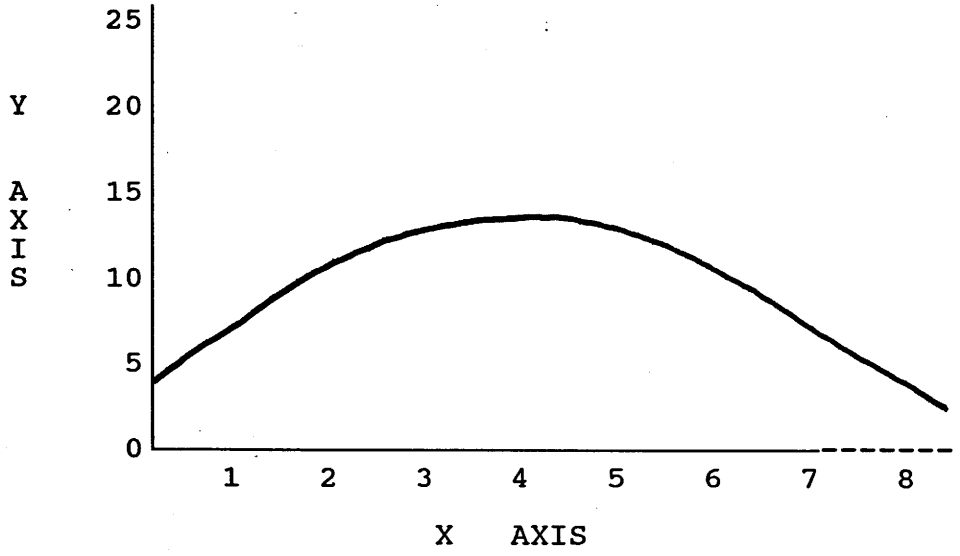
## Polynomial Variable Transformations

In the OSHA example, the direction of the relationship between X and Y never changed. For example, if you look at Figures 1 and 2 on page 98, while Figure 1 depicts a non-linear relationship and Figure 2 depicts a linear relationship, both Figures 1 and 2 show negative relationships. In other words, in every portion of both Figures 1 and 2 on page 98, if the score on X increases the score on Y decreases. This situation does not always occur in political science. In some instances, the relationship between X and Y changes direction (from positive to negative or vice versa). The diagram immediately below depicts a relationship between X and Y in which changes from a positive to a negative relationship.




The purpose of a polynomial ("polynomial" is defined later) transformation is to allow us to estimate relationships between X and Y which change direction (either from positive to negative or vice versa). The notion of a relationship between X and Y which changes direction will become clearer if we use an example from political science. Over the past several decades there have been a number of studies done by scholars in comparative politics on the causes of domestic violence. In such studies the dependent variable is the level of domestic violence (crime, riots, violent labor strikes, etc.) in a nation. One of the primary independent variables in such studies is the level of economic development in a nation. The basic expectation, which has been supported in a number of studies, is that as less developed nations start to industrialize (i.e., move from an agriculturally based economy to an industrially based economy) the level of domestic violence should increase. The theory is that as citizens are economically dislocated (i.e., those who leave the farms and move into cities and work in factories) this causes both fear and removes the social networks that were previously in place. Just keep reading!! The "fear" comes from fearing the unknown (i.e., what city and manufacturing life will be like). When people are physically moved, as they are when they leave the rural farms and move to a city, the social networks (i.e., friends, church, etc.) that they were accustomed to are lost. These networks would often provide support for them when they had encountered troubles in the past. Assuming that "contented" individuals are less violent, removing social networks and raising "fears" should increase the level of

violence.

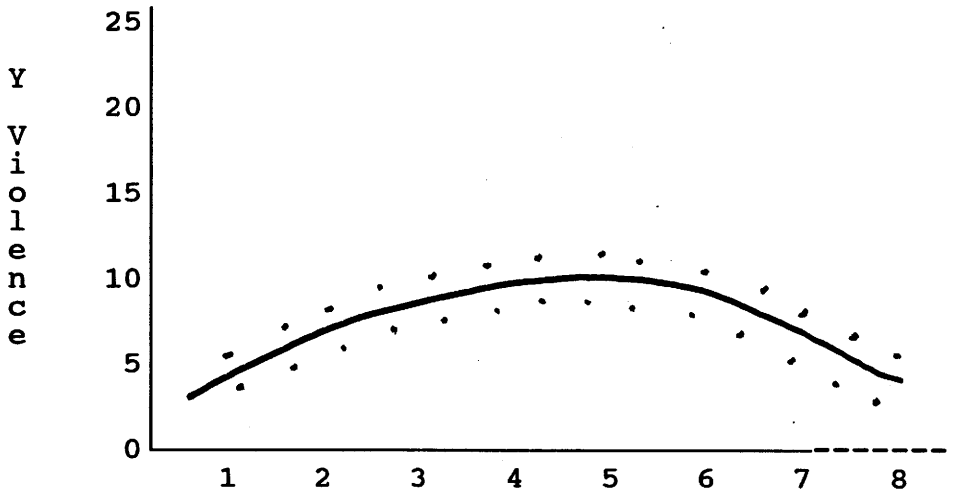
After the bulk of the population has moved into more urban areas and gained the wealth produced by the new industries, the level of violence should start to recede (i.e., diminish). This reduction in violence is likely to occur because after those who have moved from the farms to the city become "settled" in the city, they will feel more secure, social networks will be developed and they will gain wealth from the new industries they work in. All of these factors should produce a lower level of violence than when the nation converted from a rural agriculturally based economy to an economy based more on manufacturing.

Let us say that variable X is a nation's level of economic development and variable Y is the rate of domestic violence in that same nation. According to the preceding theory, we should expect a positive relationship between the level of economic development and domestic violence from low to mid-levels of economic development. Thus, as a nation starts to industrialize, its level of domestic violence should increase. Since an increase from a low level of economic development to a middle level of economic development (i.e., as a nation moves from a low level of economic development to a middle level of economic development) should be associated with a higher level of domestic violence, the relationship should be positive. However, as a nation moves from a middle level of economic development to a high level of economic development, the relationship should be negative. That is, nations with high levels of economic development should have lower rates of domestic violence than nations with middle levels of economic development.

Typically in such studies variable Y, the dependent variable (domestic violence), is measured by the number of acts of collective protest (e.g., number of strikers arrested, demonstrators arrested, etc.) per 100,000 people in a given year. Thus, if a nation had a score of "15" on variable Y in the year 1996, this would mean that there were 15 acts of domestic violence per 100,000 people in that nation in 1996. Variable X, the independent variable (level of economic development), is often measured by energy consumption per person (i.e., per capita) in a given year. Energy consumption per person is a good measure of the level of economic development because as a nation develops and attains a higher level of income, people consume more products which use energy (automobiles, mass transit, television sets, etc.).

The standard approach in measuring energy consumption is to convert all energy to metric (i.e., using the British metric system instead of ounces, pounds, gallons, etc.) tons of what are termed "coal equivalents." For example, we would convert gasoline consumption to an equivalent (i.e., equal) amount of coal. Instead of having gasoline measured in gallons, coal in tons, etc., this approach puts all energy consumption on the same scale. Such measures are available for many nations in the world over a relatively long time period. Using such a measure, if a nation had a score of "3" this would mean that the total energy consumption of the average person in that nation was the equivalent of 3 metric tons of coal in that particular year.

So, let us say that using the year 1995, we had data on both X and Y for 90 nations. A scatter plot of the data might look like the diagram on the next page.




X - Level of Economic Development in Metric Tons of Coal Equivalents Per Person

As you can see in the scatter plot above, it seems that the relationship between X and Y is positive from scores of 0 through 5 on X. That is, the higher the score on X, the higher the score on Y. However, for scores from 5 through 8 on X, the relationship is negative. Thus, for scores greater than 5 on X, the higher the score on X, the lower the score on Y.

Now we know that there is a "change of direction" (i.e., from positive to negative) in the relationship between a nation's level of economic development and its level of domestic violence. If we estimate this relationship via regression, what do we tell the computer to do? The computer is assuming that the relationship between X and Y continues in one direction (i.e., always positive or always negative).

The method by which political scientists typically handle a relationship between X and Y which changes direction (i.e., from positive to negative or vice versa) is through the use of powers of X. Just keep reading!! For example, to predict a score on Y our basic equation has been: $\hat{Y} = a + bX$ (see page 78)

To handle a situation like the one depicted in the scatter plot above, a political scientist would ask the computer to estimate the following equation to predict Y:

$$\hat{Y} = a + b_1X + b_2X^2$$ Just keep reading!!

While the preceding equation has two letter "b's" and two letter "X's" do not let this bother you! There is still only one "conceptual" independent variable (variable X). What is different about the equation above is that in addition to "X" there is also "$X^2$" (i.e., X to the second power or X squared). The $X^2$ term says that the computer will just take the score on X and square it (i.e., multiply it times itself). An equation with X raised to a power other than 1 (i.e., $X^2$) is called a polynomial transformation of X. Just keep reading!!

Let us say that a nation had a score of 3 on X. This would tell us that the average person in that nation consumed energy equal to 3 metric tons of coal equivalents. Applying this in the

equation:

$$\hat{Y} = a + b_1X + b_2X^2$$

means that the computer would read "3" for the score on X (i.e., for "X" in the term $b_1X$ above) and "9" for the score on $X^2$ [i.e., (3)(3) = 9 for $X^2$ in the term $b_2X^2$ above). Thus, for each nation, the computer is reading two scores on variable X (the score on X and that same score squared, i.e., for $X^2$).

How does using $X^2$ help us model the change in direction from positive to negative? The answer is that if we estimate the equation at the top of this page and $b_1$ is positive and $b_2$ is negative, then the regression line will eventually turn from a positive slope to a negative slope. Let us see how this works by using data for three nations. Suppose we ask the computer to estimate the above equation with both "X" and "$X^2$" and the results are as follows: $a = 1$ $b_1 = 2.000$ $b_2 = -.200$

So, as before, our prediction equation for Y is then:

$$\hat{Y} = a + b_1X + b_2X^2$$

Now let us insert the values for "a" (i.e., 1), "$b_1$" (i.e., 2.000) and "$b_2$" (i.e., -.200) in the prediction equation for Y:

$$\hat{Y} = 1 + [(2.000)(X)] + [(-.200)(X^2)]$$

Note: Adding a negative number is the same as subtracting a positive number. Therefore: $+ [(-.200)(X^2)]$ is equivalent to $- (.200)(X^2)$ which, in turn, is equivalent to $(-.200)(X^2)$

Now let us take data for three nations and see what the predicted value for Y is in each nation.

---

If X = 4 the prediction for Y is:

$$\hat{Y} = 1 + [(2.000)(4)] + [(-.200)(4^2)]$$

$$\hat{Y} = 1 + [8] + [(-.200)(16)]$$

$$\hat{Y} = 1 + 8 + [-3.2]$$ {Note: + [-3.2] is equal to - 3.2}

$$\hat{Y} = 1 + 8 - 3.2$$

$$\hat{Y} = 9 - 3.2$$

$$\hat{Y} = 5.8$$

---

If X = 5 the prediction for Y is:

$$\hat{Y} = 1 + [(2.000)(5)] + [(-.200)(5^2)]$$

$$\hat{Y} = 1 + [10] + [(-.200)(25)]$$

$$\hat{Y} = 1 + 10 + [-5]$$ {remember: + [-5] is equal to - 5}

$$\hat{Y} = 1 + 10 - 5$$

$$\hat{Y} = 11 - 5$$

$$\hat{Y} = 6$$

If X = 6 the prediction for Y is:

$\hat{Y} = 1 + [(2.000)(6)] + [(-.200)(6^2)]$

$\hat{Y} = 1 + [12] + [(-.200)(36)]$

$\hat{Y} = 1 + 12 + [-7.2]$ {note: + [-7.2] = - 7.2}

$\hat{Y} = 1 + 12 - 7.2$

$\hat{Y} = 13 - 7.2$

$\hat{Y} = 5.8$

So, given our values for "a" (i.e., 1), "$b_1$" (i.e., 2.000) and "$b_2$" (i.e., -.200) the predicted value for Y if X = 4 is 5.8. If X = 5 the predicted value for Y is 6. If X = 6 the predicted value for Y is 5.8. Did you notice what happened? If X increases from 4 to 5, the predicted value of Y increases from 5.8 to 6.0. This indicates a positive relationship between X and Y because a higher score on X (5 as opposed to 4) leads to a higher predicted value for Y (6.0 as opposed to 5.8). However, if X increases from 5 to 6 the predicted value of Y decreases from 6.0 to 5.8. This indicates a negative relationship between X and Y because a higher score on X (6 instead of 5) leads to a lower predicted value for Y (5.8 as opposed to 6.0). This is a "change" in direction (i.e., from a positive relationship to a negative relationship).

This "change of direction" occurred because the "sign" on "$b_1$" (which was positive) was the opposite of the "sign" on "$b_2$" (which was negative). Notice that the relationship between X and Y over the higher scores on X (i.e., 6, 7, 8) is negative in the diagram on page 103. Notice also that the coefficient of $X^2$ (i.e., "$b_2$") is negative. This is not a coincidence. In an equation with a non-squared value of X (i.e., X) and a squared value of X (i.e., $X^2$), the relationship for the higher scores on X will invariably be the same as the sign on the coefficient of $X^2$. In our example this means that since "$b_2$" is the coefficient of $X^2$ and $b_2$ is negative (i.e., -.200), the relationship between X and Y will be negative for the highest values on X (i.e., 6, 7, 8). This is because the value of the squared term (i.e., $X^2$) grows at such a greater rate than the value of the non-squared term (i.e., X) that it eventually dominates the calculation. For example, if X increases from 3 to 4, the value of X increases by 1 (i.e., from 3 to 4) but the value of $X^2$ increases by 7 (i.e., from 9 to 16, thus $3^2 = 9$ and $4^2 = 16$). An increase of 7 is much greater increase than a increase of 1. That is why $X^2$ will ultimately dominate the calculation of Y.

Notice that neither "$b_1$" nor "$b_2$" has been raised to a power other than 1 (i.e., there is not a term such as $b_1^{.7}$) As in the logarithmic transformation we previously discussed, by keeping "b" raised to the first power (i.e., just "b") we make the job of estimating and interpreting the results easier. However, as you might guess from the computations I went through above, interpretation of "b" would be difficult. Typically, political scientists just interpret the sign (i.e., positive or negative) and the level of statistical significance.

Political scientists handle a "change of direction" by the polynomial transformation method that we just used. The polynomial transformation is important because a number of relationships in political science "change direction" (i.e., from positive to negative or vice versa).

## Multiple Regression

Up to this point we have worked with only one independent variable at a time. For example, we used variation in county youth unemployment rates (variable X) to explain variation in county delinquency rates (variable Y). When we use regression to estimate the relationship between one independent variable and the dependent variable we call this "bivariate" (i.e., two variable) regression. In the term bivariate regression "bi" refers to "two" and "variate" refers to "variables."

How realistic is our bivariate regression model of delinquency? Don't other factors than the youth unemployment rate effect the delinquency rate? For example, supposing a county adopts a gang prevention program. Assuming the program has some beneficial effect, shouldn't the delinquency rate be lower in this county than it otherwise would have been? Yes! But we did not include a measure of a county's anti-gang activity in our model of delinquency. This points up the need for having more than one independent variable. In the current example, both the youth unemployment rate and the county's anti-gang activity would be independent variables. The term "multiple" regression means we have more than one (hence "multiple") independent variables.

Since few dependent variables are influenced (or, if our theory is strong enough, "caused") by only one independent variable, multiple regression permits us to build more realistic models. Thus, incorporating a measure of a county's anti-gang activity in our model would provide a fuller and more realistic picture of why delinquency rates vary among counties.

Furthermore, when we include additional independent variables it provides a "truer" reading of the impact of the one independent variable we used in a bivariate regression. Just keep reading!! For example, on page 76, we found that if the youth unemployment rate increased by one unit (1% in this case), the delinquency rate would, on average, be expected to increase by almost two-tenths of one unit (.198, almost two-tenths of 1% in this case). However, if we include a measure of county anti-gang activity as a second independent variable, the impact of the youth unemployment rate on delinquency is likely to change. Thus, in a multiple regression model it is highly likely that the impact of youth unemployment will be either greater or lesser (probably lesser) than approximately two-tenths of a unit (i.e., some number other than .198). So, multiple regression provides both a more realistic portrait of the world we are trying to picture, and provides a more accurate assessment of the impact of the one independent variable we used in a bivariate regression model.

While adding additional independent variables provides a more realistic view of the world, we should not just continually add independent variables to our model. Any independent variable we use needs to have a firm theoretical justification. In the above example, the reason to include a measure of a county's anti-gang activity as an independent variable is that there is a clear theoretical reason to do so. Being a gang member is likely to expose a young person to activities that would be thought of as delinquent. All other factors being equal, being in an organization that both pursues illegal activities, and fosters a strong sense of group allegiance toward such activities, should increase the chances of a member becoming delinquent. On the other hand, the percentage of a county's youth who prefer basketball to baseball is probably irrelevant to delinquency. Why would preferring either basketball over baseball, or vice versa, effect the probability that a youth would become delinquent? I have no

idea. So, including a measure of county anti-gang activity makes sense in a model of county delinquency, but sports preference probably does not. Any independent variable we include in a model needs to have a firm theoretical justification.

The meaning of "b" is somewhat different in multiple regression than in bivariate regression. As a first step in understanding this "revised" meaning of "b," let us compare a bivariate regression equation with a multiple regression equation. In the multiple regression equation I will use three independent variables.

Bivariate Regression Equation:

$Y = a + bX + e$

Multiple Regression Equation with Three Independent Variables:

$Y = a + b_1X_1 + b_2X_2 + b_3X_3 + e$

Now let me describe the variables in the multiple regression equation above. The following list of variables is very similar to Assignment #2 in POSC 300B. Let us say that we are examining voting in the U.S. Senate. Specifically, suppose we are attempting to explain why some senators are much more supportive of tax legislation that shifts the tax burden more to high income groups than other senators. The "unit of analysis" is a senator. Since there are 100 senators "N" = 100. Variable "Y" is the percentage of times the senator voted to shift the tax burden more toward high income earners. Variable "$X_1$" is the degree of conservatism of the senator. On this measure senators could score from 0% to 100% conservative (i.e., higher scores denote greater conservatism). Variable $X_2$ is the senator's party affiliation (scored "1" if the senator is a Democrat, "0" if the senator is a Republican). Variable $X_3$ is the state median family income in the senator's state. This variable is scored in thousands of dollars. Thus, if the computer reads a score of 20.3 this would mean that the median family income in the senator's state is $20,300. Remember that the median is a positional measure that divides the scores into two equal groups. Thus, a score of 20.3 on the median family income variable ($X_3$) would mean that half of the families in the senator's state had incomes above $20,300 and half the families had incomes below $20,300.

As an example of how to understand the meaning of "b" in a multiple regression, let us examine "$b_1$" in the above equation. We know that $b_1$ is the coefficient of $X_1$ (i.e., notice the term $b_1X_1$ above). Given this, you would probably guess that $b_1$ would indicate how many units, on average, Y would increase or decrease (depending upon whether $b_1$ is positive or negative) if $X_1$ increased by one unit. So far, so good!! However, there is more! What $b_1$ represents is, on average, the number of units of increase or decrease in Y (depending upon whether $b_1$ is positive or negative) if $X_1$ increases by one unit and the score on all other independent variables (in this example, $X_2$ and $X_3$) remains constant (i.e., the same). Just keep reading, it will become clear!!

Now let us take a practical example of this from our U.S. Senate study. Suppose we had two senators who were members of the same political party and were also representing the same state. In this situation we know that the scores on both political party affiliation ($X_2$) and state median family income ($X_3$) must be the same (they are both members of the same political party and

represent the same state). Now suppose the two senators differ on the degree to which they support shifting the tax burden to higher income earners. We know that the difference in their support for raising taxes on high income earners can not be explained by differences in either their party affiliation ($X_2$) or the median family income in their states ($X_3$). We know this because their scores on party affiliation and state median family income are the same. The difference in their support for shifting the tax burden to high income earners is either explained by differences in their level of conservatism ($X_1$), or is left unexplained by our model.

Now, let us say that these two senators differed by 1 unit in their level of conservatism. For example, let us say that senator #1 had a score of 70 on $X_1$. This would mean that the senator was 70% conservative. Let us then say that senator #2 had a score of 71% on conservatism. Now suppose that senator #1, who is 1% less conservative than senator #2, scores three-tenths of a unit higher on support for shifting the tax burden to high income earners. This would suggest that $b_1$ should be -.300. Why? First, because the relationship between $X_1$ and Y is negative for these two senators. We know this because the senator with the higher score on $X_1$ (conservatism) scored lower on Y (support for shifting the tax burden to high income earners). Additionally, the magnitude of this effect over these two senators is -.300 because the senator who was 1 unit higher on $X_1$ (conservatism) was three-tenths of 1 unit lower on support for shifting the tax burden to high income groups. Additionally, since there were no differences between the senators on either $X_2$ or $X_3$ (i.e., they both had the same score on party affiliation and median family income), we know that neither of these two other independent variables can account for this three-tenths of 1 unit difference in their support for shifting the tax burden to high income earners. The interpretation of $b_1$ is then as follows: $b_1$ represents the average number of units of increase or decrease (depending upon whether $b_1$ is positive or negative) in Y if $X_1$ increases by one unit and the level of all other independent variables are held constant (i.e., remain at the same level - as did political party affiliation and state median family income in the preceding paragraph). Just keep reading!!

The notion of "holding the level of all other independent variables constant" is critically important. Suppose in the example above senators #1 and #2 differed on both their degree of conservatism and their political party affiliation (i.e., one was a Democrat and the other a Republican). Now, if their scores on Y were three-tenths of 1 unit different, which variable accounts for this? Since they differ on both conservatism and political party affiliation, we can not tell. We need to hold their party affiliation constant (by making them both Democrats for instance) to then see how much impact their differing scores on conservatism have on their willingness to support shifting the tax burden to high income earners. Since our goal is to find the magnitude of the impact of each of the independent variables on the dependent variable, we must "control" (or eliminate) the impact of all other independent variables on the dependent variable. Keep in mind that just because $X_2$ and $X_3$ do not explain some portion of Y, does not automatically mean that $X_1$ does. $X_1$ may also fail to explain the portion of Y that is not explained by $X_2$ or $X_3$. If this occurs, then the portion Y that is not explained by $X_1$, $X_2$ or $X_3$ becomes the value of "e" (the error term).

Beginning with this reading assignment you should change your interpretation of "a," "b" and $R^2$. After I "re-define" "a," "b" and $R^2$ I will show you how to use these "new" definitions on the type of quiz I will give you.

"a" is the Y intercept and is also the predicted value of Y if all independent variables are simultaneously zero. Just keep reading!! Do not worry if the definitions seem "abstract" and incomprehensible. All the definitions ("a," "b" and $R^2$) will become clear when I work the sample quiz. So, in those now famous words: just keep reading!!!

In the current situation, the "new" definition for "a" would mean that whatever value "a" is would be the predicted value for Y if the senator scored zero on conservatism (which would mean they had "no" conservatism - thus they were as liberal as measurement would permit), scored zero on political party (which means they are a Republican), and came from a state where the median family income was zero dollars. As you can tell, this is a highly unrealistic set of outcomes. For example, how could a state have a median income of zero? Half the families had either no income or a negative income? Not plausible. Also, scoring zero on both political party and conservatism would mean a Republican who had no conservatism! If they have no conservatism why are they in the Republican party? Remember, all three of the outcomes would have to occur simultaneously (i.e., at the same time) for "a" to actually be the predicted value for Y. Obviously, the value of "a" depicts a state of the world which is extremely unlikely to occur. Typically, political scientists are not really interested in the value of "a." Generally speaking, political scientists do not use "a" alone. Rather, "a" is used in conjunction with "b" and "X" to make predictions for Y (as I will show you later). Just keep going!!

"b" is the average number of units of increase or decrease (depending upon whether "b" is positive or negative) in Y if the variable which "b" is a coefficient of (i.e., $X_1$ if we are referring to "$b_1$") increases by one unit and the level of all other independent variables remain constant. Don't panic, just keep reading!!

$R^2$ is now the percentage of variation in the dependent variable explained by all independent variables together (it is a good idea to reread the second paragraph on page 84 on $R^2$).

Now, let me show you a hypothetical quiz and answers. Suppose I asked a question such as: Given the following regression results, what statements could you make?

a = 49 $b_1$ = -.389 standard error of $b_1$ = .115

$R^2$ = .56 $b_2$ = .479 standard error of $b_2$ = .275

Here is what I hope you would answer:

a: If the value of $X_1$ and $X_2$ are both zero, the predicted value of Y is 49. We could also say that the Y intercept is 49.

$b_1$: Holding the level of all other independent variables constant, if $X_1$ increases by one unit, on average, Y will decrease by almost four-tenths of one unit. Since $b_1$ has an absolute value of more than twice its own standard error (i.e., -.389/.115 has an absolute value greater than 2.0), we would reject the null hypothesis that the true value of $b_1$ is equal to .000 because less than 5 times out of 100 is the null hypothesis actually true.

$b_2$: Holding the level of all other independent variables constant, if $X_2$ increases by one unit, on average, Y will increase by almost five-tenths (i.e., one half) of a unit. Since $b_2$ does not have an absolute value of at least twice its own standard error (i.e., .479/.275 is less than 2.0), we would not reject the null hypothesis that the true value of $b_2$ is equal to .000 because the null hypothesis is true more than 5 times out of 100. Notice I did not say we accept the null hypothesis. All I said was that we could not reject the null hypothesis.

$R^2$: Variation in all the independent variables together explains approximately 56% of the variation in the dependent variable.

Please Note: If I name the variables and give you the units of measure I would expect you to use such information. For example, suppose that variable Y was the percentage of times a senator voted to shift the tax burden to high income earners, variable $X_1$ was the level of the senator's conservatism (measured in percentage points from 0 - least conservative, to 100 - most conservative) and $X_2$ is the senator's party affiliation (Republican = 0, Democrat = 1). I would expect you to say for $b_1$ that, holding the level of $X_2$ constant, a one percentage point increase in a senator's conservatism ($X_1$) is associated, on average, with approximately a four-tenths of a percentage point decline in support for increasing the tax burden on high income earners. Remember that "b" is always in the units of the dependent variable. So, if Y is measured in percentage points, all the "b's" will be in percentage points. However, remember to check the units of measure for each variable (if they are given). All variables are not measured in percentage points. For example, is political party measured in percentage points? No! So, it would not make sense to say a one percentage point increase in political party affiliation. For $R^2$ I would expect you to say that variation in senatorial conservatism and political party affiliation explains approximately 56% of the variation in senatorial support for shifting the tax burden more to high income earners (reread the second paragraph on page 84 on $R^2$).

Remember that the regression model predicts a value for the dependent variable for each observation (page 78). Thus, in our example of voting in the U.S. Senate we have scores on the dependent variable (the percentage of times the senator voted to shift the tax burden to higher income earners) for all 100 senators. The regression model will "predict" 100 scores on the dependent variable (one score for each senator). Using a set of results somewhat similar to what those of you in POSC 300B will have in Assignment #2, let us see how the computer does this. The variables are as follows: Y is the percentage of times the senator voted to shift the tax burden more to high income earners; $X_1$ is the senator's degree of conservatism (from 0 - least conservative to 100 - most conservative); $X_2$ is the senator's political party affiliation (Democrat = 1, Republican = 0); and $X_3$ is the median family income in thousands of dollars in the senator's state (i.e., a score of 22.3 would mean that the median family income in that particular state was $22,300). Our basic equation is:

(equation 1) $Y = a + b_1X_1 + b_2X_2 + b_3X_3 + e$

If you remember from the past, $\hat{Y}$ is the symbol for the predicted value of Y. Also, remember that "e" is the residual. The residual (or "e") is the difference between the actual value on Y (i.e., the percentage of times the senator voted to shift the tax

burden to higher income earners) and the predicted value on Y (i.e., the regression model's prediction of the percentage of times the senator voted to shift the tax burden to higher income earners). Rather than go through a series of algebraic manipulations, just believe me when I tell you that the formula for the predicted value of Y is:

(equation 2) $\hat{Y} = a + b_1X_1 + b_2X_2 + b_3X_3$

In order to show you how the computer predicts a value for Y, we need two more pieces of information. First, we need values for "a," "$b_1$," "$b_2$" and "$b_3$." So, let us say those values are as follows: $a = 85$ $b_1 = -.477$ $b_2 = 5.757$ $b_3 = -.625$

Second, we need values for $X_1$, $X_2$ and $X_3$. Suppose we found out that senator #1 scored the following:

$X_1$: 40 (meaning senator #1 voted 40% of the time in a conservative direction);

$X_2$: 1 (meaning senator #1 is a Democrat);

$X_3$: 19.7 (meaning the median family income in senator #1's state is $19,700).

To obtain a predicted value on Y for senator #1 we need only insert the values for $X_1$, $X_2$ and $X_3$ into equation 2. So, let us do that:

(equation 2) $\hat{Y} = a + b_1X_1 + b_2X_2 + b_3X_3$

for senator #1 this becomes:

$\hat{Y} = 85 + [(-.477)(40)] + [(5.757)(1)] + [(-.625)(19.7)]$

$\hat{Y} = 85 + [-19.08] + [5.757] + [-12.31]$

$\hat{Y} = 85 - 19.08 + 5.757 - 12.31$

$\hat{Y} = 59.37$

So, our model predicts that senator #1 would support shifting the tax burden to higher income earners 59.37% of the time. Suppose that senator #1 actually supported shifting the tax burden to higher income earners 65% of the time. If this was the case then our model predicted a score for senator #1 that was less than the score actually attained (i.e., 59.37% is less than 65%). The value of "e" (the residual) for senator #1 would be:

$$e = Y - \hat{Y}$$
$$e = 65 - 59.37 = 5.63$$

Remember that "e" (5.63 in this instance) is that portion of the dependent variable (i.e., the senator's support for shifting the tax burden to higher income earners) that could not be predicted from knowledge of the senator's conservatism, the senator's party affiliation or the median family income in the senator's state. Since the value of "e" for senator #1 is 5.63, the value of "$e^2$" for senator #1 is 31.70 [i.e., (5.63)(5.63) = 31.70]. Remember also

that the values for "a," "$b_1$," "$b_2$" and "$b_3$" were selected to minimize the total sum of squared errors (i.e., the sum of "$e^2$" for all 100 senators - if confused, review pages 85-87). Put another way, any values for "a," "$b_1$," "$b_2$" and "$b_3$" other than those just used in making the prediction of senator #1's score on support for shifting the tax burden to higher income earners would have produced a greater total sum of squared errors (i.e., a great sum of scores on "$e^2$") than the values selected by the computer.

Previously in our discussion of multiple regression I explained the logic behind "controlling" for the level of each independent variable. In other words, of taking two senators from the same state (hence the same score on median family income) and from the same political party and then seeing how their differing scores on conservatism were related to their differing scores on support for shifting the tax burden more to high income earners. While I explained the logic of "controlling" for other independent variables, I did not explain how the computer actually accomplishes this. That is what I would like to do now.

The example that I am going to use was inspired by Michael S. Lewis-Beck's, Applied Regression: An Introduction, page 50. To explain the process, let us take the four variable regression model (i.e., $X_1$, $X_2$, $X_3$ and Y) we discussed earlier. The variables are as follows: Y is the percentage of times the senator voted to shift the tax burden more to high income earners; $X_1$ is the senator's degree of conservatism (from 0 - least conservative to 100 - most conservative); $X_2$ is the senator's political party affiliation (Democrat = 1, Republican = 0); and $X_3$ is the median family income in thousands of dollars in the senator's state (i.e., a score of 22.3 would mean that the median family income in that particular state was $22,300). Suppose we try to estimate the following equation:

(equation 3) $Y = a_1 + b_1X_1 + b_2X_2 + b_3X_3 + e_1$.

Please Note: I used a subscript on "a" and "e" (i.e., $a_1$ instead of just "a" and $e_1$ instead of "e") because I will need to use several equations, each with their own "a" and "e." Thus, having subscripts helps us differentiate the various equations that will follow.

The goal of the following discussion is to explain how the computer estimates the value of "$b_1$" in equation 3 above. If you understand how "$b_1$" is estimated, you will understand the process for estimating any "b" (e.g., "$b_2$"). The key to understanding this process is to remember that "e" (or "$e_1$" in equation 3) represents that portion of the dependent variable (Y in equation 3) that can not be explained by the independent variables ($X_1$, $X_2$ and $X_3$ in equation 3). If you keep this in mind the following discussion will be easier to follow.

When we previously discussed the notion of "controlling," or holding the level of other independent variables constant (i.e., the same), I used the example of having two senators from the same state (thus having the same score on median family income because there is only median family income in any state) and being members of the same political party. I then compared how differences in these two senators level of support for shifting the tax burden to high income earners (the dependent variable, Y) was related to differences in their level of conservatism. The impact of the senator's conservatism on their support for shifting the tax burden more to high income earners is represented by "$b_1$." So, how does the computer hold the level of $X_2$ and $X_3$ "constant" in calculating

the value of $b_1$? To do this we need to remove the effects of both $X_2$ and $X_3$ from both Y and $X_1$. This is what I did "verbally" in the above example when I said that the level of both $X_2$ and $X_3$ would remain the same and we would see how differences in $X_1$ were related to differences in Y. Remember, "e" is that part of the dependent variable that can not be explained by the independent variables. If you find the going a little "tough," you know what to do, just keep reading!!! Now here's what the computer would do. First, let us remove the effect of $X_2$ and $X_3$ from $X_1$. To do this we tell the computer to estimate the following equation:

(equation 4) $X_1 = a_2 + b_4X_2 + b_5X_3 + e_2$

Do not let equation 4 confuse you. First, Y does not always have to be the dependent variable. Typically, Y is the dependent variable. However, in equation 4, $X_1$ is the dependent variable. The dependent variable is whatever variable is to the left of the equal sign (i.e., to the left of "="). Second, do not be confused by the subscripts (i.e., $a_2$, $b_4$, $b_5$ and $e_2$). Didn't I use "$a_1$," "$b_1$," "$b_2$," "$b_3$" and "$e_1$" in equation 3? Yes!! So, if I used those same subscripts in equation 4 it would be more difficult to follow. I will eventually use "$a_3$," etc. to differentiate later equations from equation 4. What else could I do?

Let us examine equation 4 closely. In equation 4, "$e_2$" is that portion of $X_1$ (the dependent variable in equation 4) that can not be explained by $X_2$ and $X_3$. Thus, "$e_2$" is $X_1$ "freed" from the effects of $X_2$ and $X_3$. I use the term "freed" because "$e_2$" represents that portion of $X_1$ that can not be explained by $X_2$ and $X_3$. By definition, the "error term" (in this case "$e_2$") is what "errors" (i.e., mistakes) we make in predicting scores on $X_1$ from our knowledge of scores on $X_2$ and $X_3$. Therefore, whatever impact $X_2$ and $X_3$ have on $X_1$ is found in "$b_4$" and "$b_5$," but not in "$e_2$" (the error term in equation 4). Thus, the values of "$e_2$" are "freed" from influence from either "$X_2$" or "$X_3$."

Now, if we can do the same thing to "Y," we would then be able to explore the relationship between $X_1$ and Y after removing the effects of both $X_2$ and $X_3$. This is exactly what we are going to do! To remove the effects of $X_2$ and $X_3$ from Y, we estimate equation 5 immediately ahead:

(equation 5) $Y = a_3 + b_6X_2 + b_7X_3 + e_3$

To apply the same interpretation to equation 5 that we did to equation 4, "$e_3$" represents that portion of Y (the dependent variable in equation 5) that can not be explained by $X_2$ and $X_3$. Thus, "$e_3$" is Y "freed" from the effects of $X_2$ and $X_3$. Remember, that when I explained this process "verbally" I said that after we knew that $X_2$ and $X_3$ were at the same level for both senators (i.e., "controlled") we would then see how $X_1$ related to Y. That is exactly what we will have the computer do. At this point, we have two pieces of valuable information, "$e_2$" and "$e_3$." I am saying these two "pieces of information" (which are really variables) are valuable because they give us the means to estimate "$b_1$" in equation 3. This has been our goal. So, let us use "$e_2$" and "$e_3$" to estimate "$b_1$" in equation 3. The formula is as follows:

(equation 6) $e_3 = a_4 + b_1e_2 + e_4$ ($b_1$ in equation 6 is the estimate of $b_1$ in equation 3 on page 112)

What is different in equation 6 above from the typical bivariate regression equation: $Y = a + bX + e$ is that both the independent and dependent variables are error terms from previous equations. Thus, the independent variable in equation 6 ($e_2$) is the error term from equation 4 on page 113. Similarly, the dependent variable in equation 6 ($e_3$) is the error term from equation 5 on page 113. As noted above, $b_1$ in equation 6 is the estimate of $b_1$ in equation 3 on page 112.

To estimate $b_1$ in equation 6 above, all we need to do is apply the same formula we used for "b" on page 76. In applying the formula for b on page 76, it is important to note that the mean scores for both $e_2$ and $e_3$ are zero. This simplifies the formula for b on page 76 to the following (just keep reading):

$$(equation\ 7) \qquad b_1 = \frac{\Sigma(e_2)(e_3)}{\Sigma e_2^2}$$

Do not panic! I will show the entire process for estimating $b_1$ in class using the data set for Assignment #2 in POSC 300B.

Let me mention that if you took a statistics or econometrics course you would probably not use the formula for $b_1$ that I just did. The formula they would use produces the same answer as the formula I used. I prefer the formula for $b_1$ just presented because it is easier for you to grasp the logic of the process. The formula the computer actually uses would not shed much light on the logic of calculating $b_1$. The calculation of $b_2$ and $b_3$ is the same as $b_1$ except that the position of the variables in equations 4 and 5 is different. Do not worry about calculating $b_2$ and $b_3$. If you understand the process for calculating $b_1$ that is sufficient.

Why do we want to know the impact of $X_1$ on Y apart from the impact of $X_2$ on Y and of $X_3$ on Y? We want to know this because if $X_1$ is related to either $X_2$ or $X_3$, we do not want $X_1$ (i.e., through $b_1$) to receive credit for an impact on Y which really belongs to either $X_2$ or $X_3$. For example, if senators from wealthier states are more conservative than senators from poorer states, we need to assess how much effect each of these two factors (i.e., state wealth and conservatism of the senator) has on the willingness of senators to shift the tax burden to higher income earners. If we do not "control" (i.e., remove) the effects of each other independent variable, we can never accurately estimate the individual impact of any one independent variable. Thus, we might falsely conclude that the senator's degree of conservatism ($X_1$) has an important impact on how willing the senator is to support shifting the tax burden to higher income groups when, in reality, it might be differences in state median family income (i.e., state wealth) that is truly the important factor. Political scientists will never be able to build realistic models of any political process unless we can "control" for various influences.

For quizzes: (1) Make sure you know how to interpret a, $b_1$, $b_2$ and $R^2$ (see pages 109-110); (2) Pay attention to the example in the middle of page 110 which interprets specific units of measure for the variables; (3) Make sure you can explain in words both what a polynomial variable transformation is and why a political scientist might use one.

# Multicollinearity

Suppose we have the computer estimate "a," "$b_1$," "$b_2$" and "$b_3$" for the following equation: $Y = a + b_1X_1 + b_2X_2 + b_3X_3 + e$. Now further suppose that the results show that $b_1$ is statistically insignificant. That is, that the "t ratio" for $b_1$ (i.e., $b_1$/the standard error of $b_1$) has an absolute value of less than 2.0. From past readings, we know that our "decision rule" is not to reject the null hypothesis that $b_1 = .000$ because the null hypothesis is true greater than 5% of the time. Thus, if we reject the null hypothesis, we have greater than a 5% chance of committing a "type I" error (i.e., rejecting the null hypothesis when the null hypothesis is true). Accordingly, our "decision rule" in such a situation is not to reject the null hypothesis.

Why might $b_1$ have been statistically insignificant? One possible answer is that our hypothesis is simply incorrect. Thus, it just maybe that $X_1$ has no effect on Y. However, for the sake of discussion, let us stipulate that $X_1$ actually does influence Y. Therefore, the "truth" is that $b_1$ does have a value other than .000. If $b_1$ really does have a value other than .000, why did the "t ratio" produce such a low score (i.e., an absolute value less than 2.0) that we could not plausibly reject the null hypothesis?

One possible answer to this question is "multicollinearity." "Multi" means more than one (i.e., more than one independent variable). "Collinearity" means sharing the same (i.e., "co") line (i.e., "linearity"). Multicollinearity means that the independent variables are correlated with each other. If multicollinearity is extremely high (i.e., the independent variables are very highly related to each other), then the computer can not reliably estimate the impact of each independent variable on the dependent variable.

For example, take the U.S. Senate example we have been working with. The variables are as follows: Y is the percentage of times the senator voted to shift the tax burden more to high income earners; $X_1$ is the senator's degree of conservatism (from 0 - least conservative to 100 - most conservative); $X_2$ is the senator's political party affiliation (Democrat = 1, Republican = 0); and $X_3$ is the median family income in thousands of dollars in the senator's state (i.e., a score of 22.3 would mean that the median family income in that particular state was $22,300). Let us suppose that every Democrat had the same low score on conservatism (say 30%). Furthermore, suppose that every Republican had the same high score on conservatism (say 80%). Now put yourself in the computer's place. If Democrats and Republicans differed on how willing they were to shift the tax burden to higher income earners (variable Y), would this be because of their party affiliation or their conservatism? This is hard to answer because each time we have a Democrat we also have a senator who is low in conservatism and each time we have a Republican we have a senator who is high in conservatism. What we really need is a group of Democrats who score "high" on conservatism (e.g., greater than say 60%) and a group of Republicans who score low on conservatism (say below 40%). Additionally, it would be useful if every Republican who scores high on conservatism did not score exactly 80% and every Democrat who scores low on conservatism did not score exactly 30%. If we had this additional information, the association between political party affiliation and conservatism would be greatly weakened. Then we could get a better idea of the impact of both party affiliation and conservatism on the willingness of senators to support shifting the tax burden to higher income groups. Unfortunately, in this hypothetical example, we lack such information. In such a

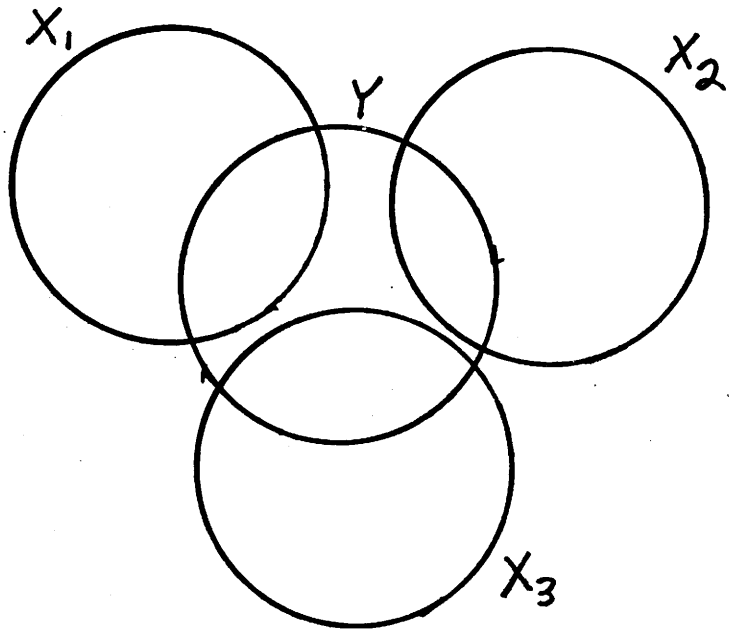
situation the computer would be very unsure of the impact of either political party affiliation or conservatism on the senator's willingness to support shifting the tax burden to higher income earners. If so, the computer is likely to report low "t ratios" (i.e., below 2.0) for both conservatism ($b_1$) and political party affiliation ($b_2$).

Why does multicollinearity result in lower "t ratios"? To avoid a long mathematical exercise, let me just tell you the consequences of high multicollinearity. Remember that the "t ratio" is "b" divided by the standard error of "b" (i.e., "b"/standard error of "b"). What high multicollinearity does is increase the size of the standard error of "b." To see the effect of this consider the following example. Let us say the "b" is 5 and the standard error of "b" is 2. The "t ratio" would be 2.5 (because 5/2 = 2.5). Since the "t ratio" has an absolute value of greater than 2.0 (i.e., 2.5) our "decision rule" is to reject the null hypothesis. Under high multicollinearity what would likely happen is that "b" would remain at "5" but the standard error of "b" might increase to 4. Now the "t ratio" would be 1.25 (5/4 = 1.25). Since the "t ratio" now has an absolute value of less than 2.0 (i.e., 1.25) our "decision rule" is not to reject the null hypothesis. This is the typical effect of high multicollinearity.

It is important that you understand that multicollinearity is a lack of information, not a lack of data. For example, suppose that instead of data on 100 senators we had data on 300 senators. The additional 200 senators only help us if there is not a perfect association between political party affiliation and conservatism. Thus, if in these additional 200 senators, every Democrat has the same low score on conservatism (say 30%) and every Republican has the same high score on conservatism (say 80%), our situation is no better than when we had just 100 senators. We still have no information about Democratic senators who score high on conservatism and Republican senators who score low on conservatism. One of the reasons larger samples are preferable to smaller samples is that the larger the sample size, the less likely it is that the independent variables will be perfectly, or very highly, correlated with each other (on correlation see page 75). In other words, with 300 senators it is less likely that every Democrat would have the same low score on conservatism and that every Republican would have the same high score on conservatism than with 100 senators.

Perhaps the best way of understanding multicollinearity is through diagrams. The three diagrams we will examine show various degrees of multicollinearity. In the upcoming diagrams, each variable ($X_1$, $X_2$, $X_3$ and Y) will be represented by separate circles. The circle for each variable represents the variation in scores for that particular variable.

Figure 1 - No Multicollinearity




There is no multicollinearity in Figure 1 because while each independent variable ($X_1$, $X_2$ and $X_3$) overlaps with the dependent variable (Y), no two independent variables overlap with the dependent variable in the same place. For example, none of the portion of Y (the dependent variable) which is explained by $X_1$ is also explained by either $X_2$ or $X_3$.

The situation depicted in Figure 1 is extremely unlikely to occur in political science. The independent variables are almost always correlated with each other. By contrast, in Figure 1, each pair of independent variables has a correlation of .00 (on correlation see page 75). For example, in Figure 1 the correlation of $X_1$ and $X_2$ = .00 and the correlation of $X_1$ and $X_3$ = .00. Such an occurrence is extremely unlikely in political science research. It is very rare when the portion of Y which overlaps $X_1$ (i.e., the portion of Y which is "explained" by $X_1$) would not also be at least partially overlapped by either $X_2$ or $X_3$.

If the relationships between the variables in our U.S. Senate example were as depicted in Figure 1, we would not even need to use multiple regression. The next sentence may be confusing. Just keep reading (the example that follows the next sentence will "clear it up")! If there is absolutely no association between any of the independent variables (which is extremely unlikely) then we would obtain the same estimates for the impact of each independent variable if we used each independent variable in a bivariate regression. For example, our "main equation" (or equation of interest) is:

(equation 1) $Y = a + b_1X_1 + b_2X_2 + b_3X_3 + e$

If there is no association at all between any of the independent variables (i.e., $X_1$, $X_2$ and $X_3$), then we could obtain the same estimates of "$b_1$," "$b_2$" and "$b_3$" from either of two methods. The first method would be to estimate equation 1. Equation 1 is a multiple regression model. Alternatively, we could estimate equations 2 through 4 ahead and we would obtain the same estimates of "$b_1$," "$b_2$," and "$b_3$" as from equation 1.

(equation 2) $Y = a_1 + b_1X_1 + e_1$

(equation 3) $Y = a_2 + b_2X_2 + e_2$

(equation 4) $Y = a_3 + b_3X_3 + e_3$

Again, if there is absolutely no relationship between any of the independent variables (i.e., no relationship between $X_1$, $X_2$ and $X_3$) then estimating equation 1 would produce the same values for "$b_1$," "$b_2$" and "$b_3$" as would estimating equations 2, 3 and 4. However, since it is extremely unlikely that there is no relationship between $X_1$, $X_2$ and $X_3$, multiple regression (i.e., equation 1) will almost invariably produce different results for the impact of each independent variable on the dependent variable (i.e., for "$b_1$," "$b_2$" and "$b_3$") than will bivariate regression (i.e., equations 2, 3 and 4). This is why we use multiple regression so frequently.

Figure 2 below depicts a far more common relationship between variables than that shown in Figure 1.

Figure 2 - Moderate Multicollinearity

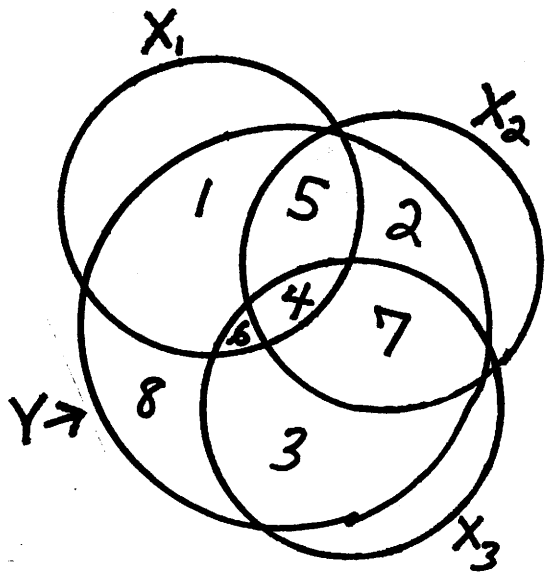



Figure 2 above represents a "moderate" degree of multicollinearity. As you can see, the three independent variables ($X_1$, $X_2$ and $X_3$) each overlap the dependent variable (Y). However, the independent variables also overlap each other. For example, portion 1 of Figure 2 shows that portion of Y which is "explained" (i.e., overlapped) only by $X_1$. If Figure 2 represented the relationships between $X_1$, $X_2$, $X_3$ and Y, the coefficient "$b_1$" (the impact of $X_1$ on Y) in equation 1 on page 117 would be estimated from portion 1 of Figure 2. Similarly, "$b_2$" would be estimated from portion 2 of Figure 2. Finally, as you probably guessed, "$b_3$" would be estimated from portion 3 of Figure 2.

Portions 4-7 of Figure 2 represent parts of Y that are explained by more than one independent variable (i.e., "multicollinear"). For example, portion 4 is that part of Y which is "explained" jointly by $X_1$, $X_2$ and $X_3$. Alternatively, portion 4 represents "multicollinearity" between $X_1$, $X_2$ and $X_3$. Portions 5-7

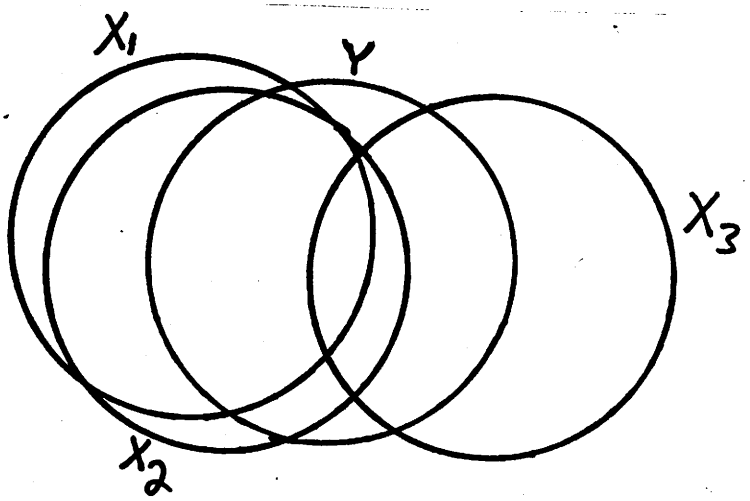
represent areas of Y which are explained by two of the three independent variables. For example, portion 5 represents that part of Y which is explained by both $X_1$ and $X_2$. Since portions 4-7 represent parts of Y that are explained by more than one independent variable, they are not used in calculating "$b_1$," "$b_2$" and "$b_3$." The computer would not know which independent variable to "credit" with explaining any section of portions 4-7. Therefore, only those areas of Y which are explained by just one independent variable (i.e., portions 1-3) are used in calculating "$b_1$," "$b_2$" and "$b_3$." This is why it is so important to include all independent variables theory suggests are important. For example, if $X_2$ were not included, then $X_1$ would receive "credit" for explaining portion 5 in Figure 2 when it really should not.

While the computer would not use portions 4-7 in calculating "$b_1$," "$b_2$" and "$b_3$," it would use portions 4-7 (along with portions 1-3) in calculating $R^2$ (i.e., the percentage of variation in Y explained by $X_1$, $X_2$ and $X_3$ together - see pages 109-110). Even though we are not sure which independent variable deserves credit for any part of portions 4-7, we do know that our independent variables have "explained" portions 4-7 of Y in Figure 2. Portion 8 is that part of Y which cannot be explained by any of the three independent variables. If $R^2$ were 1.0 (i.e., if $X_1$, $X_2$ and $X_3$ explained 100% of the variation in Y), then portion 8 would contain no area (i.e., portion 8 would not exist).

In the situation represented by Figure 2, the computer still has sufficient information to work with in order to reliably estimate the impact of $X_1$, $X_2$ and $X_3$ on Y (i.e., the value of "$b_1$," "$b_2$" and "$b_3$"). In this situation, if the computer reported a low "t ratio" (i.e., under 2.0) it would probably not be the result of multicollinearity. Rather, the independent variable may not have a noticeable impact on the dependent variable.

Figure 3 below represents a case of very high multicollinearity. This is the situation we are hoping to avoid. In a situation such as Figure 3, the computer lacks the necessary information to reliably estimate the impact of either $X_1$ or $X_2$ on Y. Hence both the coefficients of $X_1$ and $X_2$ (i.e., $b_1$ and $b_2$) are likely to have low "t ratios" (i.e., much less than 2.0). In this situation, it may well be that high multicollinearity has prevented the computer from uncovering statistically significant relationships between both $X_1$ and Y as well as between $X_2$ and Y. Put another way: if the degree of multicollinearity between $X_1$ and $X_2$ were lower, each might have been statistically significant in equation 1. Notice however, that while $X_1$ and $X_2$ are highly multicollinear, $X_3$ is not. Thus, the computer has an easy time of estimating the impact of $X_3$ on Y (i.e., of estimating $b_3$).

Figure 3 - High Multicollinearity between $X_1$ and $X_2$ but Little Multicollinearity involving $X_3$



How do we detect multicollinearity? First, let me mention that if our independent variables are statistically significant, there is no need to worry about multicollinearity. For example, if in equation 1, "$b_1$," "$b_2$" and "$b_3$" were statistically significant (i.e., each had a "t ratio" with an absolute value of 2.0, or greater), I would not worry about multicollinearity. For statistically significant independent variables, the computer obviously had enough information to conclude with a fairly high degree of certainty that there was a significant relationship between that particular independent variable and the dependent variable. Multicollinearity is only a potential problem for statistical insignificant independent variables. So, the first method of checking for multicollinearity is to see if any of the independent variables are statistically insignificant.

One almost certain situation where you will have high multicollinearity is if the $R^2$ is "high" (e.g., say .80), you have several independent variables and none of them are statistically significant. In other words, in such a situation, your model would have "explained" 80% of the variation in the dependent variable and yet none of the independent variables is thought to have a statistically significant impact on the dependent variable. Watch out! It is extremely likely that you have very high multicollinearity between the independent variables.

Let us return to equation 1. For convenience I have reprinted it below. I will sometimes refer to equation 1 as the "main equation" (i.e., the equation we are most interested in).

(equation 1) $Y = a + b_1X_1 + b_2X_2 + b_3X_3 + e$

Suppose we estimate equation 1 and find that both "$b_2$" and "$b_3$" have "t ratios" with absolute values of over 2.0. Our decision rule says that for both "$b_2$" and "$b_3$" we should reject the null hypothesis and conclude that both $X_2$ and $X_3$ have a statistically significant impact on Y. In such a situation, I would not worry about multicollinearity for either $X_2$ or $X_3$. Suppose that $b_1$ is statistically insignificant (i.e., has a "t ratio" with an absolute value below 2.0). Before concluding that my hypothesis is "wrong" (i.e., that $X_1$ has no effect on Y), I would check for multicollinearity. My first approach would be to use what I will refer to as the "explained variance test" for multicollinearity. The next sentence may be confusing. Just keep reading (you will find an easy to follow example)! In the "explained variance test" for multicollinearity, we regress each independent variable that is statistically insignificant in the "main equation" (e.g., $X_1$ in equation 1) on all the other independent variables and assess how much variance in each statistically insignificant independent variable is explained by the other independent variables. In our case this would mean to execute equation 5 below.

(equation 5) $X_1 = a_4 + b_4X_2 + b_5X_3 + e_4$

I used the letters "$b_4$" and "$b_5$" so you would be sure not to confuse them with "$b_1$," "$b_2$" and "$b_3$" from equation 1. The logic of the "explained variance test" (i.e., equation 5) is to see how much of the variation of a statistically insignificant variable in the "main equation" (equation 1) is explained by all the other independent variables in the "main equation" (equation 1). Let us say that we estimate equation 5 and find that the $R^2$ for equation 5 is .85. This would tell us that 85% of the variation in $X_1$ was explained by variation in $X_2$ and $X_3$. This indicates a level of multicollinearity similar to what was pictured for $X_1$ in Figure 3.

In such a situation, high multicollinearity is probably the reason that $X_1$ was statistically insignificant in equation 1.

On the other hand, suppose that the $R^2$ in equation 5 is only .25. This would mean that only 25% of the variation in $X_1$ is explained by variation in $X_2$ and $X_3$. This is too low a figure to say that if $b_1$ is insignificant in equation 1, multicollinearity is the likely reason. More probable reasons for the insignificance of $b_1$ are: (1) our hypothesis is incorrect (i.e., $X_1$ really doesn't influence Y); (2) we have committed a "type II" error (i.e., the null hypothesis is incorrect but because of our "decision rule" we choose not to reject the null hypothesis); (3) our measure of $X_1$ is not a valid measure of the trait we are trying to measure (thus we haven't really tested the hypothesis); or (4) we are lacking another independent variable (or two or three) that we should have included and by excluding them are hiding a "true" relationship that does exist between $X_1$ and Y.

Other than perhaps finding a more valid measure for the concept that $X_1$ is suppose to represent (in our case the senator's conservatism) or rethinking our model (i.e., concluding that we omitted independent variables that theory suggests we should have included in the first place), there is probably not much we can do. Our decision is not to accept the null hypothesis that $X_1$ does not influence Y as true. Rather, all we can say is that the evidence against the null hypothesis was not sufficiently strong to warrant rejecting the null hypothesis in favor of the alternative hypothesis that $X_1$ does influence Y.

Let me propose the following "standard" for using the "explained variance test" for multicollinearity: if the $R^2$ in an "auxiliary regression" (i.e., an equation where an insignificant independent variable in the "main equation" is now the dependent variable - such as equation 5 above) is above .70, we probably have a sufficiently high degree of multicollinearity (such as between $X_1$ and $X_2$ in Figure 3 on page 119) to conclude that multicollinearity is the likely reason the independent variable ($X_1$) is insignificant in the "main equation" (e.g., equation 1); if the $R^2$ in an "auxiliary regression" is between .40 and .70, we probably have a moderate degree of multicollinearity (as in Figure 2 on page 118) but not high enough to conclude it is the likely reason the particular independent variable (e.g., $X_1$) is statistically insignificant in the "main equation" (e.g., equation 1); if the $R^2$ in an "auxiliary regression" is below .40 multicollinearity is not nearly high enough to use as a reason for the insignificance of the particular independent variable (e.g., $X_1$).

A second test for multicollinearity is the "deletion test." The next sentence may be confusing. Just keep reading (an example will follow)! The logic of the "deletion test" is if any independent variable that is statistically insignificant in the "main equation" (e.g., $X_1$ in equation 1) becomes significant if we "delete" one other independent variable and re-estimate the "main equation," then multicollinearity is probably the reason. Let us use the "deletion test" in our current situation. Again, let me reprint equation 1 below.

(equation 1) $Y = a + b_1X_1 + b_2X_2 + b_3X_3 + e$

As previously, let us assume that "$b_1$" is statistically insignificant (i.e., has a "t ratio" of less than 2.0) when equation 1 is estimated. There are three independent variables in equation 1 ($X_1$, $X_2$ and $X_3$). Since only one of them is statistically insignificant ($X_1$), we would need to run two auxiliary equations to

perform the "deletion test." Specifically, equations 6 and 7 below re-estimate equation 1 deleting a different independent variable other than $X_1$.

(equation 6) $Y = a_5 + b_6X_1 + b_7X_2 + e_5$

(equation 7) $Y = a_6 + b_8X_1 + b_9X_3 + e_6$

Notice that equation 6 has the same variables as equation 1 (the "main equation") except that $X_3$ is omitted in equation 6. Further note that equation 7 has the same variables as equation 1 except that $X_2$ is omitted. Thus, what I have done is "re-estimate" equation 1 "deleting" one independent variable (other than $X_1$) from each of the "re-estimates." The question becomes: Is $X_1$ statistically significant in either equations 6 or 7? That is, does either "$b_6$" (in equation 6) or "$b_8$" (in equation 7) have a "t ratio" with an absolute value equal to 2.0, or greater? If so, multicollinearity is probably the reason that "$b_1$" in equation 1 (i.e., the "main equation") is statistically insignificant (although we can not be completely sure). For example, suppose that "$b_8$" in equation 7 is statistically significant. This would probably mean (although we do not know for sure), that $X_1$ and $X_2$ greatly overlap each other in the portion of Y that they each "explain" (as we saw in Figure 3 on page 119). In such a situation, if $X_2$ is removed (as in equation 7), $X_1$ might become statistically significant in explaining variation in Y. Just picture Figure 3 on page 119 without $X_2$. If neither "$b_6$" in equation 6 or "$b_8$" in equation 7 is statistically significant, then multicollinearity is probably not the reason that "$b_1$" is statistically insignificant in equation 1.

In some cases an independent variable that is highly related to the other independent variables is still statistically significant in the "main equation" (e.g., equation 1). For example, I have found in models similar to equation 1 that 75% or more of the variation in senatorial conservatism ($X_1$) is "explained" by a senator's political party affiliation and the median family income of the senator's state. Thus, in an equation such as equation 5 on page 120, the $R^2$ was likely to be about .75 However, in the vast bulk of equations like equation 1, senatorial conservatism is almost always statistically significant. Why? To answer this, think of multicollinearity as analogous to a reduction in sample size. Thus, if 75% of the variation in senatorial conservatism is explained by political party affiliation and state median family income, it is like reducing the sample from 100 senators to perhaps 10 or 15. Now think of tossing a coin. If you tossed a coin 10 times and heads came up all 10 times, what would you conclude? Since the chance of a fair coin flipping 10 consecutive heads is less than 1 in 1,000, my guess is you would conclude that the coin is not fair. Even though you had a small sample size (i.e., 10 flips) the results were so contrary to what the null hypothesis would be (5 heads and 5 tails) that you reject the null hypothesis. It is the same in the situation I am depicting. Differences in senatorial conservatism are typically so strongly related to how senator's vote (whether on taxes, defense spending, etc.) that even in the face of high multicollinearity the results are almost always statistically significant. As I mentioned previously, if the independent variable is statistically significant in the "main equation" (e.g., equation 1) forget about multicollinearity.

What can we do if high multicollinearity is causing one or more of the independent variables to be statistically insignificant? Although I will mention some alternatives, the basic answer is "not much." If you are writing an article for other political scientists and you have high multicollinearity, you often end up mentioning the results of the "explained variance" and "deletion" tests in a footnote and living with the consequences.

One possible remedy for high multicollinearity is to gather more data. If the sample size increases, perhaps you can obtain enough information for the computer to produce statistically significant results. The reason this option is typically of little value is that presumably the political scientist would have used all the observations they had when they initially estimated the results. For example, suppose instead of using 100 senators, we had only used 70 senators when we estimated equation 1. We could then increase the sample size to 100 and re-estimate equation 1. Perhaps the multicollinearity that was present with 70 senators would be noticeably reduced with 100 senators. However, this option is usually of no value because a political scientist would not have a reasonable defense for only using 70 senators if all 100 were available. Remember, a larger sample provides more accurate estimates of the coefficients (i.e., the "b's") than a smaller sample. Therefore, we would not estimate equation 1 with just 70 senators. Thus, a political scientist would have used all 100 senators when equation 1 was initially estimated. So, if high multicollinearity was a problem when all 100 senators were used, there would be no additional observations (i.e., senators) to add.

If two independent variables are highly correlated with each other, a second possible remedy for high multicollinearity is to combine these two independent variables into one variable. Thus, we might be able to reduce the number of independent variables. However, typically there is not any logical method of combining two or more independent variables to form one independent variable. Hence, this possibility is not often useful.

A final possible response to high multicollinearity is to simply delete one of the independent variables. For example, we could claim that equation 1 should only have contained the senator's conservatism ($X_1$) and state median family income ($X_3$). If political party affiliation ($X_2$) was causing senatorial conservatism to be statistically insignificant in equation 1, eliminating political party affiliation might well "solve" the problem. However, the "cure" is probably worse than the disease. If there was not a strong theoretical reason for including political party affiliation, it should never have been in equation 1 to begin with. Assuming that the theoretical justification for including political party affiliation is strong, it did not suddenly grow weak because of multicollinearity.

Additionally, as mentioned on page 119, if we delete an independent variable (e.g., $X_1$) then the values of the "b's" for the remaining independent variables (e.g., $b_2$ - the coefficient for $X_2$ and $b_3$ - the coefficient for $X_3$) will likely change (just keep reading). Thus, if the model we should estimate is equation 1: $Y = a + b_1X_1 + b_2X_2 + b_3X_3 + e$ but due to high multicollinearity between $X_1$ and $X_2$ we omit $X_1$ and instead estimate the following equation: $Y = a + b_2X_2 + b_3X_3 + e$, we will get different values for both "$b_2$" and "$b_3$" than we would from equation 1. Omitting $X_1$ will produce different, and probably biased, estimates for "$b_2$" and "$b_3$." Generally speaking, it is a mistake to try to eliminate the effects of high multicollinearity by re-specifying our model to eliminate an independent variable that theory indicates should be included just to make the remaining independent variables statistically significant.

## Dummy Variables

Political scientists often use variables that are qualitative and not quantitative. For example, international relations scholars who study conflict resolution often have a variable in their model concerning whether or not particular nations involved in a dispute have or do not have nuclear weapons. Such a variable is usually scored either "1" (if the nation has nuclear weapons) or "0" (if the nation does not have nuclear weapons). Variables that measure a qualitative difference (such as the presence or absence of nuclear weapons) but do not provide mathematical precision are often termed "dummy" variables. To continue with the example of nuclear weapons, how could such a variable be made quantitative? I am not sure. I do not know if we could construct a meaningful scale for a variable such as nuclear weapons. For example, although we could measure the destructive capability of a nation's nuclear arsenal, it might be irrelevant. If you were the leader of a nation, would it really matter to you if your adversary could demolish your nation 20 or 45 times? I doubt it. The more important question is probably whether your adversary does or does not have nuclear weapons. For this reason, nuclear weapons is usually treated as a two category (i.e., dichotomous) dummy variable in the international relations literature. Concerning non-nuclear weapons, the international relations literature often uses "balance of power" ratios and other measurement approaches that do take into account differences in degree (i.e., are ratio level measures - see page 11). Since differences in degree likely matter in non-nuclear weapons (such as the type of forces, location of forces, etc.) such measurement makes sense.

We have already used a dummy variable. In the U.S. Senate example we have previously studied, a senator's political party affiliation was measured with a dummy variable. In this example Democrats (members of the good party!) were coded as "1" and Republicans as "0." When this example was used you might have thought, what difference would it make if Republicans were coded "1" and Democrats were coded "0"? The answer is that as long as we know the coding scheme, it makes no difference at all. Let me show you. In this example, I will use one of the data sets typically used in Assignment #2 in POSC 300B. To simplify the presentation, I am just going to use one independent variable, the senator's party affiliation. In the first example, Democrats are coded "1" and Republicans are coded "0." The dependent variable is the percentage of times the senator voted for tax changes favorable to families with incomes equal to or less than the median family income. The results are as follows: a= 24.657 b = 35.293

Remember, the prediction equation is:

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

For a Democrat [i.e., score on X (political party affiliation) is "1"] this would mean:

$$\hat{Y} = 24.657 + [(35.293)(1)]$$

$$\hat{Y} = 24.657 + [35.293]$$

$$\hat{Y} = 59.95$$

So, the average Democrat would be expected to support tax changes favorable to families with incomes equal to or less than the median income approximately 60% (i.e., 59.95%) of the time. For a Republican [i.e., score on X (political party affiliation) is "0"] this would mean:

$$\hat{Y} = 24.657 + [(35.293)(0)]$$

$$\hat{Y} = 24.657 + [0]$$

$$\hat{Y} = 24.657$$

So, the average Republican would be expected to support tax changes favorable to families with incomes equal to or less than the median income approximately 24.7% (i.e., 24.657) of the time.

Next I changed the coding scheme so that Republicans were coded "1" and Democrats were coded "0." I re-estimated the equation and here are the results: a = 59.95 b = -35.293

So, for a Republican (X = 1) the prediction becomes:

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

$$\hat{Y} = 59.95 + [(-35.293)(1)]$$

$$\hat{Y} = 59.95 + [-35.293]$$

$$\hat{Y} = 24.657$$

For Democrats (X = 0) the prediction becomes:

$$\hat{Y} = 59.95 + [(-35.293)(0)]$$

$$\hat{Y} = 59.95 + [0]$$

$$\hat{Y} = 59.95$$

<u>As you can see, changing the coding scheme did not effect the final result</u>. Regardless of whether Democrats were coded "1" or "0" they still, on average, vote almost 60% (59.95%) of the time in a direction favorable to those with family incomes equal to or less than the median. Republicans, regardless of whether they are coded "1" or "0" vote, on average, almost 25% (24.6575) of the time in a direction favorable to those with family incomes equal to or less than the median.

In the previous example of political party affiliation we were dealing with a dichotomous (i.e., two category) variable. Thus, all senators were either Democrats or Republicans. However, occasionally political scientists use qualitative variables that have more than two categories. For example, scholars in comparative politics often code the selection of the chief executive of a regime as: (1) unregulated - the chief executive comes to power by force; (2) transitional - the chief executive is chosen by the political elite and without formal competition; or (3) regulated - the chief executive is either elected or comes to power through hereditary succession (i.e., the son of a King becomes King after his father dies).

Recalling the levels of measurement (pages 10-11), the classification of selection methods for the chief executive discussed above is clearly "ordinal." The codes fit a definite continuum from "least regulated" (category 1) to most regulated

(category 3). However, we can not say that the difference between categories is the same. For example, is the difference between category 1 (unregulated) and category 2 (transitional) the same as the difference between category 2 (transitional) and category 3 (regulated)? We simply can not say. Given this, we probably have an ordinal level of measurement. How do we have the computer estimate the impact of the selection method for the chief executive?

Since we have three categories of responses on the variable measuring the selection method for the chief executive, a comparative politics scholar would likely create two dummy variables. The first dummy variable might be called "transitional" and would be coded "1" if the nation used a "transitional" method of selection for the chief executive and "0" if it did not (i.e., if the nation used either the "unregulated" or the "regulated" method it would be coded as "0"). The second dummy variable might be called "regulated" and would be coded "1" if the nation used the "regulated" method of selecting a chief executive and "0" if the nation did not (i.e., if the nation used either the "unregulated" or "transitional" methods it would be coded "0"). With these two dummy variables we have covered all three possibilities. For example, nations using an "unregulated" method would score "0" on both dummy variables, nations using a "transitional" method would score "1" on the first dummy variable and "0" on the second dummy variable and nations using a "regulated" method would score "0" on the first dummy variable and "1" on the second dummy variable.

You were thinking of asking: Since we have three categories of responses (unregulated, transitional and regulated), why don't we use three dummy variables? The answer is because if we did we would have perfect multicollinearity between the variables. Thus, if a nation scored "0" on the transitional dummy variable and also scored "0" on the regulated dummy variable, wouldn't the nation have to be using an "unregulated" selection method? Yes!! If not, the nation would have scored "1" on either the "transitional" or the "regulated" dummy variables. So, if we created a third dummy variable (1 = unregulated, 0 = either transitional or regulated) it would be perfectly related to the other two dummy variables. Once the computer knew a nation's score on the transitional and the regulated dummy variables, it could perfectly predict the nation's score on the unregulated dummy variable. This would be perfect multicollinearity. In such a situation, the computer could not estimate the impact of any of the three dummy variables. So, as a rule: always use one less dummy variable than the number of categories of responses in the trait you are trying to measure.

For example, when we measured the political party affiliation of a senator there were two possible categories of responses (Democrat or Republican) and how many dummy variables did we use? One!! If we had used one dummy variable for Democrats (1 = Democrat, 0 = Republican) and one for Republicans (1= Republican, 0 = Democrat) we would have had perfect multicollinearity. The computer would have known that if a senator scored "0" on being a Democrat, they would have to have scored "1" on being a Republican. So, if we would have used two dummy variables for party affiliation, they would have been perfectly correlated at the -1.0 level (i.e., the higher the senator's score on being a Democrat the lower their score on being a Republican, see page 75). This would be perfect multicollinearity. Hence, the computer could not have estimated the impact of party affiliation of the senator on voting to shift the tax burden to higher income groups. In order to avoid this, we used only one dummy variable to measure party affiliation.

## The Relative Importance of the Independent Variables - Standardized Coefficients

It is often useful to assess the relative importance of the independent variables. For example, it would be useful to be able to say that a senator's conservatism is twice as important as the median family income of his state in determining how willing the senator is to support shifting the tax burden to higher income earners. Unfortunately, because the variables are measured in different units (conservatism is a percentage and state median family income is in thousands of dollars), we can not directly compare their relative importance. Just keep reading!!

For example, let us revisit the U.S. Senate voting example that we have been using. The variables are as follows: Y is the percentage of times the senator voted to shift the tax burden more to high income earners; $X_1$ is the senator's degree of conservatism (from 0 - least conservative to 100 - most conservative); $X_2$ is the senator's political party affiliation (Democrat = 1, Republican = 0); and $X_3$ is the median family income in thousands of dollars in the senator's state (i.e., a score of 22.3 would mean that the median family income in that particular state was $22,300).

Using the variables discussed above, suppose we estimate the following equation:

$$Y = a + b_1X_1 + b_2X_2 + b_3X_3 + e$$

Let us further suppose that among the results were the following:

$b_1 = -.899$ $\quad$ $b_2 = 5.750$

Thus, $b_1$ is the coefficient for the senator's degree of conservatism and $b_2$ is the coefficient for the senator's political party affiliation. Which variable is more important in explaining how willing the senator is to support shifting the tax burden to higher income earners, conservatism or political party affiliation? Looking at the results, our temptation is to say, political party affiliation. Why? Because the coefficient for political party affiliation (5.750) is so much larger than the coefficient for conservatism (-.899). However, since conservatism and political party affiliation are measured in different units (i.e., conservatism is a percentage and political party affiliation is a 0 or 1 - thus, dichotomous - dummy variable) we should not jump to such a hasty conclusion. What we need to do is to adjust both $b_1$ and $b_2$ so that they are comparable.

Remember when we discussed standard scores (i.e., "Z" scores) on pages 25-27? Wasn't the idea to divide the difference between a particular score and the mean score by the standard deviation? Yes! This procedure "standardized" the difference between a particular score and the mean score because it placed this difference relative to the standard deviation for that particular variable. We can use a very similar procedure to "standardize" "b."

The results (i.e., the b's) that the computer generates are typically referred to as "unstandardized" regression coefficients. Thus, the definition that we have used for "b" has been for "unstandardized b." Before I give you the formula for "standardizing b," let me mention that if you do not grasp the formula, just keep reading!! I will provide an example that will "clear it up." "Standardized b" is sometimes referred to in political science and statistical literature as "beta." The

formula to "standardize b" is as follows:

$$standardized\ b = (unstandardized\ b)\left(\frac{standard\ deviation\ of\ x}{standard\ deviation\ of\ y}\right)$$

Verbally, the formula above says that in order to "standardize" b, take "unstandardized" b and multiply it by the ratio of the standard deviation of that independent variable to the standard deviation of the dependent variable. Just keep reading!!! Let us take an example. From the previous page, we know that the unstandardized coefficient for conservatism (i.e., $b_1$) is -.899 and the unstandardized coefficient for political party affiliation is 5.750. In order to use the formula above we need the standard deviation for each of our variables. Suppose the standard deviations are as follows: standard deviation of conservatism = 31.5; standard deviation of political party = .5; standard deviation of willingness to shift the tax burden to higher income earners = 25.7. Putting the unstandardized coefficients and standard deviations into the formula above would yield the following the following standardized coefficient for conservatism:

[(-.899)] [(31.5/25.7)] = [(-.899)] [(1.23)] = -1.106

In case you just had trouble, it is -.899 multiplied by 1.23 which equals -1.106. For political party affiliation the calculation is as follows:

[(5.750)] [(.5/25.7)] = [(5.750)] [(.019)] = .109

Now let us interpret the above results. Since the coefficients (i.e., the b's) have been standardized we can compare each one relative to the absolute size of the other. Thus, I am going to compare 1.106 (not -1.106) to .109. The direction (i.e., positive or negative sign) is not important here. At this point, we already know whether each independent variable is positively or negatively associated with the dependent variable. What we want to know is their relative importance. So, 1.106/.109 = 10.15. We can say that a senator's conservatism has a little more than 10 times the impact of the senator's political party affiliation on the senator's willingness to shift the tax burden to higher income earners. The relative comparison that I just made is the interpretation of standardized coefficients most commonly used by political scientists. There are several other interpretations of standardized coefficients, but they are rarely used in political science. So, I will not "bore" you with them.

Several notes of caution are in order. First, an implicit assumption of standardized coefficients is that each variable has the same opportunity to change by one standard deviation. In the instance of political party this is untenable. Given that senators are either Republicans or Democrats (coded as "0" or "1"), how could a senator's party affiliation change by a fraction of a unit (the standard deviation for political party is .5)? Obviously, it could not. For conservatism this is not a problem. Second, some political scientists do not believe in using standardized coefficients. The "crux" of the argument against using standardized coefficients is that since the standard deviations of

both X and Y, which are both used in the standardization formula (see formula on the top of page 128), are specific to the particular sample used in the study, the results from standardizing coefficients will vary from sample to sample and, hence, will not be valid (just keep reading - it will become clear).

For example, suppose we did a study using the same variables as in Assignment #2 in POSC 300B (i.e., conservatism, party affiliation and state median family income explaining senatorial votes on tax legislation) for two different senates (e.g., 1995 and 1996). Since the conservatism scores change every year, the standard deviation of the conservatism variable will likely be different in each year. If "$b_1$" (i.e., the unstandardized coefficient for conservatism) is the same in both years (e.g., -.560) but the standard deviation is different in both years, then the results from applying the formula on page 128 for standardizing "$b_1$" will be different for the two senates (e.g., the standardized "$b_1$" might be -.800 in 1995 but -.650 in 1996). Thus, the unstandardized coefficient might be the same in each year (e.g., -.560), which indicates that conservatism has the same impact on senatorial voting on tax changes favorable to those with incomes at, or below, the median income in both 1995 and 1996. However, the standardized coefficients would likely be different (e.g., -.800 vs. -.650). Let me also mention that since the standard deviation of Y (in our example, the standard deviation of the percentage of times the senator voted in favor of tax changes favorable to those with incomes at, or below, the median family income), which is also part of the standardization formula on page 128, would also likely change from one year to the next, this could also effect the results from standardizing coefficients.

For this reason, some political scientists recommend against using standardized coefficients. I should also mention that since $R^2$ also uses the standard deviations of both the independent and dependent variables, it is also vulnerable to the same problem mentioned above (i.e., that the results from different samples might not be comparable).

Although I think the critics of standardized coefficients and $R^2$ are correct, since standardized coefficients are occasionally used and $R^2$ is frequently used in political science, I will ask you to calculate and interpret them for Assignments #2 and #3 in POSC 300B, the term paper in POSC 550 and quizzes in both courses (you have already had at least one quiz asking you to interpret $R^2$).

Finally, as I mentioned above, standardized coefficients are not used that frequently in political science. Probably 95% or more of the regression coefficients you see in political science research are unstandardized. That is why when I introduced you to regression coefficients (i.e., "b") I did not mention standardized coefficients. If I give you a value for "b" on a quiz, always assume it is unstandardized and interpreted as on pages 109-110. If I gave you a standardized coefficient on a quiz I would specifically mention that it was standardized. Thus, if nothing to the contrary is stated on a quiz or the final examination, always assume that the coefficient (i.e., "b") is unstandardized.

Potential quiz questions include: (1) What is multicollinearity? (Be ready to explain what multicollinearity is in words, not a diagram.) (2) Why are political scientists concerned about multicollinearity? (3) Explain the logic of the "explained variance" and "deletion" tests for multicollinearity. (That is, how do the tests operate?) (4) Define "dummy variable." (Don't define it by just giving an example.) (5) Under what circumstances would a political scientist use a standardized, instead of an unstandardized, regression coefficient?