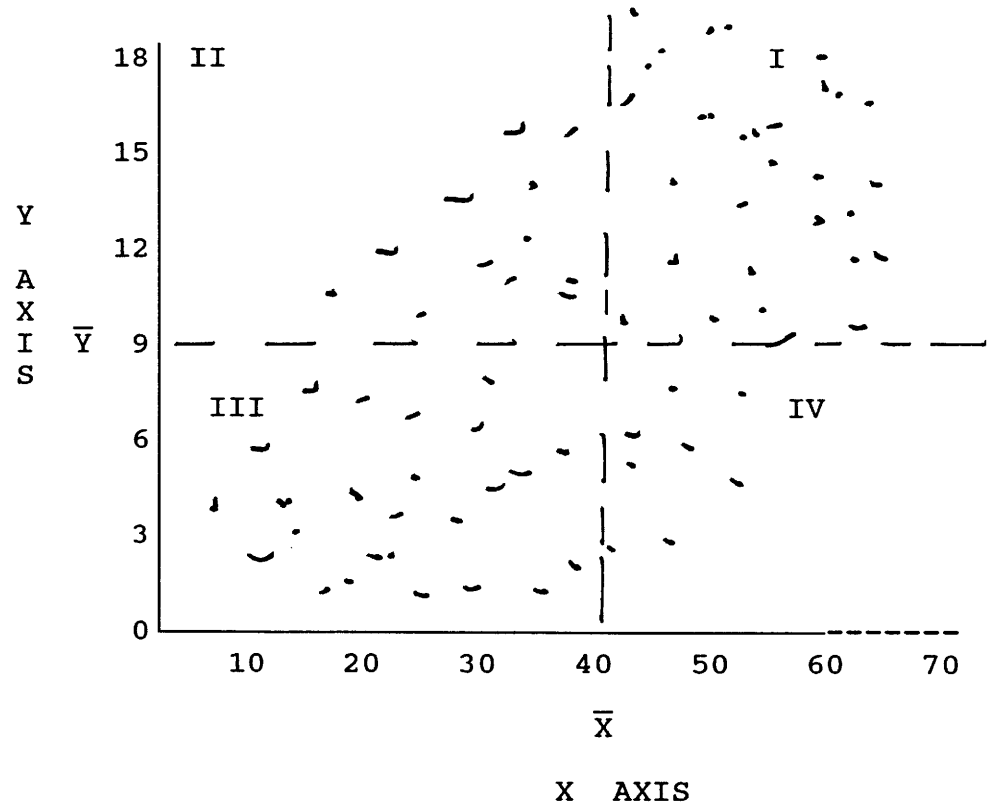
In the scatter plot below, dots in quadrant IV indicate that when a county (i.e., observation) scores above the mean on variable X the same county scores below the mean on variable Y (see observation #1). Additionally, dots in quadrant II indicate that if a county scores below the mean on variable X the same county scores above the mean on variable Y (see observation #2). Since higher scores on variable X are associated with lower scores on variable Y and lower scores on variable X are associated with higher scores on variable Y, the relationship between X and Y is "negative." As few of the dots do not fit this pattern (i.e., are in quadrants I and III), we can say that this is a "strong" (or even "very strong") negative relationship.

Strong Negative Covariance



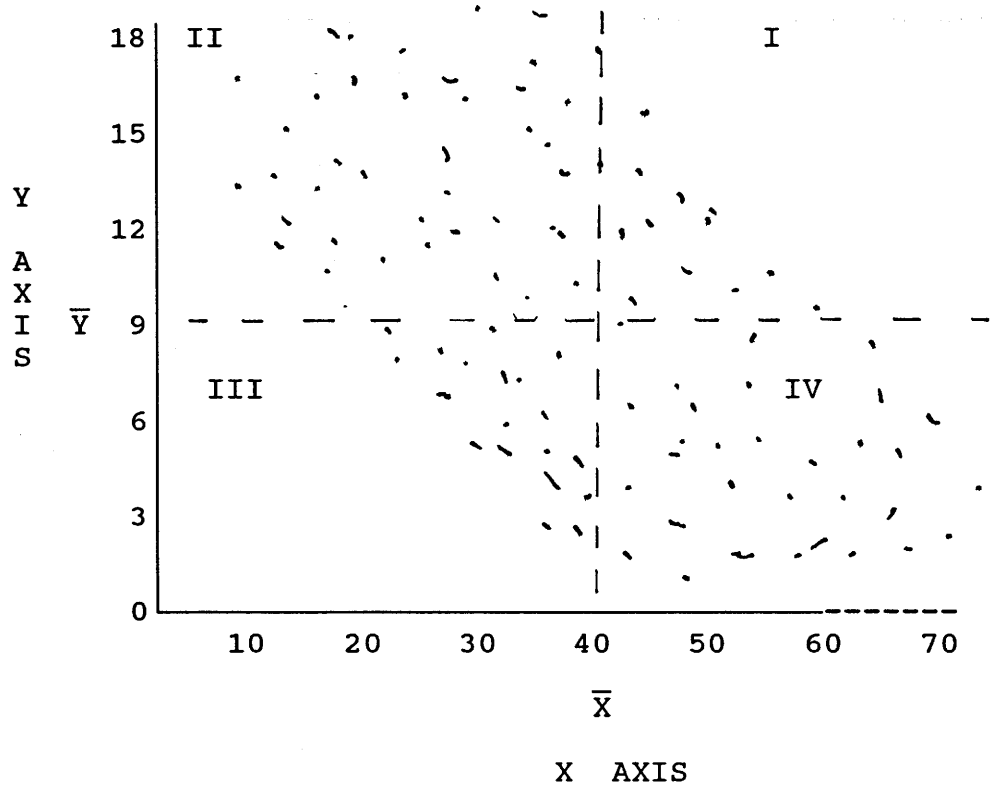
Weak Positive Covariance

The scatter plot below has the same basic pattern as the "strong positive covariance" (page 70). However, notice that we now have a number of counties in quadrants II and IV. Thus, although "positive" the relationship between X and Y is "weaker" (i.e., less "strong") than on page 70. From page 69 we know that the covariance of X and Y in our sample is 92.2. Since this number is positive (remember the covariance is not -92.2), we know that (except for the number of dots) the graphical representation of this covariance must be similar to the drawing on either page 70 or 72. We know this because the graphs on pages 71 and 73 are for negative covariances. As the graph on page 74 is for a covariance of approximately 0, it probably does not represent our findings either. Unfortunately, from the answer of 92.2 we can not say whether the graph on page 70 or 72 is closer to our result. This is one reason why we will later (page 75) calculate Pearson's Product Moment Correlation. Correlation, when used in conjunction with the table on the lower half of page 34, will let us determine whether the graph on page 70 or 72, is closer to a pictorial representation of our covariance of 92.2.



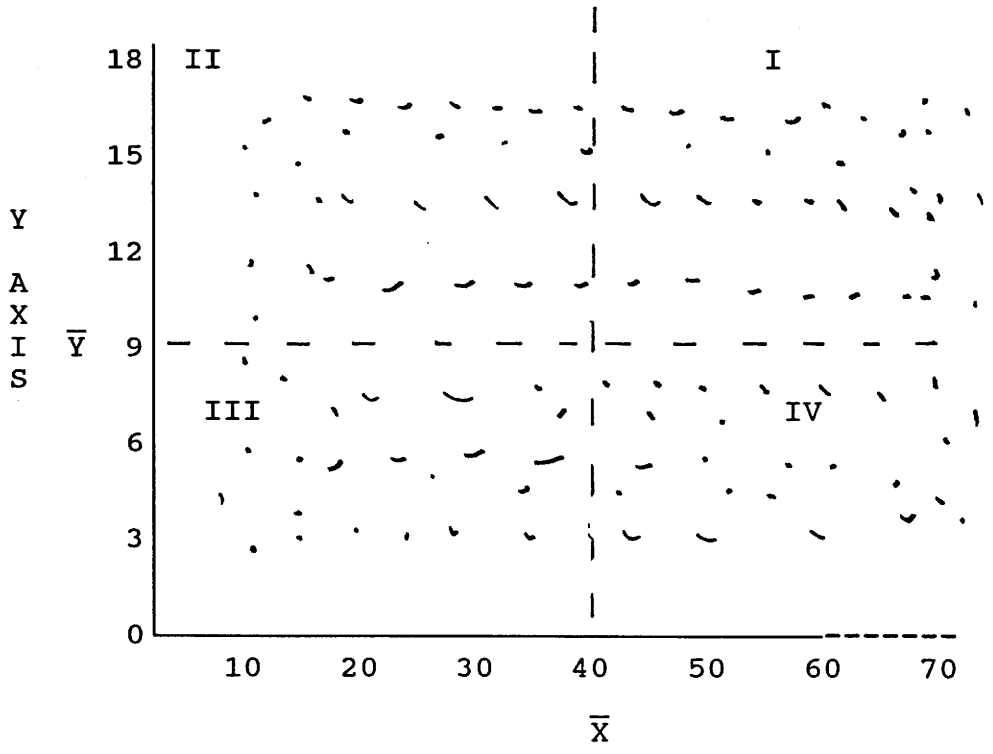
Weak Negative Covariance

The scatter plot below has the same basic pattern as the "strong negative covariance" (page 71). However, notice that we now have a number of counties in quadrants I and III. Therefore, while "negative," the relationship between X and Y is "weaker" (i.e., less "strong") than on page 71.



In the scatter plot below, if a county is above the mean on variable X the same county is about equally likely to be either above, or below, the mean on variable Y. Therefore, knowing whether or not the county scored higher than average (i.e., the mean) on variable X is of no help in predicting whether or not the county would score higher than average on variable Y. Thus, we can say that X and Y are "not associated," "unrelated," or "independent" (three interchangeable terms) of each other.

No Association



Remember that covariance (page 69) was our first attempt to calculate a measure that assessed the degree of association between variables X and Y. This is extremely important because one of the major goals of this course is to estimate the magnitude of the relationship(s) between the variables (see page 5). The discussion on the last two paragraphs of page 69 explained why covariance is not a good method of assessing the relationship between variables X and Y. Pages 75-76 discuss correlation and regression, which are better methods of assessing the relationship between scores on variable X and scores on variable Y. So, let's examine these "better" methods!

When we calculated the covariance (see page 69), we discovered that the relationship between variables X and Y is "positive" (i.e., 92.2 and not -92.2). However, we could not say how "strong" or "weak" this positive relationship is. In the formula below, notice how the numerator is the covariance from page 69, while the denominator is the product (i.e., multiplication) of the standard deviation of X and the standard deviation of Y. Let me mention that the covariance of two variables (the numerator below) must be equal to or less than the product of their individual standard deviations (the denominator below). If the numerator must either be equal to or less than the denominator, the ratio (i.e., the correlation formula below) must produce an answer between +1.0 and -1.0 (if the numerator is negative the correlation is negative because the denominator must be positive - all standard deviations are positive - and a negative numerator divided by a positive denominator yields a negative answer). Accordingly, we can compare the "strength of different correlations" (e.g., .87, the correlation below, is "stronger" than a correlation of .29, see the table on the bottom of page 34). Be sure you do not confuse the "direction" (i.e., positive or negative) with the "strength" of association (re-read the bottom half of page 33 before continuing).

Pearson's Product Moment Correlation of X and Y:

$$r_{XY} = \frac{Cov(X,Y)}{\left(\begin{matrix}Standard\\Deviation\\of\ X\end{matrix}\right)\left(\begin{matrix}Standard\\Deviation\\of\ Y\end{matrix}\right)} = \frac{92.2}{(21.6)(4.9)} = \frac{92.2}{105.8} = .87$$

see page 69

see page 67

Since our answer is .87 (again, not -.87) we know that we have a "very strong" positive relationship between scores on variable X and scores on variable Y (see the bottom half of page 34 on interpreting measures of association). Thus, the graph on page 70 is much closer to our result than the graph on page 72.

However, correlation has the same weakness as Gamma and Kendall's Tau (pp. 35-38) in that it can not tell us the magnitude of the relationship between variables X and Y. For example, a one percentage point increase in unemployment is likely to lead to how many percentage points of increase in delinquency? We do not know. All we can say is that the county unemployment rate is highly correlated with the county delinquency rate. Since one of the major goals of an empirical analysis is to estimate the magnitude of the relationships between variables (see page 5), this is a critical limitation. Regression, the subject of the next series of pages and the bulk of the rest of the semester, will tell us the magnitude of the association between X and Y.

The Computational Basis of Regression

As has been discussed on pages 36 and 75, Pearson's Product Moment Correlation does not tell us the magnitude of the relationship between X and Y. Thus, suppose you were asked the following question: if the youth unemployment rate increased by 1% in a particular county, how much should we expect the delinquency rate in that same county to increase? As of yet, we do not know. All we know so far is that since the correlation between the youth unemployment rate and the delinquency rate is positive (i.e., .87), we should expect an increase in the youth unemployment rate to be associated with an increase in the delinquency rate. However, we do not know by how much. The magnitude of the relationship between the youth unemployment rate and the delinquency would tell us how many units we should expect the delinquency rate to increase if the youth unemployment rate increased by one unit (just keep reading). Since both the youth unemployment rate and the delinquency rate are measured in percentage points (e.g., a "unit" of unemployment is a percentage point of unemployment). The magnitude of the relationship between the youth unemployment rate and the delinquency rate would tell us how many percentage points the delinquency rate would be expected to increase if there was a 1% increase in the youth unemployment rate. In order to find the magnitude of the relationship between the youth unemployment rate and the delinquency rate, we turn to regression and use the following formula.

Estimate of the Magnitude of the Relationship between X and Y - Symbolized by "b":

$$b = \frac{Cov(X,Y)}{Var(X)} = \frac{92.2}{466.7} = .198$$

see page 69

see page 67

Now we have the magnitude of the relationship between variables X and Y. If the unemployment rate in a particular county (variable X) increases by one unit (since unemployment is measured in percentage points this would be a one percentage point increase) we can expect the delinquency rate in that same county to increase, on average, by .198 units. This is approximately one-fifth of a unit (since .198 is almost .200 and .200 is one-fifth of 1.00). Since a "unit" of delinquency is a percentage point this would mean approximately a one-fifth of one percentage point increase. The magnitude, or "b," is always stated in the units of the dependent variable. Now we not only know the magnitude of the relationship between variables X and Y, but we can forecast the impact of various changes in variable X on variable Y. For example, if a recession produces a three percent increase in the unemployment rate in a county, we would predict (i.e., expect) the delinquency rate in that particular county to rise (increase) by approximately six-tenths of one percent [(.198)(3) = approximately .6].

It is important to note that when we estimated "b" we had to specify which variable was the independent variable and which variable was the dependent variable. Thus, the .198 estimate of "b" is the impact that youth unemployment has on delinquency, not the impact that delinquency has on youth unemployment. Look back at the third paragraph on page 38 and you will see that we do not have to specify which variable influences which other variable (i.e., which variable is independent and which is dependent) when using either cross tabulation or a measure of association (e.g., Pearson's Product Moment Correlation on page 75). By forcing the political scientist to state which variable is independent and which is dependent, regression (and all other statistical procedures we will discuss) requires us, for example, to think through why and how youth unemployment might influence delinquency and why delinquency might not influence youth unemployment. Such thinking is a very important step in the development of political science as a science.

Make sure you know how to interpret "b." The next several quizzes may well ask you to interpret specific values of "b." For example, if I give you a quiz that asks you to interpret a value for "b" of -.577, what would you write? I hope you would say that if X increases by one unit, then Y, on average, will decrease by almost six-tenths of one unit. In the last sentence, note the term "on average." The relationship between X and Y (which is "b") is not deterministic (just keep reading!). If X increases by one unit Y does not have to decrease by almost six-tenths of one unit. Y may decrease more or less than approximately six-tenths of one unit. It just means that "on average" Y will decrease by almost six-tenths of one unit. Furthermore, make sure that you do not interpret "b" as you would a measure of association. Thus, you should not use the diagram on the bottom of page 34 in interpreting "b." Remember, "b" is a regression statistic, not a measure of association (i.e., not gamma, Kendall's Tau_b or correlation).

If I give you the specific units of measure for X and Y on a quiz I expect you to use them. For example, if variable Y is thousands of dollars of annual income (i.e., a score of 12.1 = $12,100), variable X is education measured in years and the value of "b" is .975 this would mean that each additional year of education was associated (on average) with an increase of over 9/10s of a unit in annual income. However, since a "unit" of income is a $1,000 couldn't we say an increase of 9/10s of $1,000? Yes! However, isn't 9/10s of $1,000 also equal to $900? Yes! If you were talking to someone wouldn't you say $900 rather than 9/10s of $1,000? Yes! Thus, to score the maximum points on a quiz you would need to say that a one year increase in education is associated (on average) with a $900 increase in income. Since "b" is .975 it's actually a $975 increase. I can't say the following too strongly: on the day this reading assignment is due make sure you understand everything discussed in this paragraph, the previous paragraph, the interpretation and weakness of correlation on page 75, the meaning and interpretation of "a" in the upcoming discussion on pages 78-79, the meaning of "e" in the upcoming discussion on pages 79-80 and, most importantly, how the covariance on page 69 is calculated and interpreted. In reviewing page 69, make very sure you know how each calculation that results in the 830 at the bottom of column 7 on page 65, as well as all other calculations necessary to obtain the covariance, are performed and in what order (review paragraphs #1 and #2 on page 69). On the day this reading assignment is due, you will need to calculate, as well as interpret, the covariance. You will not need a calculator or to memorize the formula.

The Y Intercept - Symbolized by "a"

On page 76 we calculated the magnitude of the relationship between X and Y. As you remember, the magnitude, or "b," for our data is .198 (see page 76). For our data, this means that if the youth unemployment rate (variable X) increases by 1% we expect the delinquency rate (variable Y) to increase by almost .2% (i.e., .198 is almost .200 which is two-tenths of one percent). It is often easier to grasp the concept of the magnitude, or "b," by using a graph (just keep reading, it will become clear over the next several pages). The magnitude, or "b," can be thought of as the slope (i.e., angle) of a line depicting the relationship between X and Y. Such a line is called the regression line. In order to graph the regression line we need two items: the slope (i.e., angle) which is the magnitude, or "b," and the point from where the regression line originates. The point where the regression line originates is called "a," or the Y intercept (i.e., the point on the Y axis where the regression line "intercepts" or intersects, the Y axis - just keep reading). Both "a" (the Y intercept) and "b" (the slope of the line between X and Y) are depicted in the diagram on the top of page 81 ("a" is on the left side of the diagram and "b" is the slope of the line). The formula for "a," or the Y intercept is as follows:

$$a = \overline{Y} - b\overline{X}$$

See page 66 *See page 76*

$$a = 9 - (.198)(40)$$

$$a = 9 - 7.92$$

$$a = 1.08$$

One of the many useful features of regression is that once we have values for "a" and "b," we can then use this information, together with the score on variable X for a particular observation (in our situation, a county) to predict the score on the dependent variable for that same observation (in our situation this would mean to predict the delinquency rate in a particular county). Just keep reading!! Remember that prediction is one of the goals of science (see bottom of page 4). Also remember that we were unable to make predictions when we used either cross tabulation or measures of association (see page 37). In order to use the values for "a," "b" and "X" to predict a score on Y, we will employ the following formula.

see page 65, column 8

$$\text{Predicted Value of } Y = \hat{Y} = a + bX$$

For County #1, with an unemployment rate of 10% (see the first score in on page 65, column 2), this is as follows:

$$\hat{Y} = a + bX$$

$$\hat{Y} = 1.08 + (.198)(10)$$

$$\hat{Y} = 1.08 + 1.98$$

$$\hat{Y} = 3.06$$

Note: the first entry in page 65, column #8 is 3.07 (the discrepancy is due to rounding). If X = 0, the term "bX" in the above formula would reduce to 0 [because (.198)(0) = 0] and we would be left with "a" (1.08) as the predicted value for Y. Thus, 1.08 would be the "predicted" delinquency rate in a county with a zero percent youth unemployment rate. As no county has a youth unemployment rate of zero percent (the lowest rate is 10 percent for county #1) it is risky to predict outside the range of our data.

Looking at the line on the top of page 81, place your finger on the line at the point marked 3.06 (it is about one inch above the number 10 on the X axis - i.e., a county with an unemployment rate of 10%). If you move your finger across to the left side of the page you will intersect the Y axis at approximately 3.06 (i.e., a county with a predicted delinquency rate of 3.06%). Thus, a county with a 10% unemployment rate would be predicted, by the line, to have a delinquency rate of 3.06%. A county with a 20% unemployment rate would be predicted to have a delinquency rate of 5.05%. Since higher scores on X (unemployment) are associated with higher predicted scores on Y (delinquency) the relationship between X and Y, "b" (i.e., the regression line), has a positive slope.

Make sure you know how to interpret "a" (the Y intercept). I frequently ask this on quizzes. For example, if the quiz says interpret "a" = 5, what would you write? I hope you would say that if X is zero, then the "predicted" value of Y = 5. Note that it is the "predicted" value of Y that is 5. The prediction may be incorrect. Do not say that if X = 0 then Y = 5. Also, you should say that "a" is the Y intercept. If you are given the units of measure, you should use them in your interpretation. For example, if variable Y is thousands of dollars of annual income (i.e., a score of 12.1 = $12,100), variable X is education measured in years and the value of "a" is 5, this would mean if someone had no formal years of education (i.e., X = 0) then they would be "predicted" to have an income of $5,000. Also, the Y intercept would be $5,000.

"The Residual" or "Error Term" - Symbolized by "e"

Let us return to our example where X is the percentage of the youth in a county who are unemployed and Y is the percentage of the youth in a county who are delinquent. In the social sciences the relationship between X and Y is usually inexact. Thus, when X = 10 (in this instance a county youth unemployment rate of 10%), Y may assume many different values. Notice the normal curve drawn at the point where X = 10 in the diagram on the top of page 81. Our best estimate is that a county with a youth unemployment rate of 10% will have a juvenile delinquency rate of 3.06%, but its actual rate may (and probably will) differ (as it does in our sample). For

this reason, we can not say that Y = a + bX. Instead, we say: Y = a + bX + e where "e" (called "the residual") is the difference between the individual's actual score on variable Y and the score that we would have predicted for this same individual on variable Y based upon this individual's score on variable X (i.e., given a county unemployment rate of 10% we would predict that this county would have a juvenile delinquency rate of 3.06% - see page 79).

The "residual" ("e") is composed of the effects of omitted variables (other factors besides a county's youth unemployment rate that would undoubtedly alter our prediction of the county's juvenile delinquency rate - for example, whether or not the county had a gang prevention program - we do not have data on this) and measurement error in the dependent variable (we may have inaccurately measured the juvenile delinquency rate in this particular county). Think of it this way: for each value of X there is a whole distribution of possible scores on Y with the average (or "mean") score in this distribution of Y (for that particular value of X) equal to the value designated by the regression line (for X = 10 this "average" value for Y is 3.06%). Since the value for "b" is positive (.198 as opposed to -.198) the distribution of Y for each value of X is "centered" on progressively "higher" values of Y as X increases (again, see the diagram of the top of page 81). The symbols and computation of the residual appear below.

$$Residual = e = Y - \hat{Y}$$

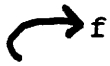
For county #1 this is as follows:

$$e = Y - \hat{Y}$$

see bottom of page 78

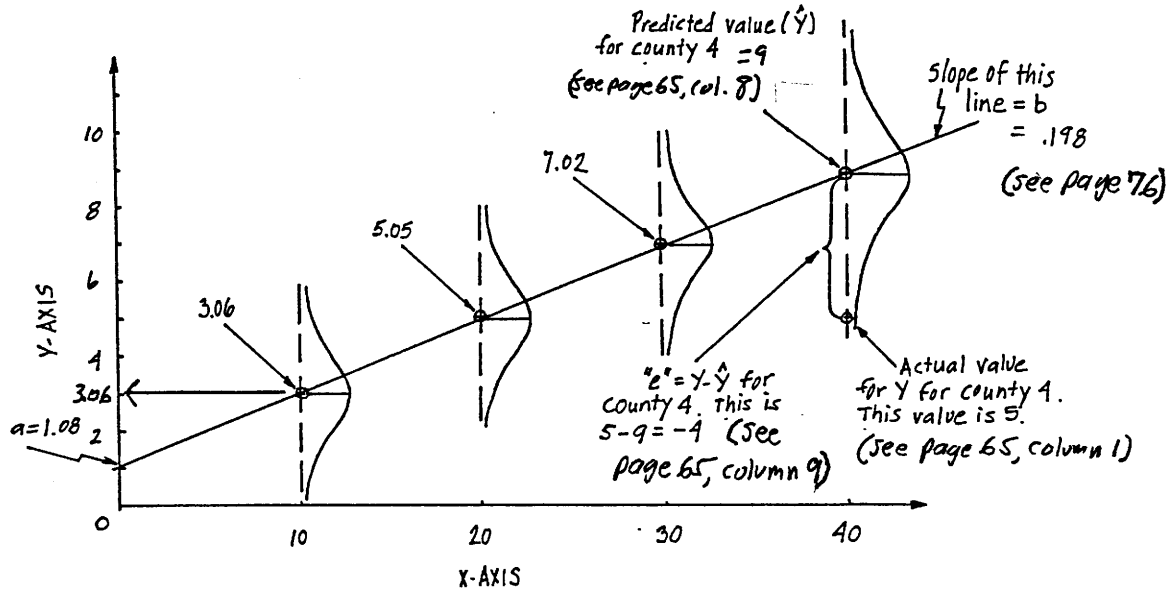
$$e = Y - (a + bX)$$

first entry on p. 65, col. 1 → $e = 2 - 3.07$ ← first entry in p. 65, col. 8



first entry in p. 65, col. 9

$$e = -1.07$$



Several points concerning the above diagram are important. First, think of the "mean" point of each of the normal curves as representing the "average" score on variable Y among a whole group of counties that all had the same score on variable X. For example, if we had a sample of 1000 counties (instead of 10) we would probably have 10-15 counties which each had a score on variable X of 10 (i.e., a youth unemployment rate of 10%). The results diagrammed above would suggest that the "average" (i.e., "mean") score on variable Y (of these 10-15 counties for whom X=10) would be 3.06%. Some counties would score higher, or lower, than 3.06% on Y, but the "average" (i.e., "mean") score on Y of these 10-15 counties (all scoring 10 on variable X) would be 3.06%. As you can tell from page 65, column 1, county #1 (the only county in our data set that scored 10 on X) had a different score on Y than 3.06% (i.e., "2"). If we had many counties that each scored 10 on variable X, most of them would probably score something other than 3.06% on Y. However, the "average" score on variable Y of all the counties scoring 10 on variable X would be 3.06%. Second, think of "b" (i.e., the slope, or the regression line) as the "line of all predictions." Thus, every predicted value for Y (i.e., every value of Y) will be on the "b" line (or regression line). Look at the mean values under the normal curves where X = 20 (5.05), X = 30 (7.02) and X = 40 (9.00). Aren't those the same values that appear on page 65, column 8 (i.e., the $\hat{Y}$ column) for the counties where X = 20 (counties #2 and #3), X = 30 (counties #5 and #6), and X = 40 (county #4)? Yes!!

For the quizzes, review: (1) page 69 on the calculation, interpretation and weakness of covariance; (2) page 75 on the interpretation and weakness of correlation; (3) pages 76-77 on the meaning and interpretation of "b"; pages 78-79 on the meaning and interpretation of "a"; pages 79-80 on the meaning of "e".

Assessing How Well Our Model Works - Symbolized by R^2

One of the goals of any science is to explain why something occurs (see page 4). In our case we are attempting to explain variation in county delinquency rates (variable Y). In other words, we are trying to explain why some counties have a lower than average (mean) level of delinquency and other counties have a higher than average level of delinquency. For example, let us look at county #10. According to the data on page 65, column 1, the delinquency rate in county #10 is 17%. This rate is 8% higher the mean rate of 9% (on page 66 the mean value of Y is calculated to be 9%). Why did county #10 have such a high delinquency rate? Our hypothesis is that this occurred because the youth unemployment rate (score on variable X) in county #10 (60%) was so much higher than the average (mean) youth unemployment rate (40% - see page 66 on the mean of X).

The above reasoning suggests the following question. What percentage of the variation in county delinquency rates (i.e., all counties do not have the same delinquency rate) can be attributed to variation (i.e., difference) in county youth unemployment rates? The higher the percentage of the variation in county delinquency rates that can be attributed to variation in county youth unemployment rates, the more successful our model (i.e., the greater the support for our hypothesis). While "b" (.198 - see p. 76) tells us how much delinquency increases as youth unemployment increases (on average, each one percentage point increase in youth unemployment is associated, on average with almost .2 percentage point increase in delinquency), it does not tell us what percentage of the variation in county delinquency rates is explained by variation in county youth unemployment rates. Instead, "b" tells us the slope of the relationship between X and Y, but not how closely the points are to whatever slope is drawn. Review the diagram on page 37. The diagram on page 37 illustrates the difference between the slope of a line and how closely the points are to a particular line. For example, the slope could be steep, but the points might not lie close to the line (i.e., diagram on page 71). Alternatively, the slope could be relatively flat, but the points could cluster very closely to the line (e.g., line "B" on page 37). Thus, the steepness of the line is unrelated to how close the points are to the line. Basically what we now need is something similar to correlation (i.e., the "strength" of the relationship between X and Y). Remember that correlation (pages 37, 75) does assess the closeness of the points to the line. However, correlation does not tell us the percentage of the variation in Y which is explained by X.

Let us now formulate an answer to the question of what percentage of the variation in county delinquency rates is explained by variation in county youth unemployment rates. If we tried to assess how well the Dodgers were playing during the season we might look to see how many games they had won. If they had won 20 games, what would we decide? I think we would want to know how many games they had played. For example, if they won only 20 out of 100 games, this would probably not be thought of as "successful." On the other hand, 20 wins in 20 games would be as successful as they could have been. We might think of it this way: how well did they do in relation to how well they could have done? Applied to our situation, this question would be how much of the variation in Y did we explain relative to how much variation in Y there was to explain.

Let us first calculate how much variation in Y there was to explain. Since variation is generally measured in terms of a deviation from the mean (as in "standard deviation"), wouldn't the answer to our question be found by subtracting the mean score on Y from each individual score on Y, squaring this difference, repeating this process for each observation and then adding up all of these squared deviations? Yes! Page 65, column 3, shows us how each score on Y deviated from the mean score on Y. Page 65, column 4, squares each of the values in column 3. Remember, if you just added the deviation without first squaring, you would end up with zero (note the total at the bottom of page 65, column 3). This would lead us to the erroneous conclusion that there was no variation on Y (i.e., the mean score on Y occurred because every county scored the mean value). Therefore, we square before we sum. The sum (total) at the bottom of page 65, column 4 (216), is the total amount of variation in Y (i.e., the largest amount of variation that could be explained by variation in X). The total amount of variation in Y is also referred to as TSS (total sum of squared deviations). Alternatively, 216 is "how well we could have done."

Now we need to determine "how well we did." Our theory says that scores on variable X are suppose to influence scores on variable Y. But suppose we did not have scores on variable X. If not, what delinquency rate would we logically predict for each of our 10 counties? As it turns out, in the absence of additional knowledge (i.e., not having scores on a logically related independent variable), our best "guess" would be to predict that each county would score the mean value (i.e., 9%) on variable Y. Looking at page 65, column 1, we can see that predicting each county would have a 9% delinquency rate would be rather inaccurate. Many of the scores in column 1 are quite far from 9%. Nevertheless, if we predicted the same score for each county, and this score was other than the mean value (i.e., other than 9%), we would produce larger prediction errors than simply predicting that each county would score the mean value on variable Y. This suggests a useful test.

Does knowledge of the scores on variable X lead us to make more accurate (i.e., less inaccurate) predictions about scores on variable Y than if we predicted the mean value score on Y for each county? Let's see. For example, as we noted previously, county #10 has a delinquency rate of 17% (see page 65, last score in column 1). If we did not know the youth unemployment rate in county #10 we would have predicted that it would have had a delinquency rate of 9% (i.e., the mean score on variable Y). Obviously, had we predicted a score for county #10 of 9% on variable Y we would have made an "error" of 8% (17% - 9% = 8%). However, because we knew the youth unemployment rate in county #10 was 60% (last score on page 65, column 2), we changed our prediction from the mean value on variable Y (i.e., 9%) to 12.95% (see the last entry on page 65, column 8). Remember the predicted value of $Y = a + bX$ (see page 78). This prediction of 12.95% comes from our knowledge of "a" (1.08 - page 78), "b" (.198 - see p. 76), and the score on X (county #10 has a score on X of 60%). Thus, $12.95 = 1.08 + \{(.198)(60)\}$. While our prediction is incorrect (12.95% does not equal 17%), it is closer to the actual value of Y (17%) than is the mean value of Y (i.e., 9%).

Doesn't variable "X" deserve the credit for reducing our prediction error from 8% (17% - 9% = 8%) to 4.05% (17% - 12.95% = 4.05%). Yes! Page 65, column 11, shows the difference between our

predicted value on Y given (i.e., utilizing) knowledge of variable X (for county #10 this is 12.95) and the mean value of variable Y (9). The entry for county #10 in column 11 is 3.95 (12.95 - 9 = 3.95). Page 65, column 11, shows the reduction in prediction errors resulting from using each county's score on variable X to predict its score on variable Y as opposed to predicting the mean value on variable Y for that same county. Do not be concerned with whether the value in column 11 is "positive" or "negative." Page 65, column 12, takes each entry in column 11 and squares it [for county #10: (3.95) (3.95) = 15.60]. The total of these "squared prediction improvements" is called either the regression sum of squares (RSS) or the "explained" sum of squares. As we see from the bottom of page 65, column 12, this total is 163.8. This is "how well we did."

Now we can answer our basic question, "How well did we do in relation to how well we could have done"? The answer would be to divide RSS by TSS. If we do this we get an answer of .76 (163.8/216 = .76). This result tells us the proportion of variation in variable Y which is explained by variable X. If we multiply .76 by 100, we can then say that approximately 76% of the variation in variable Y is "explained" by variation in variable X. The percentage of variation in variable Y explained by variation in variable X is called R^2. Make sure you learn this interpretation of R^2. The quizzes are highly likely to ask you to give a percentage interpretation of R^2. Just so there isn't confusion, if the quiz says to interpret $R^2 = .76$ you should say that 76% of the variation in variable Y is explained by variation in variable X. Make sure you remember the following: explaining variation in the dependent variable is not the same as perfectly predicting the score on the dependent variable. Column 9 on page 65 shows the difference between the predicted score on Y and the actual score on Y for each county. If we perfectly predict the score on variable Y for a county, the entry in column 9 on page 65 for that county will be zero (i.e., there would be no difference between the predicted and actual values). Notice that our model does not perfectly predict any of the scores on Y (i.e., there is some non-zero number in each cell in column 9 on page 65). What the R^2 of .76 tells us is that knowledge of variable X reduces the squared prediction errors that we would make by just predicting the mean score on Y (i.e., 9) for each county by 76%. However, this is a statistical explanation which is only as good as our theory. Other factors besides youth unemployment no doubt effect county juvenile delinquency rates (e.g., the presence of a gang prevention program in a county - hence the need for "multiple" regression - i.e., more than one independent variable - which will be discussed soon).

Not surprisingly, the difference between the actual value of variable Y and the predicted value of Y based upon knowledge of variable X is referred to as "error" (i.e., these are the values for "e"). Page 65, column 9, shows the prediction errors for each county. Column 10 contains the squared values of each entry in column 9. The total squared errors is referred to as the error, or "unexplained" sum of squared deviations (ESS). The total at the bottom of column 10 on page 65 tells us that we have 51.96 squared units of error. It is worth mentioning that "RSS" (the explained or regression sum of squares) plus "ESS" (the sum of squared prediction errors) must equal "TSS" (the total sum of squares). If we add what we can explain (RSS) to what we can not explain (ESS) it should equal all that there was to explain (TSS). You can see that this relationship does hold [163.8 (which is RSS) + 51.96 (which is ESS) = 216 (which is TSS)]. The figures would exactly equal except for differences due to rounding.

The Least Squared Errors Principle

Since we place so much emphasis on "b," it is important to understand why the formula for "b" reads as it does. From past discussion (page 78) we know that in regression there is a predicted value for Y (i.e., $\hat{Y}$) for each observation (e.g., each county in the data set on page 65). We also know that "b" is an important component of the formula for predicting the value of each observation on Y (page 78). Remember from page 78 that the formula for predicting the value of Y is:

$$\textit{Predicted Value of } Y = \hat{Y} = a + bX$$

Furthermore, "b" is also a part of the formula for "a." Remember from page 78 that:

$$a = \overline{Y} - b\overline{X}$$

Thus, "b" has a very important influence on $\hat{Y}$. Therefore, "b" has several important functions. First, it tells us how many units Y is expected to either increase (if "b" is positive) or decrease (if "b" is negative) if X increases by one unit (see pages 76-77).

Second, "b" influences the predicted value for Y (i.e., $\hat{Y}$) for each observation in our study (e.g., each of the 10 counties in the data set on page 65). Remember from pages 79-80, that the predicted value of Y is a critically important part of the following formula for the "residual" or "error term" (i.e., "e"):

$$\textit{Residual} = e = Y - \hat{Y}$$

The "residual" (i.e., "e") is important because the nature of the prediction errors we make may well tell us something valuable about the process we are trying to model. For example, suppose that in our county delinquency study (page 65), we typically have larger prediction errors (i.e., larger values for "e") for counties that do not have gang prevention programs. This should cause us to re-think our model and add a second independent variable (to measure gang prevention programs). We will learn how to do this later (pages 106-114). The important point is that the nature of the prediction errors help us think about our model. So, since "b" tells us how much Y changes as X changes (see pages 76-77) and how large the prediction error will be for each observation (see above), calculating "b" becomes very important.

From page 76, we know the formula for "b" is as follows:

$$b = \frac{Cov(X,Y)}{Var(X)}$$

Rather than discuss the mathematics behind the formula for "b", my interest is in explaining what criteria "b" is suppose to meet. Put another way, what is the goal in calculating of "b"? The next few sentences may be confusing. The following discussion will make sense if you just keep reading through page 87.

The formula that the computer executes in order to obtain the least squared errors is the formula for "b" above. That formula was determined through the use of calculus. The interesting part of the derivation of "b" is not the calculus behind the formula, but rather deciding what criteria the formula for "b" should meet.

To explain this, let us examine the following two panels of data. Reading across the columns in panels 1 and 2 below: column 1 contains the actual scores on variable Y; column 2 contains the predicted values (i.e., scores) for Y (i.e., $\hat{Y}$ - which we know from page 85 are based, in part, on the value of "b"); column 3 contains the values for "e" (each "e" is just the score in column 1 minus the corresponding score in column 2); column 4 contains the values of e^2 (each "e^2" is just the corresponding score in column 3 multiplied times itself). Each panel below has 2 "observations" (e.g., data on 2 counties or an "N" of "2"; we had 10 counties - or 10 observations - in the data set on page 65).

Panel #1

	Column 1 Y	Column 2 $\hat{Y}$	Column 3 $e = (Y - \hat{Y})$	Column 4 $e^2 = [(e)(e)]$
Observation #1	5	5	0 = (5 - 5)	0 = [(0)(0)]
Observation #2	7	0	7 = (7 - 0)	49 = [(7)(7)]
			Σ =7 = (0 + 7)	Σ =49 = (0 + 49)

Panel #2

	Column 1 Y	Column 2 $\hat{Y}$	Column 3 $e = (Y - \hat{Y})$	Column 4 $e^2 = [(e)(e)]$
Observation #1	5	3	2 = (5 - 3)	4 = [(2)(2)]
Observation #2	7	1	6 = (7 - 1)	36 = [(6)(6)]
			Σ =8 = (2 + 6)	Σ =40 = (4 + 36)

In both panels on page 86, the scores for variable Y in column #1 are the same. In each panel, the score on variable Y for observation #1 is 5 and for observation #2 is 7 (see column #1 in both panels). The differences in the two panels are entirely the result of column #2. For panel #1, the predicted score for observation #1 on variable Y is 5 (see panel #1, column #2). For panel #2, the predicted score for observation #1 on variable Y is 3 (see panel #2, column #2). Since the actual values on Y are the same in each panel (again, see column #1 of each panel) and the predicted values for observation #1 in each panel are different, the value of the "error" term (column #3 - which is the difference between columns #1 and #2) and the square of the "error" term (column #4 - which is column #3 times itself) will be different in each panel. The same pattern holds for observation #2.

Since the actual values of Y are the same and panels #1 and #2 are each trying to predict the scores on Y, which panel does the better job? A reasonable goal for any model that makes predictions would be to minimize error. Turn the question around, why would someone want to maximize error? In panel #1, the total sum of errors is 7 (see the bottom of column 3 of panel #1). In panel #2, the total sum of errors is 8 (see page 86, bottom of column 3 of panel #2). Therefore, if we want to "minimize" error, panel #1 is the better choice (because 7 is less than 8).

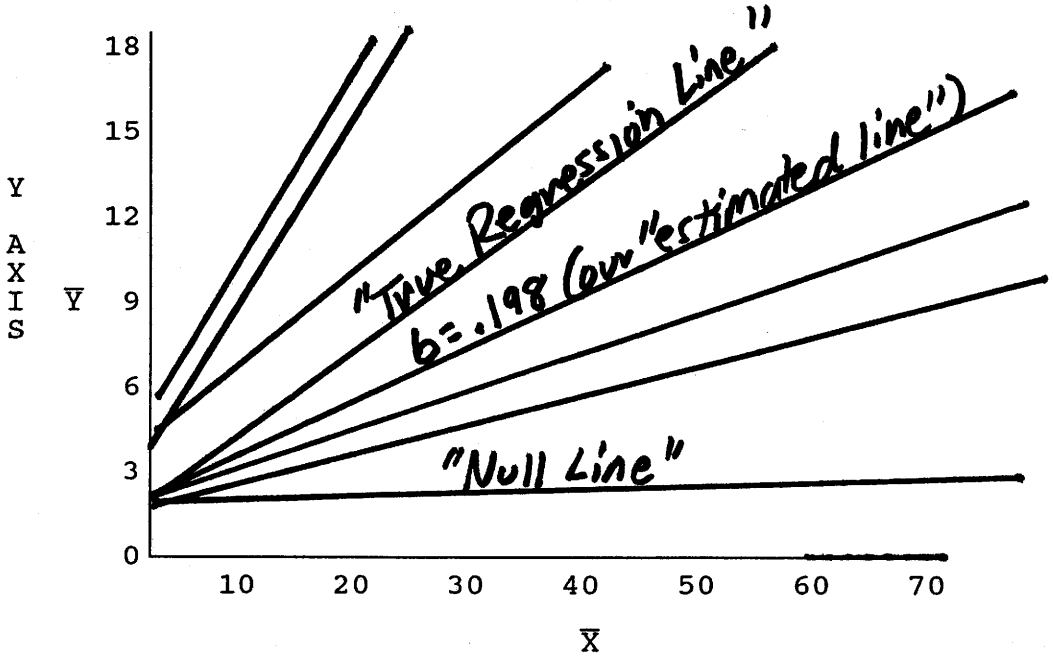
While minimizing the sum of errors seems like a worthy goal, it is not the goal that political scientists use. If we use the sum of total errors, we do not care how large any single prediction error is, only the total of all the prediction errors. One could make a compelling case that a preferable criterion would be to minimize large prediction errors. For example, two predictions errors of 3 each could be thought of as more desirable than one prediction error of 4 and one error of 2. In each case the total prediction error is 6 (3 + 3 = 6 and 4 + 2 = 6). The error of 4 is simply larger than the other errors. This "larger" error of 4 will produce a greater (i.e., less desirable) squared error total: two prediction errors of 3 each produce a "squared" error total of 18 ($3^2 = 9$; two errors of "9" each total 18) but one error of 4 and one error of 2 produce a "squared" error total of 20 ($2^2 = 4$; $4^2 = 16$ and 4 + 16 = 20). Obviously, 20 is greater than 18.

Therefore, one could argue that the goal should be to minimize the largest errors as opposed to just the total error. The way that we do this is to minimize the total "squared" errors. Look at the bottom of column 4 on page 86 for both panels #1 and #2. The total of squared errors is lower in panel #2 (40) than in panel #1 (49). If we use the lower "total" prediction error criterion, panel #1 is preferable to panel #2 (because 7 is less than 8 - see column #3). However, if we use the lower total "squared" prediction errors criterion, panel #2 is preferable to panel #1 (because 40 is less than 49 - see column #4). This difference is because panel #1 has one large prediction error (7) which when squared produces a higher squared error total (49) than panel #2 (40).

Since political scientists want to minimize large prediction errors they use the least squared errors criterion in choosing a formula for "b." Thus, political scientists would choose the value for "b" used in panel #2 instead of panel #1. Returning to page 76, if "b" were any value other than .198, then the sum of squared errors at the bottom of column 10 on page 65 would have been greater than 51.96 (i.e., 51.96 is the lowest possible total of squared prediction errors when using scores on X to predict scores on Y for the data on page 65). Using calculus to answer the question: For the lowest total of squared prediction errors what formula for "b" should we use? The answer is the formula for "b" on page 86. The least squared errors method is called "Ordinary Least Squares" (or just "OLS").

Testing for the Statistical Significance of "b" - The "t ratio"

Keep in mind that the value for "b" which we calculated on page 76 (.198) is only an "estimate" of "b." Since we do not know the "true" value of "b," it is important to determine how much confidence we can place in our estimate of "b". This is the "fundamental question of statistical inference" (page 40). The question is, How likely is the result (i.e., "b" = .198) the product of chance? This question could be rephrased as follows: How likely would we estimate the value of "b" in our sample to be .198 when the "true" (or "population") value of "b" is .000? Remember that if the "true" value of "b" were .000 then X would have no effect on Y. That is, if X increased by one unit, on average, Y would neither increase nor decrease. The diagram below puts these ideas in a visual context.



X AXIS

Please note several points about the diagram immediately above. First, there is a "true" regression line that expresses the actual relationship between X and Y. Remember that we only have data from 10 counties at one point in time, not all counties at all points in time. Therefore, we will never know the value of the "true" regression line. The regression line that is designated as the "true" regression line is just a representation to let you know that there is one "true" regression line. Second, the slope of the regression line that was estimated from our sample, .198, is our best estimate of the "true" regression line. However, keep in mind that it is only one estimate of the "true" regression line. If we draw 5 more samples of 10 counties each and estimate a value from each of these samples, we would obtain five additional estimates of "b." These five estimates of "b" might distribute themselves around the "true" value for "b" (if we could know what value that was) and the estimate of "b" from page 76 as in the diagram above. Finally, note the "null" line. If X has no relationship to Y, the

"null" line would also be the "true" regression line. This is because as scores on X increase (e.g., from 30 to 40 to 50, etc.) the predicted score on Y remains the same (in the diagram on page 88 the "null" line is drawn, and never varies, from approximately 2.75 on the Y axis). Hence, increasing the score on variable X has no effect on variable Y. Put another way: the "null" line is parallel to the X axis. Think of the "fundamental question of statistical inference" this way: How likely would the "null" line be the same value as the "true" regression line if the value of "b" in our sample was .198 (i.e., the estimated value for "b" from page 76)? If this is confusing, just keep reading, it will make sense soon!!

The ensuing discussion helps us answer the "fundamental question of statistical inference" for our estimate of "b" (.198). Just keep reading, it will become clearer! Previously, we used the standard deviation to tell us the percentage distribution of scores around the mean for a normal distribution (e.g., 95% of the scores are within 2 standard deviations of the mean - see page 24). We will use this information to assess the statistical significance of our estimate of "b." However, instead of the mean we will use our estimate of "b" (.198) and instead of the standard deviation we will use the standard error of "b." The computation for the standard error of "b" is as follows:

$$\textit{Standard error of } b = S_b = \sqrt{\frac{ESS/N-2}{\textit{Variation in } X}} = \sqrt{\frac{51.96/10-2}{4200}} = \sqrt{\frac{6.495}{4200}} = \sqrt{.001546} = .039$$

see page 65, bottom column 10

see page 65, bottom column 6

To test for the statistical significance of "b" we use the ratio of our estimate of "b" (.198) to the standard error of "b" (.039). This ratio is called the "t ratio." Although we will use the "t" distribution (hence "t ratio") instead of the "normal" distribution, they are quite similar (i.e., for both distributions there is approximately a 95% chance that the true value of "b" is within 2 standard errors of our estimate of "b"). Just keep reading, this will become clearer over the next several pages. The "t ratio" for our data is as follows:

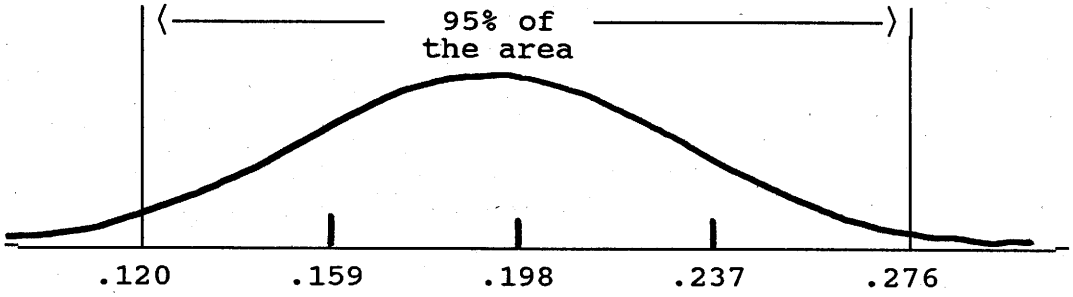
$$t \textit{ ratio} = \frac{\textit{coefficient } b}{\textit{standard error of } b} = \frac{.198}{.039} = 5.08$$

see page 76

see above

As will be shown on the next page, the "t ratio" must have an absolute value (does not matter whether it is positive or negative) of 2.0, or greater, for our estimate of "b" to be statistically significant at the .05 level. Just keep reading!! As our "t ratio" of 5.08 exceeds 2.0, we can reject the null hypothesis (that "b" equals .000) in favor of the alternative hypothesis (that "b" does not equal .000) with less than a 5% (i.e., .05) probability of committing a "type I" error (i.e., rejecting the null hypothesis when the null hypothesis is actually true). Just keep reading, this will be made clearer over the next several pages.

This paragraph is absolutely critical. If you get confused, just keep reading through the rest of the paragraph. Remember from our discussion of the normal curve that approximately 95% of the area under the normal curve lies within plus or minus two standard deviations of the mean (see page 24). To apply this information to our situation, we replace the mean with our estimate of "b" (.198) and the standard deviation with the standard error of the coefficient "b" (.039). We have approximately a 95% chance (probability) that the "true" value of "b" will lie between .120 (.198 - .039 - .039 = .120) and .276 (.198 + .039 + .039 = .276). The diagram below graphically shows this.



Think of it this way: if we replicated (i.e., "repeated") our study 100 times (i.e., a different randomly selected group of ten counties in each replication) we would obtain 100 estimates of "b." Approximately 95% of these estimates of "b" would lie between .120 and .276. Importantly, zero (.000) is not within this 95% probability region (or "zone") for "b." However, the "true" value for "b" could still be zero (.000), it is just highly unlikely (less than a 5% probability) given our results. While we may not be able to replicate (repeat) our study, the results we found on pages 41-44 (where we could replicate a study, i.e., keep drawing samples), allows us to make the assumptions of the preceding paragraph.

Our decision rule is this: if the absolute value (i.e., either positive or negative) of the "t ratio," which is "b" divided by the standard error of "b," is less than 2.0 (i.e., if "b" is less than twice the size of its own standard error) we retain the null hypothesis and if the "t ratio" is 2.0, or greater, in absolute value (either positive or negative), we reject the null hypothesis because there is less than a 5% chance that the null hypothesis is true. This 2.0 threshold for the ratio of the coefficient ("b") to the standard error of "b" is the one formula you need to memorize. Note that having 95% confidence (or a 5% chance of committing a type I error) does not mean that our estimate of "b" (.198) is the "true," or correct, estimate of "b" 95% of the time. It just means that we have a 95% probability of not committing a type I error.

Make sure you know how to interpret the "t ratio." I often give quizzes that ask you to interpret regression results. No matter what, just keep reading through this paragraph! A typical example would be the following: b = -.379, standard error of "b" = .160. How should you interpret this? I would expect you to say that if X increased by one unit, on average, Y would be expected to decrease by almost four-tenths of a unit. If I give you the specific units of measure (e.g., Y = dollars of income) I expect you to use them (review pages 76-77). The standard error (.160 in this example) is only interpreted relative to the

value of "b." Since "b" (-.379) has an absolute value of over 2 times the size of the standard error of "b" (i.e., -.379 is over twice the size of .160), then we would reject the null hypothesis because there is less than a 5% chance that the null hypothesis (i.e., that "b" = .000) is actually true. That is, we reject the null hypothesis since the null hypothesis has less than a 5% chance of being true. We would say that "b" is statistically significant at the .05 level. The "level of significance" is equal to the probability of committing a "type I" error. Thus, if "b" is statistically significant at the .05 level, it means that 5% or less of the time that we reject the null hypothesis the null hypothesis is actually true. So, we reject the null hypothesis and run a 5% or less chance that we were wrong. However, if "b" = .579 and the standard error of "b" = .357, what would you say? After interpreting "b" I would hope you would say that we would not reject the null hypothesis because .579 is less than twice the size of .357. Thus, we would say that "b" was not statistically significant at the .05 level. Alternatively, we would say that "b" was statistically insignificant.

Pages 62-63 discussed two properties of every significance test. The first property was the magnitude of the effect and the second property was the sample size. Just keep reading!!! Let us see how these two properties are reflected in the "t ratio." Since the standard error of "b" is the denominator of the "t ratio" formula (i.e., the "t ratio" is "b" divided by the standard error of "b"), let us see if either of the two properties mentioned above are reflected in the formula for the standard error of "b."

$$\text{Standard error of } b = S_b = \sqrt{\frac{ESS/N\text{-}2}{\text{Variation in } X}} = \sqrt{\frac{51.96/10\text{-}2}{4200}} = \sqrt{\frac{6.495}{4200}} = \sqrt{.001546} = .039$$

see page 65, bottom column 10

see page 65, bottom column 6

Notice the "N-2" term in the numerator of the formula for the standard error. Let us see what the "N-2" term means in the context of the data set we are working with. As you look at the middle of the above formula, you can see that the numerator of the formula is ESS/N-2 which becomes 51.96/10-2 = 51.96/8 = 6.495. If, instead of "10" we had 50 counties, the numerator would have been 51.96/50-2 = 51.96/48 = 1.08. So, if N = 10, the numerator of the standard error is 6.495 and if N = 50, the numerator is 1.08. Now, look at the denominator of the standard error formula immediately above. Doesn't it work out to be 4200? Yes! If you divide 6.495 by 4200 you get .001546. If you divide 1.08 by 4200 you get .000257. While you may need to squint to tell the difference, .000257 is smaller than .001546! So what!! Actually, this is important. The larger the size of "N" (i.e., an "N" of 50 is obviously larger than an "N" of 10) the smaller the numerator in the standard error formula. The smaller the size of the numerator of the standard error, the smaller the standard error will be. Think for a second why this is so. If the numerator of the standard error was 5 and the denominator was 2, the standard error would be 2.5 (i.e., 5/2 = 2.5). Alternatively, if the denominator remained at 2, but the numerator decreased from 5 to 3, then the standard error would be only 1.5 (i.e., 3/2 = 1.5 which is smaller than 2.5). So, all other factors being equal, a larger "N" results in a smaller standard error. Now put this in the context of the "t ratio" formula which appears ahead.

see page 76

$$t\ ratio = \frac{coefficient\ b}{standard\ error\ of\ b} = \frac{.198}{.039} = 5.08$$

see page 89

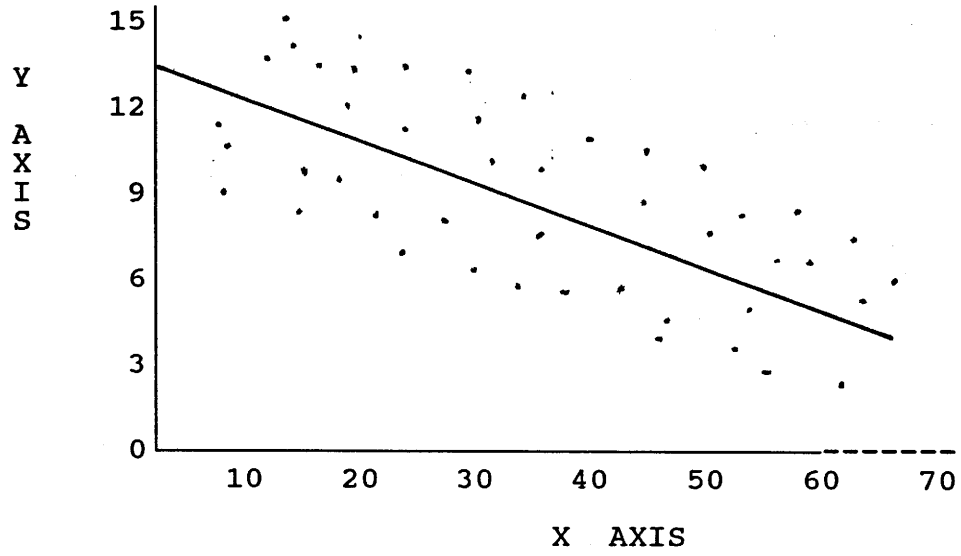
Since the standard error of "b" is the denominator in the "t ratio" formula, all other factors being equal, the smaller the standard error of "b," the larger the value of the "t ratio." For example, if the numerator of the "t ratio" formula is .198 (as above) and the denominator is .039 (as above) we can see that the value of the "t ratio" is 5.08 (i.e., .198/.039 = 5.08). However, suppose that the numerator of .198 remained the same but the denominator decreased to only .020 (instead of .039). The "t ratio" would now become .198/.020 = 9.9. Obviously, a "t ratio" of 9.9 is greater than a "t ratio" of 5.08. While both 9.9 and 5.08 are greater than the 2.0 threshold that was discussed previously, it is true that a "t ratio" of 9.9 is more statistically significant than a "t ratio" of 5.08. For example, the second principle of significance tests was the larger the sample size the more statistically significant the result (see page 63). The example we have just worked through shows that this principle does apply to the "t ratio."

The first principle of significance tests is that the greater the magnitude of the relationship between X and Y, the more statistically significant the result (see page 62). Let us see if this principle holds for the "t ratio." We know that "b" represents the magnitude of the relationship between X and Y. As is shown in the "t ratio" at the top of this page, the value of "b" is .198. If the denominator of the "t ratio" above remained at .039, but the value of "b" (i.e., the numerator) increased from .198 to .298, what would happen to the value of the "t ratio"? Come on!! You know!! Just as you thought, if the value of the denominator (.039) remained the same but the value of "b" increased from .198 to .298 the value of the "t ratio" would have to increase. In this instance, the value of the "t ratio" increases from 5.08 (.198/.039 = 5.08) to 7.64 (.298/.039 = 7.64). Although both 7.64 and 5.08 are above the 2.0 threshold, it is true that a "t ratio" of 7.64 is more statistically significant than a "t ratio" of 5.08. Thus, the first principle of significance tests does apply to the "t ratio."

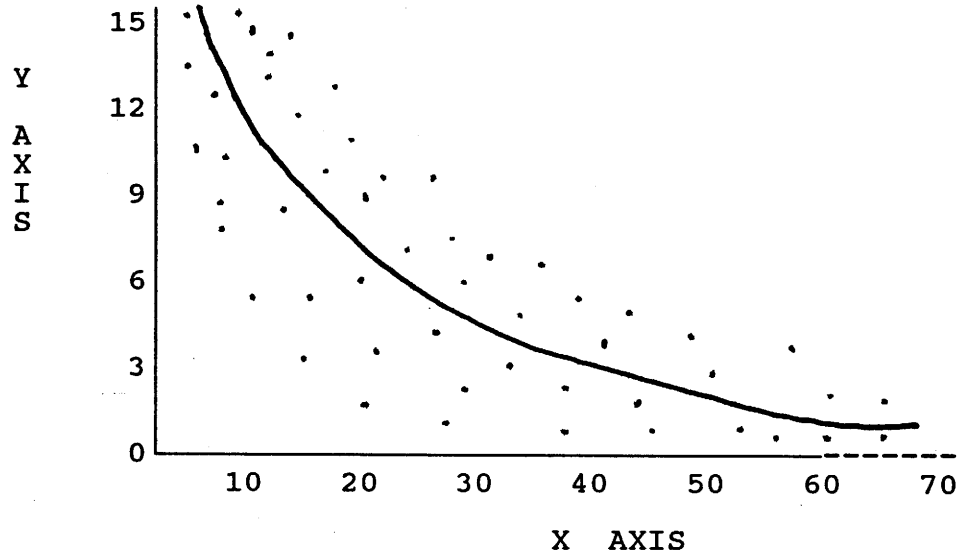
When reading the assignment it is always useful to think about upcoming quizzes. For example, how would you interpret an R^2 of .39? What is the least squared errors principle and why do political scientists use it? If "a" = 75, "b" = -.588 and the standard error of b = .200, what statements could you make? (Note that the t ratio is -.588/.200. What does this tell you?) Don't just say something is "statistically significant" because I don't know that you know what "statistically significant" means. If a relationship is significant at the .05 level, what does this mean? What happens 5% of the time? As you read pages 93-100, try to focus on both the method and logic of a logarithmic variable transformation. Thus, be able to answer in words (not with a diagram), what a logarithmic variable transformation is and why a political scientist might use one. Finally, when reading pages 93-100, pay attention to the discussion of "rates." Thus, why would a political scientist use an injury "rate" instead of the number of injuries?

Non-Linear Relationships

One of the basic assumptions of regression is that the relationship between X and Y is best represented by a straight line. A "straight line" relationship is also called a "linear" relationship. In a linear relationship the amount of increase or decrease in Y for a one unit increase in X is the same regardless of the score on X. The scatter plot immediately below depicts a linear negative relationship between X and Y.



A non-linear relationship between X and Y occurs if the regression line that best represents the relationship between X and Y is not a straight line. The line immediately below depicts a negative non-linear relationship between X and Y.



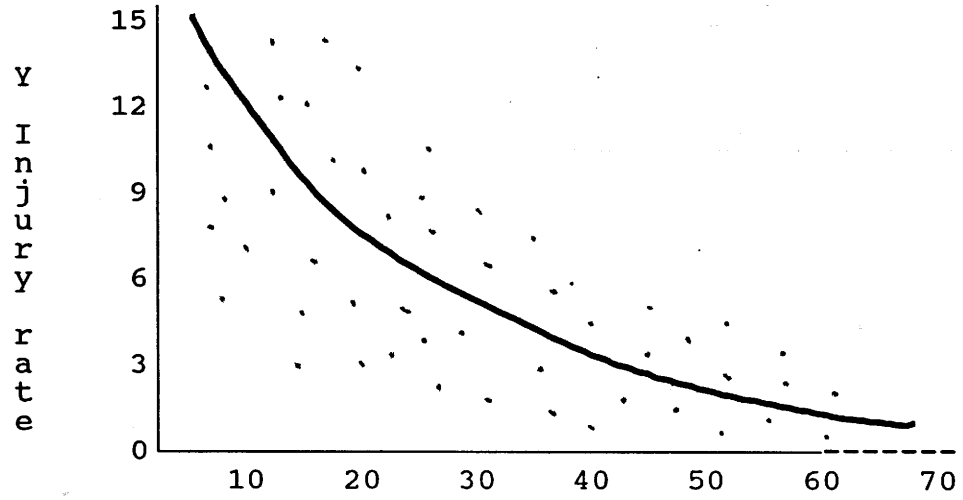
Notice the line in the scatter plot immediately above indicates that when X increases from 10 to 20 Y decreases by more that when X increases from 40 to 50. Since Y decreases by different amounts as X increases by the same amount, the relationship between X and Y is non-linear.

Political scientists frequently encounter situations where the relationship between X and Y is non-linear. For example, suppose you are a political scientist studying to what extent the federal government can regulate safety in the workplace. This is an important topic. Every year workers are injured, or killed, on the job. Every year the federal government spends millions of dollars trying to reduce work related injuries. How successful are such efforts? If health is the goal, could the money spent by the federal government on regulating the workplace be better spent on something else (e.g., medical research)? These are important questions that political scientists have attempted to answer. The example ahead provides a step in this direction. The example was inspired by Michael S. Lewis-Beck, Applied Regression: An Introduction, pages 43-47. To avoid reprinting fees, the following example is different from the one reported by Lewis-Beck. However, while the example uses a different regulatory policy area and the results are hypothetical, the results in the example below are entirely consistent with what Lewis-Beck found. Thus, the regulatory "picture" that emerges below is consistent with what actual research in a different regulatory policy area has found.

The Occupational Safety and Health Administration (OSHA) is a federal agency whose primary purpose is to regulate workplace safety. One of the prime functions of OSHA is to inspect the safety of private and public worksites. In this example, variable X is the amount of money that the federal government spends annually (i.e., per year) on OSHA inspections. Variable X is measured in millions of dollars. This means that a score of 1.0 on X would mean that the federal government spent $1,000,000 on OSHA inspections in that particular year. The scores on X are adjusted for inflation (so one dollar in 1996 has the same value as one dollar in 1986). Adjusting data for inflation is often referred to as using "constant" dollars (i.e., a dollar has a constant, or the same, value over time). Variable Y is the annual rate of serious injuries per million hours worked in the United States. Thus, a score of 1.0 on Y would mean that there was one serious injury for every one million hours worked in the United States in that particular year. We have one score on variable X and one score on variable Y for each year over a 40 year period (e.g., 1957 to 1996; hence "N" = 40).

At this point, let me raise a "measurement question." Why not measure Y as the number of workers seriously injured in that year? Why use a seriously injured "rate" (i.e., the number of workers seriously injured per million hours worked)? The answer is that using the number of workers seriously injured in a given year would not adjust for the number of hours that were worked in a year. For example, let us say that in 1995 the United States was in a recession. That wasn't the case, but for sake of discussion, let us say that 1995 was a recessionary year. One of the manifestations of a recession is increased unemployment. If more people are unemployed, then fewer people are working. If fewer people are working, the number of hours worked would decrease. If the number of hours worked decreased we should expect fewer serious injuries. Therefore, if we used the number of serious injuries as our measure for Y, the variable would show a decrease in serious injuries each time we went into a recession. Such a decline would not mean that the workplace was "safer," rather it would just reflect the fact that less work was done. Since we want to measure "safety" we need a measure which adjusts for the differing number of hours worked between any two years. Therefore a serious injury "rate" (such as the number of serious injuries per million hours worked) makes sense.

Looking at the scatter plot below, we can see that the relationship between X and Y is negative (i.e., higher scores on X are associated with lower scores on Y). We would hope that the relationship between X and Y is negative because this would mean that as the federal government increases spending on OSHA inspections (variable X) the rate of serious injuries (variable Y) decreases. Isn't the line downwardly sloping? Yes! This means the relationship between X and Y is negative.



Since the relationship between X and Y is negative, we know that as scores on X increase, scores on Y will decrease. In a negative linear relationship, the number of units that Y will decrease if X increases by one unit is the same regardless of the score on X. Thus, in the above scatter plot, if the relationship between X and Y is linear, when X increases from 10 to 20 Y should decrease by the same amount as when X increases from 40 to 50. Clearly, this is not what occurs. Y decreases much more if X increases from 10 to 20 than if X increases from 40 to 50. Therefore, the line in the scatter plot above is non-linear. As discussed on the previous page, regression assumes that the relationship between X and Y is linear. Clearly we have violated an assumption of regression. What should we do?

The situation above represents only one, of a series, of violations of regression assumptions that you will read about in this course. In the final analysis, we will comply with the regression assumption of linearity. So, ultimately, we will convert to a scoring mechanism which results in a linear relationship between X and Y. However, each time we violate a regression assumption, we should ask the following question: What does the violation of the regression assumption tell us about the process we are trying to explain? In this instance we are trying to explain why the workplace injury rate differs over the years. Thus, every year there is some workplace injury rate. Perhaps in one year 2 workers are seriously injured for every million hours worked. Perhaps in a different year 4 workers are seriously injured for every million hours worked. We are trying to explain why this difference in serious injury rates occurs. In other words, what was different when only 2 workers were seriously injured per million hours worked than when 4 workers were seriously injured per million hours worked? Our hypothesis is that OSHA

inspections were more numerous when only 2 workers were seriously injured per million hours worked than when 4 workers were seriously injured per million hours worked. As discussed on page 5, we are testing to see if, in fact, this is the case (i.e., is our hypothesis supported?). Also, we want to estimate how much each additional million dollars spent on OSHA inspections lowers the rate of serious injuries (i.e., the magnitude of the effect of OSHA inspections on the rate of serious injuries - again, see page 5).

When we test hypotheses, we should always be guided by theory. In this instance that means we should think through what the likely relationship is between the amount (i.e., level) of OSHA expenditures (variable X) and the rate of serious injuries (variable Y). We should expect the relationship between the amount of money the federal government spends on OSHA inspections and the rate of serious injuries to be non-linear. This is because the amount of federal money spent on OSHA inspections is probably a good measure of the number of visits OSHA inspectors make to each workplace. The first time an OSHA inspector visits a particular worksite in a particular year, the reduction in serious injuries is likely to be high. The inspector will note obvious violations of safety practices that management will have to correct. However, with each succeeding visit the OSHA inspector will have a harder time finding as many problems. This is why the fourth visit in a year will probably not save as many serious injuries as the first visit in a year. This is analogous to dieting. During the first week on a diet you may lose 5 pounds. However, during the third week you may only lose 2 pounds. It becomes more difficult to lose each additional pound. As I mentioned before, regression assumes that you continue to lose the same number of pounds each successive week on your diet. Since this is unlikely, a researcher should expect a non-linear relationship between the number of weeks on a diet and the number of pounds lost. A non-linear relationship should also occur between OSHA budgets and the rate of serious injuries at the workplace.

It is much easier for us to work with linear relationships than non-linear relationships. For example, it is much more straightforward to say that if X increases by one unit, on average Y will either increase (if a positive relationship) or decrease (if a negative relationship) by some constant amount (e.g., you will lose two pounds each week on a particular diet). If the relationship between X and Y is non-linear, we can not make such a statement.

Now let us think through what we can do about this. If ever larger increases in X (i.e., OSHA expenditures) are required to bring about the same rate of decrease in Y (thus the same rate of decline in the rate of serious injuries) wouldn't a logical response be to change the method by which scores on X are assigned so that they are recorded as having increased at a slower rate? Just keep reading!!

As you learned in mathematics courses (a long time ago!), we can represent 100 as 10^2 (10 times 10 = 100). Similarly, we can represent 1,000 as 10^3 (10 times 10 times 10 = 1,000). Wouldn't scores on X increase at a slower rate if an increase in X from 100 to 1,000 was listed instead as an increase from 2 to 3? Yes! A score of 1,000 is 10 times the size of a score of 100. However, a score of 3 is only 1.5 times a score of 2. Just have the computer read the exponents of 10 (i.e., 2 and 3) instead of the numbers in non-exponential form (i.e., 100 and 1,000) and you will greatly reduce the rate of increase in X. The only difference between what I just did and what we will ultimately do is that I used a "base"

of 10 (e.g., 10 to the second power, i.e., 10^2) and ultimately we will use a "base" of approximately 2.71828. The base of approximately 2.71828 is referred to as base "e" and is symbolized by the letters "ln." To overcome the nonlinearity of the diagram on page 95, we will change from the equation: $Y = a + bX + e$ to the equation: $Y = a + b(\ln X) + e$. The term "ln X" is a logarithmic term with base "e" (not to be confused with the error term "e"). First, notice that nothing has happened to "b." However, in the latter equation the computer reads "ln X" instead of X.

A logarithm is the power that a number designated as the "base" has to be raised to equal some other number. I know, that was a difficult sentence! Just keep reading!! If the number we want is 100, the logarithm in base "e" is whatever power 2.71828 has to be raised to equal 100. It turns out that 2.71828 to the 4.605 power equals 100 (or $2.71828^{4.605} = 100$). Thus, if the computer originally read a value (a score) on variable X of 100, it would now (after the logarithmic transformation in base "e") read this same score as 4.605.

Now, let us see what a logarithmic transformation of X would do to the non-linear relationship between X and Y that we saw on page 95. The two diagrams on the next page compare the relationship between X and Y when X is in millions of constant dollars spent on OSHA inspections and when X is in the natural logarithm of millions of dollars spent on OSHA inspections.

Figure 1

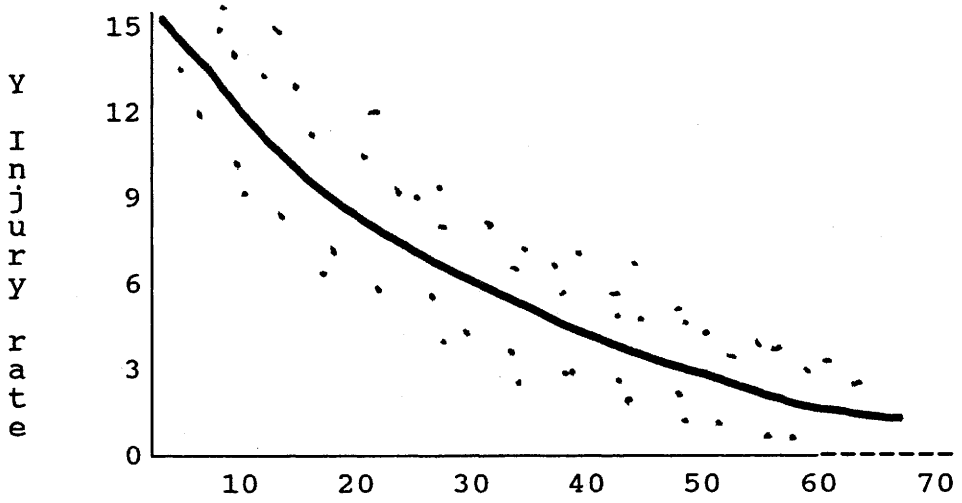
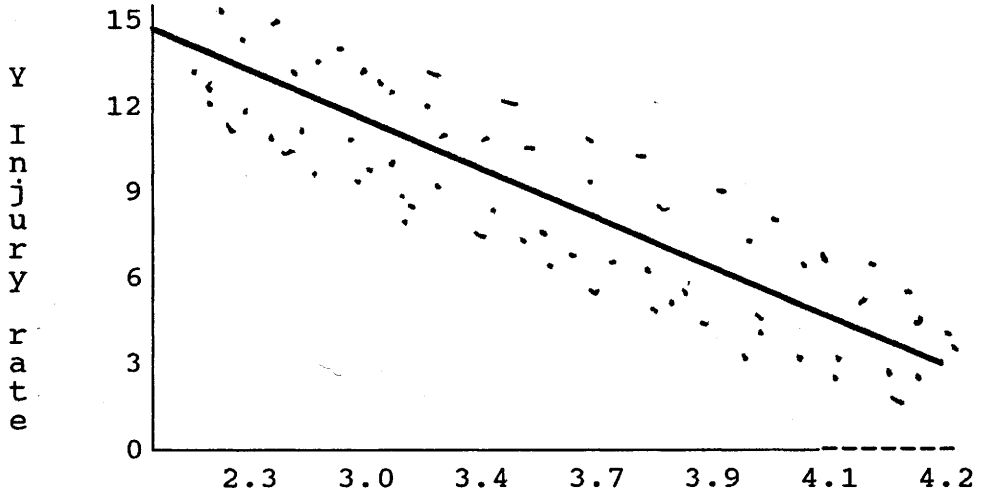


Figure 2



Where as the relationship between X and Y in Figure 1 is non-linear, the relationship between X and Y in Figure 2 is linear. Thus, if we transform variable X into natural logarithmic units a non-linear relationship can become linear. In Figure 2 above, the regression assumption of linearity has been met.

To see why Figure 2 is linear and Figure 1 is not, look at the relationship between scores on the X axis in the two diagrams. In Figure 1, the first score on the X axis is 10. In Figure 2, the first score on the X axis is 2.3. Remember that a logarithm is the power that a number designated as the "base" has to be raised to equal some other number. If the number we want is 10, the logarithm in base "e" is whatever power 2.71828 has to be raised to equal 10. It turns out that 2.71828 to the 2.3 power equals 10 (or $2.71828^{2.3} = 10$). Thus, if the computer originally read a value (a score) on variable X of 10, it would now (after the logarithmic

transformation in base "e") read this same score as 2.3. If you compared the other scores on the two X axes on page 98, you would see this same relationship. Notice how the change from millions of constant dollars to the natural logarithm of millions of constant dollars compresses scores on the X axis. Just keep reading!! In Figure 1 on the preceding page, the non-zero scores on the X axis range from a high of 70 to a low of 10. This is a 7 to 1 ratio (i.e., 70 is 7 times 10). In Figure 2 the non-zero scores on the X axis range from a high of 4.2 to a low of 2.3. This is only approximately a 1.8 to 1 ratio (i.e., 4.2 is approximately 1.8 times as great as 2.3). Converting the highest to lowest scores on the X axis from a 7 to 1 ratio in Figure 1 into a 1.8 to 1 ratio in Figure 2 is why the relationship between X and Y went from being non-linear in Figure 1 to being linear in Figure 2.

It is important to note that if X is in logarithmic units, the interpretation of "b" is somewhat different than what we have done previously. Thus, for the equation: $Y = a + b(\ln X) + e$, a one unit increase in X would mean a one unit increase of the logarithm of X. Just keep reading!! For example, let us say that we estimated the equation: $Y = a + b(\ln X) + e$ for our OSHA example and the results were that "b" = -.215 and the standard error of b = .80. The "t ratio" would be -.215/.80 = -2.68. Since the "t ratio" of -2.68 has an absolute value of greater than 2.0 we would reject the null hypothesis that X is unrelated to Y in favor of the alternative hypothesis that X is negatively related to Y. Okay, so we know that "b" is statistically significant at the .05 level (i.e., that we will reject the null hypothesis because there is a 5% or less chance that the null hypothesis is true).

Now, how do we interpret "b"? Remember that since "b" equals -.215, it means that every one unit increase in X is associated, on average, with a little over a two-tenths (.2) of one unit decrease (i.e., -.215) in Y. Since a unit of Y is 1 serious injury per million hours worked, this means that a one unit increase in X is associated, on average, with a decrease of approximately .2 of 1 serious work related injury per one million hours worked (or a reduction of approximately one-fifth of a serious injury for each one million hours worked).

But what is a "unit of X"? Originally, variable X was the amount of money that the federal government spends annually (i.e., per year) on OSHA inspections measured in millions of constant dollars. Thus, as originally measured, a score of 10 on X would have meant that the federal government spent $10,000,000 on OSHA inspections in that particular year. However, remember that after the logarithmic transformation of X, the computer is now reading the "log" of X in base "e" (i.e., the power that 2.71828 must be raised to equal the score on X) Just keep reading!! Thus, if the computer read a score of 10 on X, it would now read this as 2.3 (i.e., $2.71828^{2.3} = 10$). So, a unit increase in the log of X would be an increase of 1.0 in the exponent of 2.71828. Just keep reading!!

For example, an increase from 2.3 to 3.3 in the exponent of 2.71828 (i.e., from $2.71828^{2.3}$ to $2.71828^{3.3}$) would be an increase of one unit in "ln X" (i.e., in the natural logarithm of X - just keep reading). To convert this into the original units of X (i.e., X in millions of dollars rather than in the natural logarithm of millions of dollars) would mean an increase from 10 million ($2.71828^{2.3} = 10$) to approximately 27 million dollars ($2.71828^{3.3} = 27.11$). So an increase of one unit in the natural logarithm of X (e.g., from $2.71828^{2.3}$ to $2.71828^{3.3}$) could, for example, be an increase of from 10 million to approximately 27 million dollars. This is a much larger increase than if X went from 10 million to 11

million dollars (i.e., a one unit increase in X in the original, non-logarithmic, units in which X was measured).

Additionally, each one unit increase in the exponent of 2.71828 is not the same (just keep reading). The difference of $2.71828^{2.3}$ to $2.71828^{3.3}$ (approximately 17 million dollars: $2.71828^{2.3}$ = 10; $2.71828^{3.3}$ = 27.11; 27.11 - 10 = 17.11) is less than the difference between $2.71828^{3.3}$ to $2.71828^{4.3}$ (approximately 46 million dollars: $2.71828^{3.3}$ = 27.11; $2.71828^{4.3}$ = 73.69; 73.69 - 27.11 = 46.58; 17 million dollars is less than 46 million dollars). Just keep reading! All of this complicates the interpretation. I would recommend just looking at the sign of "b" (i.e., which is negative in this case, -.215) and then see if "b" is statistically significant (which we know it is in this case). I would recommend going through all the "gymnastics" that I have only if necessary!! Fortunately, it rarely is!!

Keep in mind the theoretical logic of the variable transformation we just made. Since greater and greater increases in X (the amount of money spent on OSHA inspections), are needed to decrease scores on Y (the rate of serious injuries) by the same amount, if we want to preserve linearity between X and Y we will need to reduce the rate of increase in recorded scores on variable X. Just keep reading!!! Again, to use the dieting example. During the first week on a diet you may lose 5 pounds. However, during the second week you may only lose 2 pounds. It becomes more difficult to lose each additional pound. It may take you three weeks (i.e., weeks 2, 3 and 4) to lose as many pounds as you did during week 1. If variable Y is the number of pounds lost and variable X is the number of weeks you have been on your diet, then we would have to record weeks 2 through 4 as the same amount of time as week 1 if we are going to have a linear relationship between X and Y. That is just what we did in the OSHA example. As long as we realize that we have converted from weeks on the diet to the natural logarithm of weeks on the diet there is no problem. By converting the relationship between X and Y from being a non-linear relationship into a linear relationship, it is much easier for us to interpret "b." That was the point of the preceding example (to keep the interpretation of "b" the same as we learned on page 76).

It is important to mention that sometimes when the "t ratio" in a linear model suggests that X and Y are unrelated (i.e., that the "t ratio" is less than 2.0) X and Y are actually non-linearly related. For example, if the "t ratio" were 1.3, our decision rule says that since the "t ratio" has an absolute value of less than 2.0, we should not reject the null hypothesis that X is unrelated to Y. However, in some of these circumstances there is a statistically significant relationship between X and Y. It is just that the relationship is non-linear instead of linear. This is one reason to examine a scatter plot of the relationship between X and Y. The scatter plot may show a non-linear pattern. However, if the scatter plot reveals non-linearity, it does not automatically mean that we should pursue either a logarithmic (or some other) variable transformation. Before we transform a variable we should think through the situation from a theoretical perspective. Thus, we should ask: What theory would support the type of non-linearity that the scatter plot reveals? If there is not a theory to support the particular type of non-linearity we find, we should not transform the variable. Notice in the OSHA example that I mentioned a particular pattern of reasoning that would lead us to expect a non-linear relationship between the amount of money spent by OSHA and the rate of serious injuries at the workplace.

Quiz Questions: see the last paragraph on page 92.