Chapter seventeen

Multivariate Regression: Additional Topics

Non-linear relationships

What are They?

How are They Included in Regression Models?

How are They Interpreted?

Interaction effects

What are They?

How are They Included in Regression Models?

How are They Interpreted?

Multicollinearity

When Does It Arise?

How is It Diagnosed?

How is It Treated?

D. Weisburd and C. Britt, *Statistics in Criminal Justice*, DOI 10.1007/978-1-4614-9170-5_17,

In the previous chapter we extended the bivariate regression approach by showing how we could include multiple independent variables simultaneously in a single model. We illustrated how we incorporate variables measured not only at the interval level of measurement, but also nominal and ordinal independent variables, into a regression model. While the models we have examined so far allow us to approximate linear relationships between various independent variables and the dependent variable, in the real world we are sometimes confronted with more complex research questions that require us to make additional modifications to our model.

For example, in the OLS regression model, our interpretation of the coefficients is based on the notion that there is a linear relationship between the independent and the dependent variable. But what if we find evidence of a curvilinear relationship? Or, theory suggests that there may be a non-linear relationship between two variables? Although the OLS regression model is based on the assumption of a linear relationship between the dependent and each of the independent variables, **non-linear relationships** can be incorporated into an OLS regression model in a straightforward manner.

Another issue in the application of OLS regression is that the interpretation of the coefficients is based on the idea that each independent variable has a constant effect irrespective of the levels of other independent variables. For example, if we include a dummy variable we assume that the effect of every other independent variable is the same for men and women. But what if there was a good theoretical or policy reason to suspect that the effect of some variable was different for men and women? How would we incorporate that into our model? In the statistical literature, these are known as **interaction effects**, and they allow us to test whether the effects of specific independent variables in a regression model vary by the level of other independent variables.

In this chapter we also introduce an important problem that researchers sometimes face when estimating multivariate regression models. We have emphasized so far that researchers must include all relevant independent variables in a model if it is to be correctly specified. Correct model specification, in turn, is necessary to avoid bias in regression models. But sometimes the inclusion of multiple independent variables can lead to a problem we term **multicollinearity** which is likely to lead to estimation of unstable regression coefficients. Multicollinearity refers to the situation where independent variables are very highly correlated with each other, which then makes it very difficult for OLS regression to determine the unique effects of each independent variable.

Non-linear Relationships

Policy-oriented research focused on the severity of punishment for convicted offenders illustrates that as the severity of an offender's prior record and the severity of the conviction offense increase, the severity of the punishment tends to increase.[1] In the interpretation of OLS regression results, we would say something, for example, about how each one unit increase in the severity of an offender's prior record results in the length of a sentence increasing by some fixed time period (e.g., 8 months). Key to the interpretation of OLS regression coefficients is the idea that the level of the independent variable does not matter – each unit change is expected to result in the same change in the dependent variable regardless of whether we are looking at small or large values on the independent variable. What if this is not always the case?

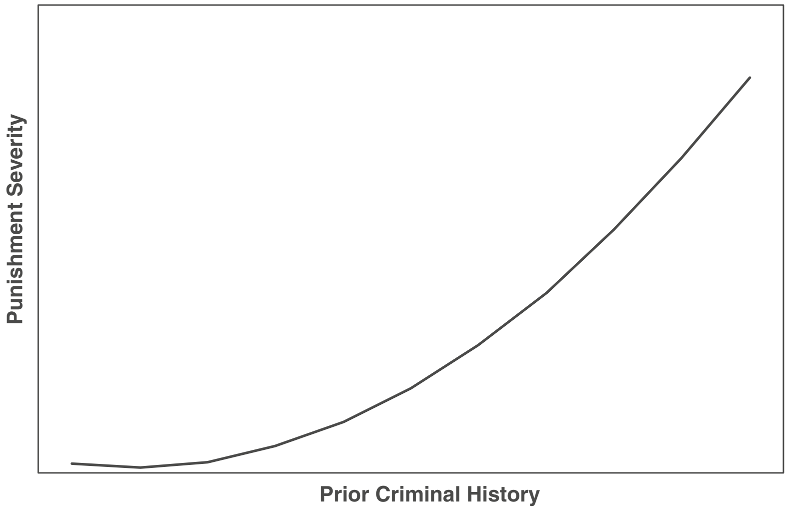
For example, an assumption often made in research on sentencing outcomes, but rarely examined, is the idea that first-time offenders (i.e., those with no prior record) or those offenders convicted of relatively minor forms of crime will be punished much more leniently than other offenders. Then, as the severity of prior record or of conviction offense increase, there is an expectation of an increasingly punitive response by the criminal justice system. Put another way, there is an expectation of a non-linear relationship between the severity of the conviction offense or the offender's prior record and the severity of punishment – changes in the level of the dependent variable may vary by the level of the independent variable. Figure 17.1 presents a hypothetical plot for punishment severity and prior criminal history that reflects increasingly harsher punishments for offenders with more extensive criminal records.

As can be seen in the figure, there is a gradual increase in the severity of the punishment as the severity of the prior record increases. Then, the increases in the severity of the punishment become larger for the same unit increase in prior criminal history.

The range of potential non-linear relationships is limitless and is bounded only by the imagination and creativity of the researcher and the theoretical basis for conducting the research. Yet, while there may be a wide range of possible non-linear relationships, most researchers will confine their analyses to a relatively limited group of non-linear possibilities,

[1] Michael H. Tonry, *Sentencing Matters* (New York, Oxford University Press, 1996).

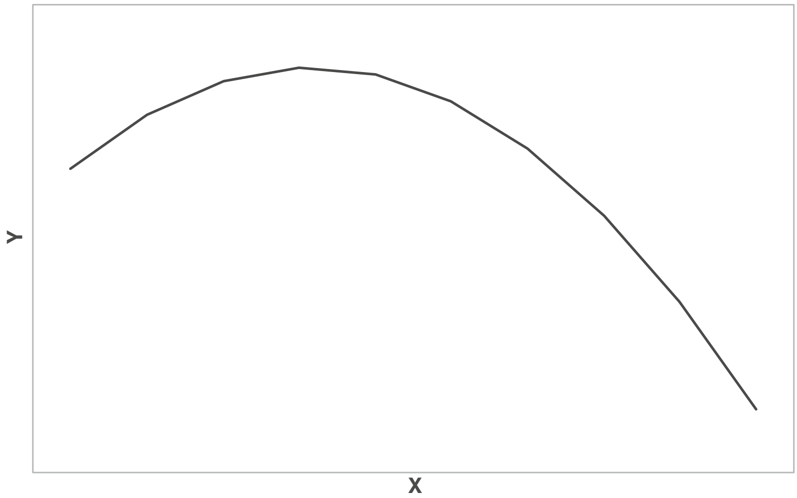
Figure 17.1 *Hypothetical Non-linear Relationship Between Punishment Severity and Prior Criminal History*



some of which are displayed in Figure 17.2. Panel (a) presents what is referred to as a quadratic equation. All that this means is that a squared term has been added to the equation to give it a form such as Y = X + X^2. The quadratic equation is one of the more commonly used transformations in criminology and criminal justice, and has had frequent application in the study of age-related behavior.

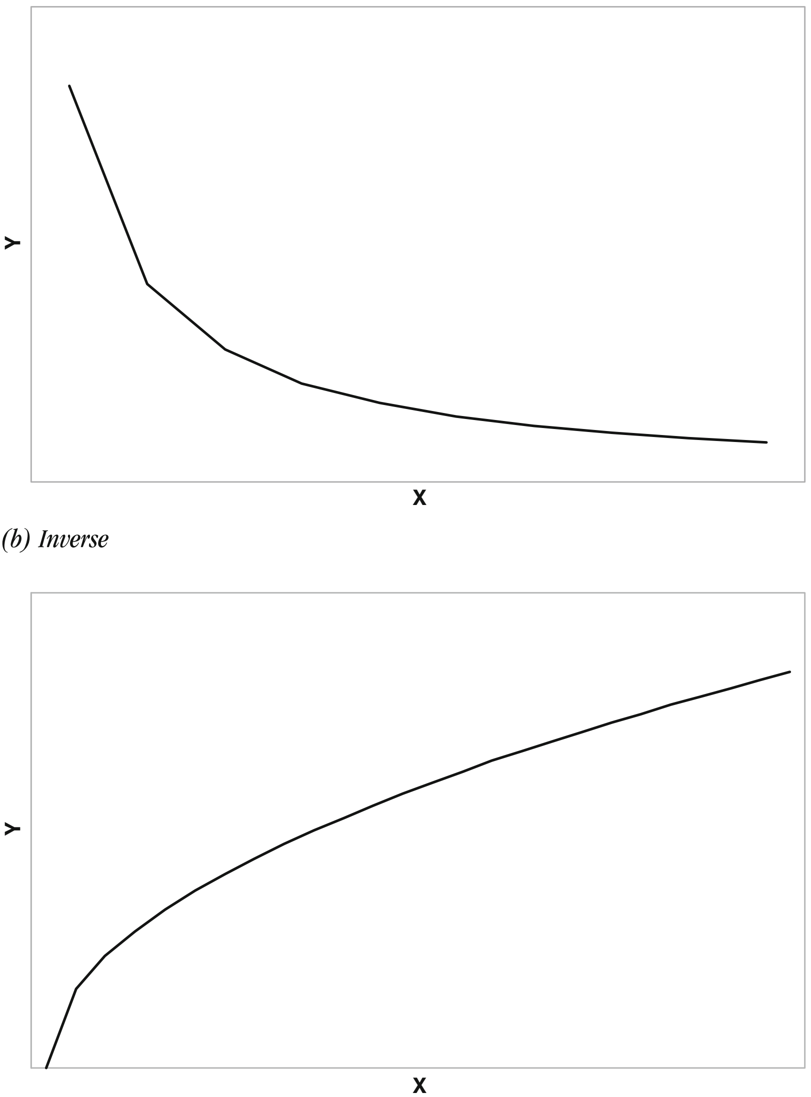
Panel (b) presents an inverse function of the form Y = 1/X. This kind of transformation helps to capture relationships where there is a decreasing negative effect of the independent variable on the dependent variable.

Figure 17.2 *Common Non-linear Relationships Used in Criminal Justice Research*



(a) Quadratic

Figure 17.2 *Continued*



(b) Inverse

(c) Square Root

Panel (c) presents a square root transformation of the form Y = $\sqrt{X}$. This kind of transformation is useful when there is a diminishing positive impact of the independent variable on the dependent variable.

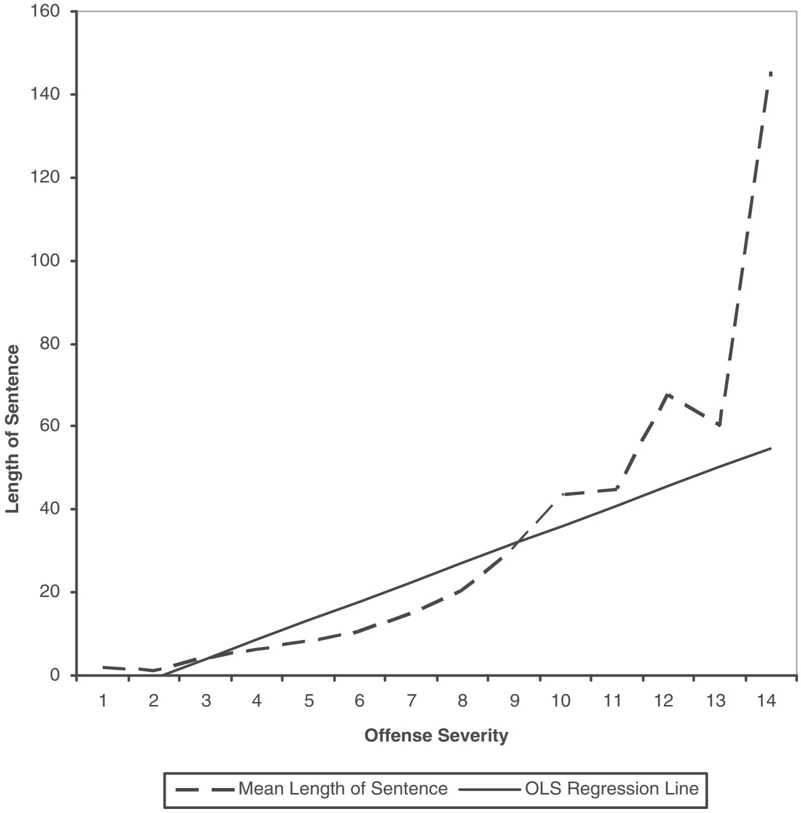
Finding a Non-linear Relationship: Graphical Assessment

Perhaps the most straightforward way of exploring data for a non-linear relationship is to use a line graph (see Chapter 3). A simple scatterplot (discussed in Chapter 14) will often contain so many data points that it is difficult, if not impossible, to discern any pattern in the data. A line graph that plots the mean of the dependent variable against the value of the independent variable will likely provide a rough indication of the nature of the bivariate relationship between the two variables. For example, Figure 17.3 presents the mean for length of sentence against the severity of the conviction offense for over 20,000 offenders sentenced in Pennsylvania in 1998.[2] As you look at Figure 17.3, you can see that there is a gradual, linear increase in length of sentence as offense severity increases to about level 6 to 7. At that point, the increases in sentence length become larger for each additional increase in offense severity. To highlight the curvilinear nature of the relationship between offense severity and length of punishment, the OLS regression line for these data is overlayed in Figure 17.3, indicating that the straight-line relationship does not capture the relationship between length of sentence and severity of offense particularly well.

Incorporating Non-linear Relationships into an OLS Model

Assuming that we have good reason for assuming that a non-linear relationship exists between the dependent variable and one or more of the independent variables, how do we incorporate this information into

Figure 17.3 *Plot for Mean Length of Sentence by Offense Severity for Offenders in Pennsylvania*



[2] These data are available through the National Archive of Criminal Justice Data and can be accessed at http://www.icpsr.umich.edu/NACJD

the OLS regression model? The first step, as noted above, is to try and gain a sense of the relationship graphically. In most circumstances, if theory suggests or if we find evidence of a curvilinear relationship, the most straightforward approach is to add a quadratic term – the squared value of the independent variable – such as that in Panel (a) of Figure 17.2. More formally, a quadratic regression equation would have the following form:

$$Y = b_0 + b_1X_1 + b_2X_1^2$$

where Y represents the dependent variable
X_1 represents the independent variable

In our example presented in Figure 17.3, we have evidence of a curvilinear relationship that might be accounted for by adding a squared term for offense severity to a regression equation. We begin by noting that the OLS regression line portrayed in Figure 17.3 is

$$Y = -9.85 + 4.62\ X_1$$

where Y represents length of sentence (in months)
X_1 represents offense severity

To incorporate a non-linear relationship, we begin by transforming the variable – in this case offense severity – and then add this transformed variable to the regression equation. In most statistical software packages this would simply involve the creation of a new variable which represents the original variable squared. When we square offense severity and add it to the regression equation, we obtain the following equation:

$$Y = b_0 + b_1X_1 + b_2\ (X_1 * X_1) = b_0 + b_1X_1 + b_2X_1^2.$$

If we then estimate this new regression equation that includes both the original measure of offense severity and the squared value of offense severity, we obtain the following results:

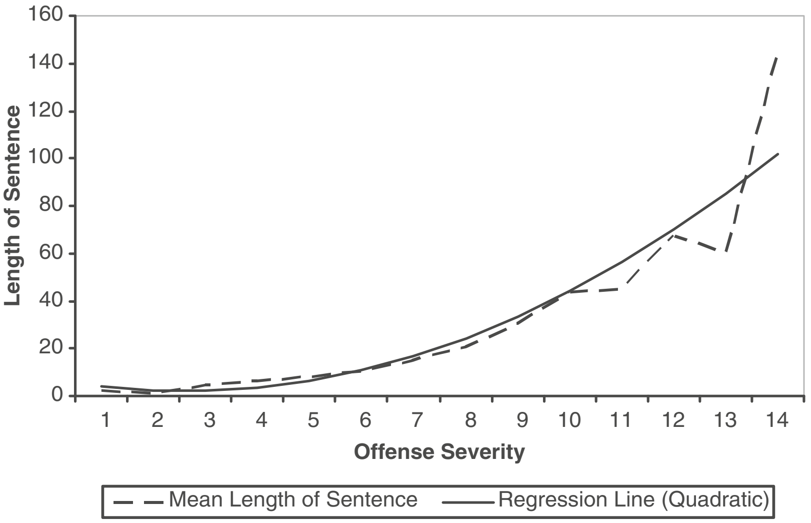
$$Y = 7.07 - 3.99\ X_1 + 0.77\ X_1^2$$

Substantively, this regression equation captures the curvilinear relationship between offense severity and sentence length much better than a straight-line relationship, since there are increasingly larger increases in sentence length for each unit change in offense severity. Figure 17.4 presents the mean sentence length by offense severity (similar to that in Figure 17.3) along with the new regression line based on including the quadratic term.

Interpreting Non-linear Coefficients

In many practical applications of adding non-linear terms to OLS regression models, there is often less emphasis on the interpretation of the individual coefficients that represent transformed variables. The reason for

Figure 17.4 *Plot for Mean Length of Sentence by Offense Severity for Offenders in Pennsylvania with Quadratic Regression Line*



this is the difficulty in making sense of individual coefficients. For example, in the example using the data on offenders sentenced in Pennsylvania, the coefficient for the squared value of offense severity is given as 0.77. Like any other OLS coefficient, this coefficient can be interpreted in the context of one unit changes of the *transformed* variable:

> For each one unit increase in the *squared value* of offense severity, the length of sentence is expected to increase by 0.77 months.

Unfortunately, this kind of interpretation tends not to have much intuitive meaning for most people – researchers and others. Consequently, the description of results using non-linear transformations in OLS regression models will often focus on the general pattern of results, rather than on the specific coefficients. This is not entirely satisfactory, however, because it still leaves the reader wondering whether the transformation added much to the analysis. Our suggestion is to use graphs, such as that presented in Figure 17.4, which do provide an effective way to convey evidence of a non-linear relationship between the dependent and independent variables. What makes this kind of plot particularly useful is that it conveys both the pattern in the observed data and the predicted values based on the estimated regression model.

Note on Statistical Significance

Estimating statistical significance for a non-linear term does not present any new problem to our understanding of multivariate regression. The statistical significance of both the individual coefficients and the overall model in an OLS regression model incorporating non-linear terms is determined in the same way as for any other OLS regression model. For individual coefficients, we use the *t*-test and for the overall model, we use the *F*-test.

Summary

How does one know whether to include a non-linear term in a regression model? In light of the many different non-linear relationships that are possible – we could transform any number of our independent variables in an OLS regression model – how do we settle on an appropriate model? The single best guide for the researcher is prior theory and research. If a theoretical perspective claims a non-linear relationship or prior research has established a non-linear relationship between two variables, then the researcher may want to examine a non-linear relationship in the regression analysis. Without the guidance of theory and prior research, the researcher is better off using an OLS model without any non-linear relationships included. If subsequent analyses, such as a residual analysis (discussed in Chapter 16) indicate a non-linear relationship, then some kind of transformation of an independent variable may be in order.

Interaction Effects

A number of different theories of crime and delinquency make statements about how the effect of one variable will vary by the level of some other variable. A perspective known as general strain theory hypothesizes that the effects of psychological strain (e.g., having one's parents file for divorce) on delinquency will vary by the ability of a youth to adapt to strain.[3] For example, if an individual characteristic, such as self-esteem, helps individuals to adapt to various forms of strain, then the effect of that strain may vary by the level of self-esteem: as the level of self-esteem increases, the effect of strain on the chances of delinquency may become smaller. Alternatively, research on criminal justice decision-making has suggested that the effects of offender characteristics, such as the offender's age, may differentially affect the severity of punishment across different racial or ethnic categories.[4]

Assuming that we have a rationale for including an interactions effect, how do we incorporate it into our regression model? Let us begin with a simple regression model that has two independent variables X_1 and X_2.

$$Y = b_0 + b_1X_1 + b_2X_2 + e$$

[3] R. Agnew, 1992, Foundation for a general strain theory of crime and delinquency, *Criminology*, 30, 47-87.

[4] D. Steffensmeier, J. Kramer, and J. Ulmer, 1995, Age differences in sentencing, *Justice Quarterly*, 12, 583-602.

To add an interaction effect to a regression model, all that we need to do is to compute the product of the two variables: $X_1{*}X_2 = X_3$. We then add this term to the regression equation:

$$Y = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + e$$

Where $X_3 = X_1{*}X_2$

Note that we now have an additional regression coefficient (b_3) in the model that represents the interaction of the variables X_1 and X_2, which we will need to interpret. The interpretation of interaction effects can be quite complicated, with the degree of complexity based on the level of measurement of the two variables.

Interaction of a Dummy Variable and Interval-Level Variable

To illustrate the process of interpreting interaction effects, it is useful to begin with a relatively simple case: the interaction of a dummy variable with a variable measured at the interval level of measurement. In the regression model above, let us assume that X_2 is a dummy variable, where the two categories are coded as either 0 or 1.

It is now possible to work through a series of regression equations, much like we did in the previous chapter in our discussion of dummy variables, by inserting different values for X_2. Note that the key difference is we now have more than one place where we need to insert values for the dummy variable.

If we set the value for $X_2 = 1$, we have the following regression equation:

$$Y = b_0 + b_1X_1 + b_2X_2 + b_3(X_1{*}X_2) = b_0 + b_1X_1 + b_2(1) + b_3\ (X_1{*}1)$$

Which reduces to:

$$Y = b_0 + b_1X_1 + b_2 + b_3X_1$$

By rearranging our terms, we can rewrite the regression equation as:

$$Y = (b_0 + b_2) + b_1X_1 + b_3X_1 = (b_0 + b_2) + (b_1 + b_3)X_1$$

As in the previous chapter, we see that when we focus our attention on the category with the value 1, the model intercept changes by the value of the coefficient for that variable (i.e., b_2). What is different in the above equation is that the effect of variable X_1 is now the sum of two different regression coefficients: the original coefficient for X_1 (i.e., b_1) and the coefficient for the interaction term (i.e., b_3).

How do we now interpret the effect of X_1? After summing the two regression coefficients b_1 and b_3, we would say that for cases that had a value of 1 on X_2 (i.e., the cases were in Group 1), for each one unit increase in X_1, Y is expected to change by b_1+b_3 units.

When we set the value for $X_2 = 0$ (the reference category for our dummy variable – Group 0), we now have the following regression equation:

$$Y = b_0 + b_1X_1 + b_2(0) + b_3\ (X_1{*}0)$$

Which reduces to:

$$Y = b_0 + b_1X_1$$

This indicates that the model intercept (b_0) and coefficient for X_1 (b_1) represent the intercept for the reference category on X_2 and the effect of X_1 for cases in the reference category, respectively.

To make the example more concrete, suppose that after estimating this regression equation, we find the following results:

$b_0 = 2.5$
$b_1 = 3.2$
$b_2 = 1.9$
$b_3 = 1.3$

By inserting the values for the regression coefficients into the regression equation, we have the following:

$$Y = 2.5 + 3.2\ X_1 + 1.9\ X_2 + 1.3\ (X_1{*}X_2)$$

For $X_2 = 1$, we have the following:

$$\begin{aligned} Y &= 2.5 + 3.2\ X_1 + 1.9\ (1) + 1.3\ (X_1{*}1) \\ &= (2.5 + 1.9) + (3.2 + 1.3)\ X_1 \\ &= 4.4 + 4.5\ X_1 \end{aligned}$$

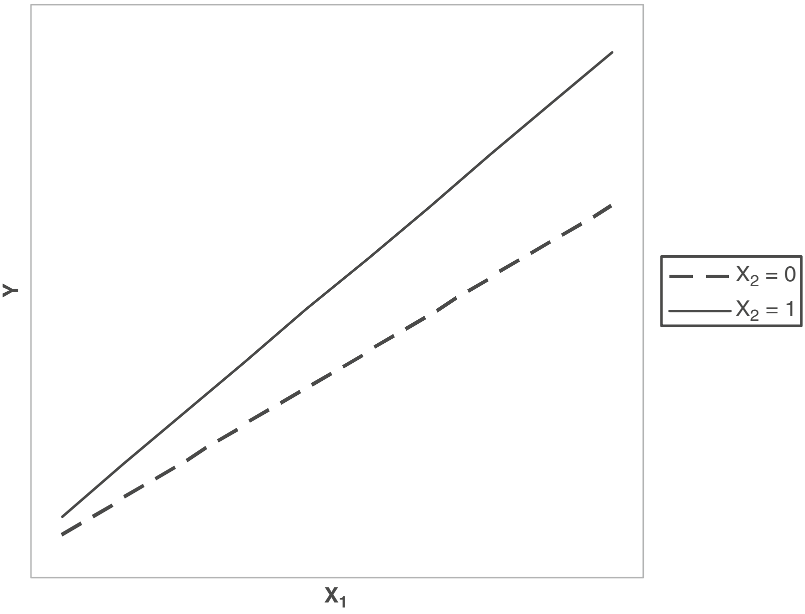
And for $X_2 = 0$, we have:

$$Y = 2.5 + 3.2\ X_1 + 1.9\ (0) + 1.3\ (X_1{*}0) = 2.5 + 3.2\ X_1$$

The interpretation of the effect of X_1 is straightforward, but we need to make sure that we are clear about the group for which we are interpreting the effect of X_1. Thus, for $X_2 = 1$, for each one unit increase in X_1, we expect Y to increase by 4.5 units. When $X_2 = 0$, for each one unit increase in X_1, Y is expected to increase by 3.2 units. Substantively, this type of result would allow a researcher to say that the effect of X_1 varied across the groups measured in X_2. As a visual aid to understanding these results, we have presented the two regression lines in Figure 17.5, where group $X_2 = 0$ is represented by the dashed line and group $X_2 = 1$ is represented by the solid line.

Up to this point, we have assumed that all of the coefficients are positive. Figure 17.6 presents several additional possibilities for various combinations

Figure 17.5 *Regression Lines for the Interaction of X_1 and X_2*

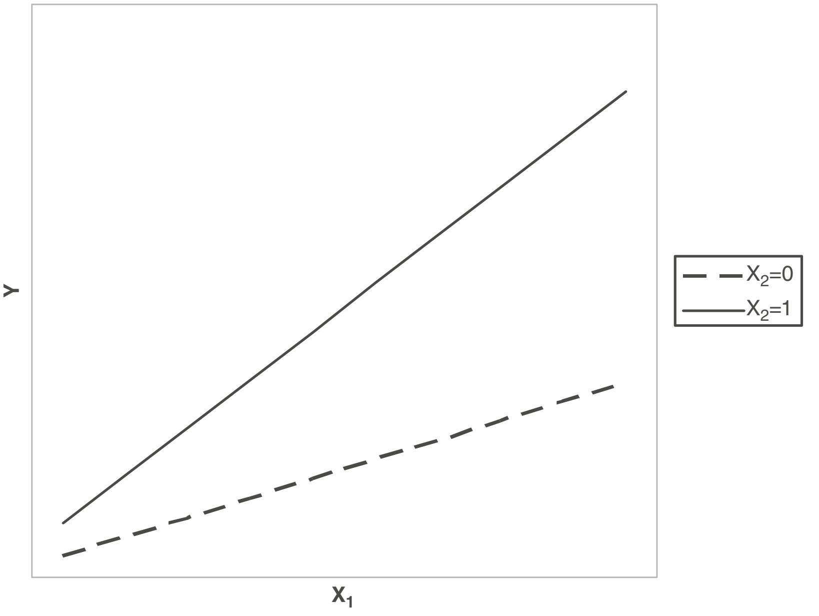


of positive and negative values for b_1 and b_3 (we assumed that b_0 and b_2 were positive in each plot). (Keep in mind that b_1 represents the effect of X_1 for the reference category of the dummy variable and b_3 represents the value of the interaction effect.) Panel (a) is comparable to the preceding example, where b_1 and b_3 are both positive. Panel (b) illustrates a hypothetical example when b_1 is positive and b_3 is negative. Panels (c) and (d) illustrate possible patterns when b_1 is negative and b_3 is positive (panel (c)) or negative (panel (d)). Clearly, there are many other possibilities, but we wanted to provide a few illustrations for different patterns that researchers have had to address in their analyses.

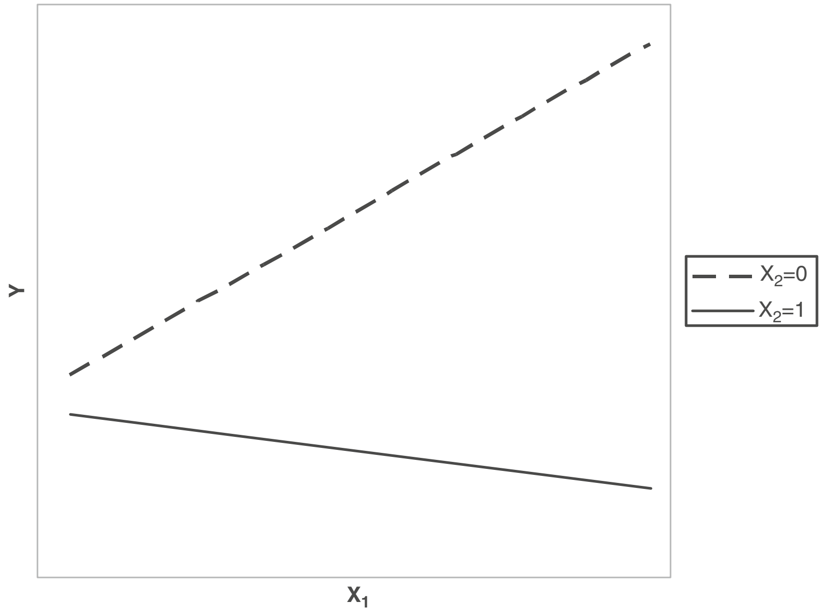
An Example: Race and Punishment Severity

Suppose that we are interested in testing whether the severity of a criminal offense differentially affects the length of time offenders are sentenced to prison by race. Put another way, does the severity of the conviction offense affect the severity of punishment in the same way for offenders of different races? We again use data on the sentences of over 20,000 offenders sentenced to prison in Pennsylvania in 1998 to illustrate the test for an interaction effect between severity of offense and race of offender. To simplify our model here, we measure race as a dummy variable (0 = white, 1 = African American). Offense severity is scored by the Pennsylvania Sentencing Commission and has values ranging from 1 to 14

Figure 17.6 *Hypothetical Interaction Effects for Different Combinations of Positive and Negative Interaction Effects*

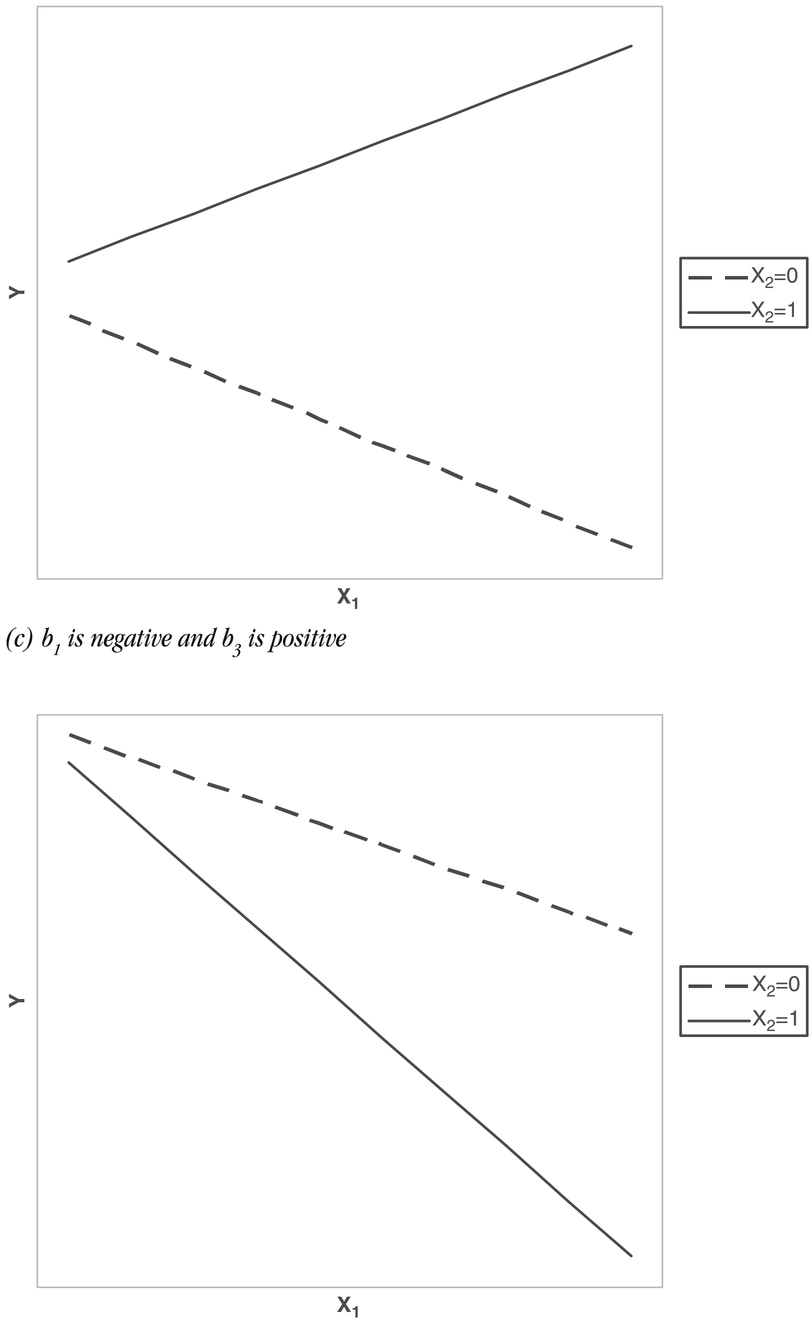


(a) b_1 and b_3 are both positive



(b) b_1 is positive and b_3 is negative

Figure 17.6 *Continued*



(c) b_1 is negative and b_3 is positive

(d) b_1 and b_3 are both negative

and sentence length is measured in months sentenced to prison. The regression model we set out to test can be written as:

$$Y = b_0 + b_1 X_1 + b_2 X_2 + b_3 X_1{*}X_2$$

where Y represents length of sentence (in months)
X_1 represents offense severity
X_2 represents race

When we estimate this regression model, we produce the following set of results:

$$Y = -8.14 + 4.12 X_1 - 5.41 X_2 + 1.31 X_1{*}X_2$$

Using the same approach as above, we begin by focusing on African Americans (X_2 = 1):

$$\begin{aligned} Y &= -8.14 + 4.12 X_1 - 5.41 (1) + 1.31 (X_1{*}1) \\ &= (-8.14 - 5.41) + (4.12 + 1.31) X_1 \\ &= -13.55 + 5.43 X_1 \end{aligned}$$

For whites, the equation is:

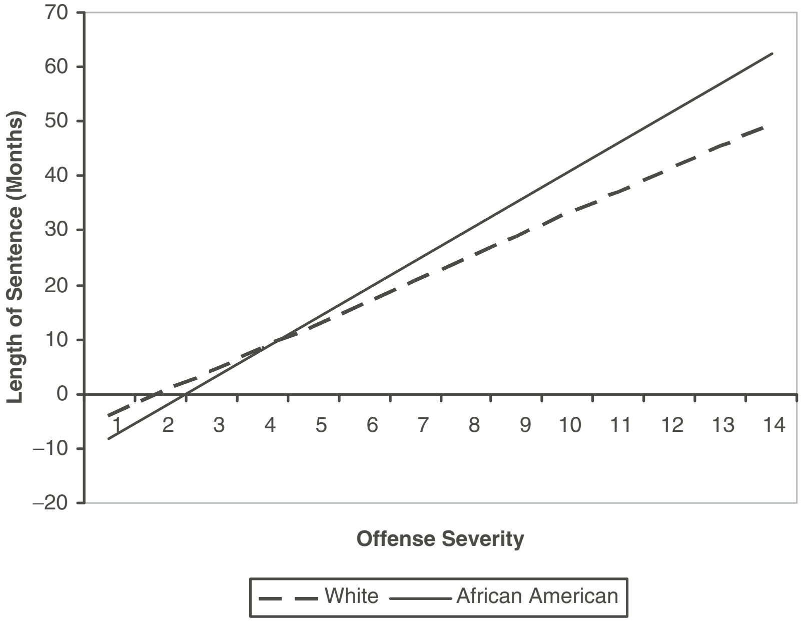
$$\begin{aligned} Y &= -8.14 + 4.12 X_1 - 5.41 (0) + 1.31 (X_1{*}0) \\ &= -8.14 + 4.12 X_1 \end{aligned}$$

Substantively, we can now directly interpret the effect of offense severity for white and African American offenders separately. Among white offenders, each one unit increase in offense severity, is expected to increase sentence length by 4.12 months, while for African American offenders, each one unit increase in offense severity is expected to increase sentence length by 5.43 months. More succinctly, these results suggest that the *effect* of offense severity on punishment severity is greater for African American offenders than for white offenders. These results are presented graphically in Figure 17.7. The dashed line reflects the slope for offense severity on sentence length for white offenders, while the solid line reflects the effect for African American offenders. As one moves further out the x-axis to greater values for offense severity, there is an increasingly greater effect for African American offenders compared to white offenders.

Interaction Effects between Two Interval-level Variables

Up to this point, our attention has been focused on interaction effects involving one variable measured at the interval level of measurement and one dummy variable measured at the nominal level of measurement. The inclusion of an interaction effect between two interval-level variables in a regression model is done in exactly the same way – we compute a product of the two variables and add the product to the regression equation. The interpretation of the interaction effect is much more complex,

Figure 17.7 *Regression Lines for the Effect of Offense Severity on Sentence Length by Race of Offender*



however, since we are no longer able to simplify the regression equation to represent the effect of one variable for two different groups.

In some cases we may not be concerned with a specific interpretation of the interaction term. For example, we may want to simply identify whether the interaction between two measures is statistically significant. A statistically significant interaction term between two measures would suggest that their effect cannot be measured only by the additive effects of each measure in the model, but rather there is an additional effect that is measured by the interaction term.

It may help to conceptualize this issue if we turn to a substantive example. Many sociologists have suggested that extra-legal variables such as income and social status impact upon sentencing outcomes.[5] In a simple additive model each of these factors would have some defined independent effect on the severity of a sentence measured in months of imprisonment. This model is illustrated in equation form below.

$$Y = b_0 + b_1 X_1 + b_2 X_2$$

where Y represents length of sentence (in months)
X_1 represents income
X_2 represents social status

[5] Donald J Black, The Behavior of Law (New York; Academic Press, 1976).

But what if the researcher believed that the effect of income and social status was not simply additive but also multiplicative, meaning that there was an added effect that was due to the interaction between the two. This theory is represented in the equation below:

$$Y = b_0 + b_1 X_1 + b_2 X_2 + b_3 X_1 * X_2$$

where Y represents length of sentence (in months)
X_1 represents income
X_2 represents social status

In this case the researcher is hypothesizing that there is not only the independent effect of income and of social class, but that there is an additional interaction effect that is measured by multiplying income by social class. What if this effect is statistically significant? What interpretation can the researcher draw? To illustrate this we take a hypothetical example of regression results as reported below:

$$Y = 7.2 - 2.4 X_1 - 1.6 X_2 - 1.1 X_1 * X_2$$

The additional interaction term in this case suggests that there is an additional benefit beyond that of the additive independent effects of income and social status that must be taken into account. In a very simple interpretation, we can say that not only does a high income high status individual receive a benefit from their income and status, but when the individual is high on both measures simultaneously they gain an added benefit above and beyond that provided by each characteristic on its own.

While models specified in this way can help us to identify additional impacts that come from the interactions between interval level variables, it is very hard to develop interpretations beyond what we have noted above. But the researcher can adjust such models to develop a more easily interpretable understanding of interaction terms.

Conceptually, when we have an interaction between two interval-level variables, we are testing the idea that the effect of one interval-level variable varies by the level of the second interval-level variable. For example, in the example noted above from general strain theory, the hypothesis is that the effect of strain varies by the level of self-esteem. In practice, the difficulty we often have in the interpretation of interaction effects between two interval-level variables is in choosing values for one variable to represent the effect of the other variable. In effect, we have already done this in our example of an interaction between the dummy variable and the interval-level variable. Recall that when we include an interaction effect between a dummy variable and an interval-level variable, we set the value of the dummy variable to either 0 or 1 and then interpret the effect of the interval-level variable for each group represented in the dummy variable.

In trying to determine how to interpret the interaction between two interval-level variables, we would encourage you to first consider which variable is of key importance for a study. The second variable would then be set at a limited number of values, which allows the researcher to see how the effect of the key variable changes across levels of the second variable. For example, if we again refer to the interaction between strain and self-esteem, the key theoretical variable is strain. Following these guidelines, we would then want to interpret the effect of strain for several designated values of self-esteem. Clearly, we could interpret the interaction effect the other way: the effect of self-esteem for specified levels of strain, but this is not a key piece of the theory.

What values do we use for the second interval-level variable? For any given interval-level variable, there may be hundreds or thousands of realistic possible values that we could use. We think that a useful place to start is to use the mean, one standard deviation above and below the mean, and two standard deviations above and below the mean. This will cover a wide range of possible values of the variable we are using and should be ample for understanding how our key variable changes across values of the second variable. In other cases, where there may be meaningful values on the second independent variable that have more intuitive meaning to the reader, these values should be used. For example, if we were to fix years of education, we might use 8, 12, and 16 to reflect the completion of junior high school, high school, and undergraduate collegiate education, respectively.

For example, suppose that we have estimated a regression model with two interval-level variables X_1 and X_2 and the interaction of X_1 and X_2:

$$Y = 2.3 + 1.7\ X_1 + 2.0\ X_2 + 0.5\ (X_1 * X_2)$$

For the purpose of this example, we will consider X_1 the key variable. We find the mean and standard deviation of X_2 to be 3.2 and 1.2, respectively.

The values that are one or two standard deviations above and below the mean of X_2 are:

Two standard deviations above: $3.2 + 2 * 1.2 = 3.2 + 2.4 = 5.6$
One standard deviation above: $3.2 + 1.2 = 4.4$
One standard deviation below: $3.2 - 1.2 = 2.0$
Two standard deviations below: $3.2 - 2 * 1.2 = 3.2 - 2.4 = 0.8$

We can now input these values for X_2 to determine the effect of X_1 on Y:

Effect of X_1 at the mean of X_2:

$$\begin{aligned} Y &= 2.3 + 1.7\ X_1 + 2.0 * (3.2) + 0.5\ (X_1 * 3.2) \\ &= 2.3 + 1.7\ X_1 + 6.4 + 1.6\ X_1 \\ &= (2.3 + 6.2) + (1.7 + 1.6)\ X_1 \\ &= 8.5 + 3.3\ X_1 \end{aligned}$$

If we wanted to interpret the effect of X_1 directly, then we would state that at the mean for X_2, each one-unit increase in X_1 is expected to increase Y by 3.3 units.

Effect of X_1 at one standard deviation above the mean of X_2:

$$\begin{aligned} Y &= 2.3 + 1.7\ X_1 + 2.0 * (4.4) + 0.5\ (X_1 * 4.4) \\ &= 2.3 + 1.7\ X_1 + 8.8 + 2.2\ X_1 \\ &= (2.3 + 8.8) + (1.7 + 2.2)\ X_1 \\ &= 11.1 + 3.9\ X_1 \end{aligned}$$

If we wanted to interpret the effect of X_1 directly, then we would state that at one standard deviation above the mean for X_2, each one-unit increase in X_1 is expected to increase Y by 3.9 units.

Effect of X_1 at one standard deviation below the mean of X_2:

$$\begin{aligned} Y &= 2.3 + 1.7\ X_1 + 2.0 * (2.0) + 0.5\ (X_1 * 2.0) \\ &= 2.3 + 1.7\ X_1 + 4.0 + 1.0\ X_1 \\ &= (2.3 + 4.0) + (1.7 + 1.0)\ X_1 \\ &= 6.3 + 2.7\ X_1 \end{aligned}$$

If we wanted to interpret the effect of X_1 directly, then we would state that at one standard deviation below the mean for X_2, each one-unit increase in X_1 is expected to increase Y by 2.7 units.

Effect of X_1 at two standard deviations above the mean of X_2:

$$\begin{aligned} Y &= 2.3 + 1.7\ X_1 + 2.0 * (5.6) + 0.5\ (X_1 * 5.6) \\ &= 2.3 + 1.7\ X_1 + 11.2 + 2.8\ X_1 \\ &= (2.3 + 11.2) + (1.7 + 2.8)\ X_1 \\ &= 13.5 + 4.5\ X_1 \end{aligned}$$

If we wanted to interpret the effect of X_1 directly, then we would state that at two standard deviations above the mean for X_2, each one-unit increase in X_1 is expected to increase Y by 4.5 units.

Effect of X_1 at two standard deviations below the mean of X_2:

$$\begin{aligned} Y &= 2.3 + 1.7\ X_1 + 2.0 * (0.8) + 0.5\ (X_1 * 0.8) \\ &= 2.3 + 1.7\ X_1 + 1.6 + 0.4\ X_1 \\ &= (2.3 + 1.6) + (1.7 + 0.4)\ X_1 \\ &= 3.9 + 2.1\ X_1 \end{aligned}$$

If we wanted to interpret the effect of X_1 directly, then we would state that at two standard deviations below the mean for X_2, each one-unit increase in X_1 is expected to increase Y by 2.1 units.

Aside from making direct interpretations of the effect of X_1 at these five values of X_2, what we see is that as the value of X_2 increases, the *effect* of X_1 on Y increases.

An Example: Punishment Severity

We again use the data on the sentencing of offenders in Pennsylvania in 1998 and modify our regression model slightly. We continue to use length of sentence as the dependent variable and severity of the offense as an independent variable. Our second independent variable is a prior criminal history score that is computed by the Pennsylvania Sentencing Commission and can take on values ranging from 0 to 8; larger values for prior criminal history reflect both a greater number of prior offenses as well as more serious prior offenses. For the purposes of this example, we have added an interaction effect between severity of the offense and prior criminal history and are interested in how the effect of offense severity varies across levels of prior criminal history.

After estimating this model, we obtain the following regression equation:

$$Y = -7.83 + 3.58\ X_1 - 1.21\ X_2 + 0.62\ X_1{*}\ X_2$$

Where Y represents length of prison sentence (in months)
X_1 represents offense severity
X_2 represents prior criminal history

Since we are primarily interested in the effect of offense severity across levels of criminal history, we have computed the mean and standard deviation of the criminal history variable to be 1.51 and 1.97, respectively. Following the same procedure as in our hypothetical example, we calculate the effect of offense severity at the mean of prior criminal history.

Effect of offense severity at the mean of prior criminal history:

$$\begin{aligned} Y &= -7.83 + 3.58\ X_1 - 1.21\ (1.51) + 0.62\ (X_1{*}\ 1.51) \\ &= -7.83 + 3.58\ X_1 - 1.83 + 0.94\ X_1 \\ &= (-7.83 - 1.83) + (3.58 + 0.94)\ X_1 \\ &= -9.66 + 4.52\ X_1 \end{aligned}$$

If we wanted to interpret the effect of offense severity directly, then we would state that at the mean for prior criminal history, each one-unit increase in offense severity is expected to increase sentence length by 4.52 months.

Since the standard deviation of prior criminal history is greater than the mean, we cannot sensibly compute values at 1 or two standard deviations below the mean, since the values would be negative, and prior record score is constrained to have a value ranging from 0 to 8. In such a case, we can choose other values for the independent variable that carry substantive meaning. For example, a prior criminal history score of 0 implies little or no prior criminal activity, while a prior criminal history score of 8 implies an extensive history of serious and likely violent offending. By using the minimum and maximum values for prior criminal history, along with the mean, we can gain a good sense of how the effect of offense severity varies by level of prior criminal history.

Effect of offense severity at a prior criminal history score of 0:

$$
\begin{aligned}
Y &= -7.83 + 3.58\ X_1 - 1.21\ (0) + 0.62\ (X_1 * 0) \\
&= -7.83 + 3.58\ X_1
\end{aligned}
$$

A direct interpretation of the effect of offense severity at a prior criminal history score of 0 indicates that each one-unit increase in offense severity is expected to increase sentence length by 3.58 months.

Similarly, if we take the maximum value for prior criminal history score of 8:

$$
\begin{aligned}
Y &= -7.83 + 3.58\ X_1 - 1.21\ (8) + 0.62\ (X_1 * 8) \\
&= -7.83 + 3.58\ X_1 - 9.68 + 4.96\ X_1 \\
&= (-7.83 - 9.68) + (3.58 + 4.96)\ X_1 \\
&= -17.51 + 8.54\ X_1
\end{aligned}
$$

Thus, among the historically most serious offenders, these results suggest that each one-unit increase in offense severity is expected to increase sentence length by 8.54 months.

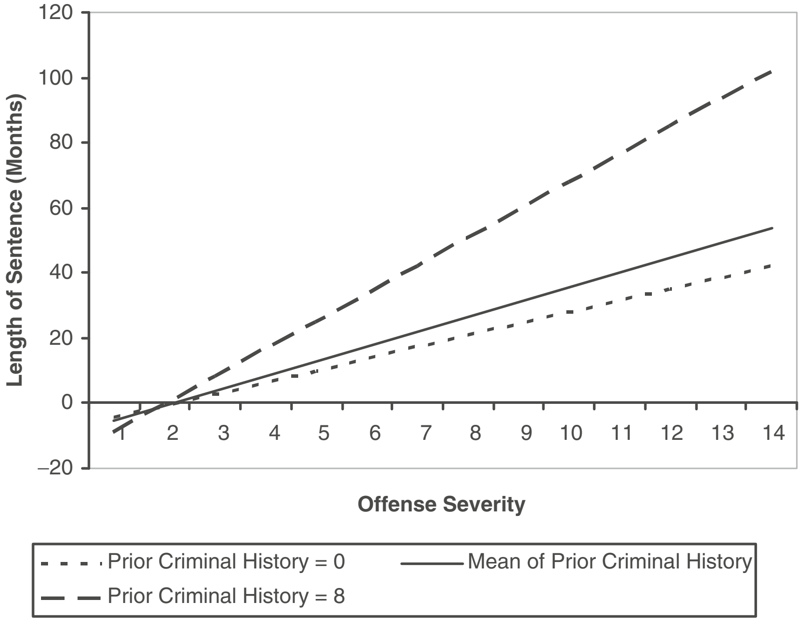
Much like the hypothetical example earlier, we see that as the value for prior criminal history increases, the *effect* of offense severity increases. To appreciate the change in the effect of offense severity, we have plotted these three regression equations in Figure 17.8. We see that at a prior criminal history score of 0, the regression line is positive, but shows modest increases over the range of offense severity. As we move to the mean of prior criminal history and then to the maximum value for prior criminal history, we see the slope of the regression line become steeper, reflecting increasingly greater effects on sentence length for any given increase in prior criminal history.

The Problem of Multicollinearity

The use of interaction terms is very likely to create a problem in regression analyses that can lead to difficulty in estimation of the regression equation. This is because we are likely to be interested not only in the interaction between two variables, but also the simple additive effects of each independently. When we include the original variables as well as the term for their interaction, the three terms are likely to be very highly correlated. Even though multivariate regression was developed in part to take into account the interrelationships among variables that predict *Y*, when independent variables in a regression model are too strongly related to one another, regression estimates become unstable. This problem is called **Multicollinearity.**

Figure 17.8

Regression Lines for the Effect of Offense Severity on Sentence Length by Level of Prior Criminal History Score



In criminal justice, the independent variables examined are generally multicollinear, or correlated with one another. Indeed, this correlation is one reason it is so important to use multivariate techniques in criminal justice research. When variables are intercorrelated, as in the case of our example of years in prison and prior arrests discussed in Chapter 16, it is important to control for the potential confounding influences of one variable on the other. Failure to do so is likely to lead to bias in our estimates of the effects of specific regression coefficients. However, the irony of multicollinearity is that when variables become too correlated, or highly multicollinear, the regression estimates become unreliable. This may happen when models do not include interaction terms, but is a particularly serious concern when interactions between variables are specified in a model.

Multicollinearity can be identified in one of two ways. A common method is to look at the intercorrelations among the independent variables included in your model. Very high correlations between independent variables are likely to lead to multicollinearity problems. What is considered a very high correlation? As with many other definitions in statistics, there is no absolute number at which multicollinearity is considered serious. As a general rule, a correlation between two independent variables of greater than 0.80 should be seen as a warning that serious multicollinearity may be evident in your model.

Multicollinearity between two variables occurs less often than multicollinearity across a series of variables. To diagnose this type of multicollinearity, we use a statistic that is usually defined as **tolerance.** Tolerance measures the extent of the intercorrelations of each independent variable with all other independent variables. It is defined as 1 minus the percent of variance in X explained by the other independent variable examined (Equation 17.1).

$$Tolerance = 1 - R_X^{\,2} \qquad \textbf{Equation 17.1}$$

Calculation of tolerance is generally provided as an option in standard statistical computing packages, but it also can be calculated by taking each independent variable as the dependent variable in a regression that includes all other independent variables. This value is then subtracted from 1. For example, let's say we defined a model for explaining sentence length among offenders in Pennsylvania that included three independent variables:

$$\begin{aligned} Y_{\text{length}} &= b_0 + b_1\ (\text{age}) + b_2\ (\text{offense severity}) \\ &\quad + b_3\ (\text{prior criminal history}) + e \end{aligned}$$

The R_X^2 for age would be estimated by calculating a regression in which age was the dependent variable and offense severity and prior criminal history were the independent variables. You would then take this R^2 and subtract it from 1. Similarly, to get R^2_X for offense severity, you would regress age and prior criminal history on offense severity and then subtract the resulting R^2 from 1. Table 17.1 presents the Tolerance statistics for this regression model.

How do we know if multicollinearity is negatively effecting our model estimates based on the tolerance statistics? A very small tolerance statistic suggests that the model is likely to include a high level of multicollinearity. Again, there is no clear yardstick for defining a level of tolerance that is likely to lead to estimation problems. In general, however, a tolerance level of less than 0.20 should be taken as a warning that serious multicollinearity may exist in your model. We see from the results in Table 17.1 that the smallest tolerance statistic has a value of 0.94, which does not indicate any serious multicollinearity in this regression model.

Beyond these diagnostic procedures for multicollinearity, there are warning signs that can be observed in the regressions that are estimated. Sometimes when multicollinearity is present, the percent of explained

Table 17.1 Tolerance Statistics for Regression of Sentence Length for Pennsylvania Offenders

INDEPENDENT VARIABLE	TOLERANCE
Age	0.940
Offense Severity	0.931
Prior Criminal History	0.975

variance in a model is high, but the regression coefficients overall fail to reach conventional thresholds of statistical significance. Sometimes multicollinearity inflates coefficients to unrealistic sizes or produces coefficients in a direction contrary to conventional wisdom. One problem in diagnosing multicollinearity is that it may have such varied effects in your model that you may have difficulty distinguishing a misleading result that is due to multicollinearity from one that represents a new and interesting finding.

When there are indications of serious multicollinearity, you can take a number of alternative corrective measures. The simplest is to exclude the variable or variables that are contributing most to multicollinearity. In the case of interaction terms this might require that one of the terms included in the interaction be dropped from the model. In the case where a small group of measures are highly collinear, you might choose to exclude the one variable that appears to present the most serious problem (i.e., that has the lowest tolerance value). The drawback of this approach is that the exclusion of such measures is likely to lead to model misspecification and may result in biased estimates of other regression coefficients that remain in the model. This approach makes sense only when other variables that remain in the model measure the same concept or theory. An approach that achieves a similar result, without excluding specific measures, is to create new indices from clusters of variables that are multicollinear. For example, if a group of measures all relating to social status are multicollinear, you may decide to create a new composite measure defined as social status and use it as an independent variable in subsequent regressions.

Chapter Summary

Non-linear relationships refer to the effect of the independent variable on the dependent variable not being a straight-line (linear) relationship. A linear relationship implies that each one unit increase in the independent variable will result in the dependent variable increasing or decreasing by some fixed amount, regardless of the level of the independent variable. A non-linear relationship implies that each one unit increase in the independent variable does not result in the same amount of change in the dependent variable – it may be larger or smaller and will vary by the level of the independent variable. A non-linear relationship can be incorporated into an OLS regression equation by transforming the independent variable.

Interaction effects reflect the varying effect of one independent variable on the dependent variable across the levels or values of a second independent variable. When we have an interaction effect between a dummy variable and an interval-level variable, we can directly interpret the effect of the interval-level variable on the dependent variable for each group measured by the dummy variable. Interpretation of an interaction of two interval level independent variables is much more difficult. One way of simplifying interpretation is to designate values for one variable, such as the mean, one standard deviation above/below the mean, two standard deviations above/below the mean, and so on, as fixed points to compute the effect of the other interval-level variable on the dependent variable.

Multicollinearity occurs when independent variables in a regression model are too strongly related. It leads to unstable results. The problem may be diagnosed by checking the bivariate correlations between the variables and by measuring **tolerance.** Multicollinearity may be dealt with either by excluding specific variables altogether or by merging several similar variables into one composite index.

Key Terms

interaction effect An interaction effect is present when the effect of one independent variable on the dependent variable is conditional on the level of a second independent variable.

multicollinearity Condition in a multivariate regression model in which independent variables examined are very strongly intercorrelated. Multicollinearity leads to unstable regression coefficients.

non-linear relationship Relationship between the dependent and the independent variable that is not captured by a straight line (linear) relationship.

tolerance A measure of the extent of the intercorrelations of each independent variable with all other independent variables. Tolerance may be used to test for multicollinearity in a multivariate regression model.

Symbols and Formulas

R^2_X R^2 obtained when an independent variable is treated as a dependent variable in a test for tolerance

To calculate tolerance:

$$\text{Tolerance} = 1 - R^2_X$$

Exercises

17.1 An analysis of shoplifting frequency among youth and young adults included a quadratic term for age of the individual and produced the following results:

INDEPENDENT VARIABLE	B
Age (in years)	0.35
Age^2 (in years^2)	–0.01

Interpret the effect of age on the frequency of shoplifting.

17.2 An analysis linking level of informal social control to frequency of delinquency produced the following results:

INDEPENDENT VARIABLE	B
Age (in years)	–0.12
Sex (1=Female, 0=Male)	–1.50
Race (1=White, 0=Non-white)	0.27
Informal Social Control (1=Low, 10=High)	–0.83

After plotting the mean level of delinquency by level of informal social control, the researcher observed what appeared to be an inverse relationship (1/X) between delinquency and informal social control. After transforming the measure of informal social control, the researcher estimated a new regression and produced the following results:

INDEPENDENT VARIABLE	B
Age (in years)	–0.11
Sex (1=Female, 0=Male)	–1.61
Race (1=White, 0=Non-white)	0.32
Inverse of Informal Social Control (1=Low, 10=High)	2.45

a. Interpret the effect of the inverse of informal social control.

b. Sketch the relationship between delinquency and informal social control using the coefficient for the inverse of informal social control.

17.3 A researcher wanted to test the hypothesis that adolescent females were more affected by parental supervision than adolescent males. In a regression analysis incorporating an interaction effect between sex and supervision, the researcher produced the following set of results:

INDEPENDENT VARIABLE	B
Sex (1=Female, 0=Male)	–2.7
Supervision (1=Low, 10=High)	–1.3
Sex * Supervision	–0.5

Interpret the effect of supervision for adolescent females and males.

17.4 A study of attitudes about punishment used a scale of punitiveness ranging in value from 1 (Low) to 100 (High). The researcher was particularly interested in whether there was an interaction effect between age and political conservatism. A regression analysis produced the following results:

INDEPENDENT VARIABLE	B	MEAN
Age (years)	1.67	44.95
Political Conservatism (1=Low, 10=High)	0.92	6.5
Age * Political Conservatism	0.56	

a. What is the effect of political conservatism at the mean age of the sample? Interpret this effect.

b. What is the effect of age at the mean level of political conservatism for the sample? Interpret this effect.

c. What is the effect of political conservatism at each of the following ages?

- 20
- 30
- 50
- 60

Describe how the effect of political conservatism changes as age increases.

d. What is the effect of age at each of the following values of political conservatism?

- 0
- 2
- 5
- 8
- 10

Describe how the effect of age changes as the level of political conservatism increases.

17.5 A study of violence in prison cell blocks was concerned about the amount of space available to each inmate and the proportion of inmates identified as gang members who had been identified as gang members. The researcher tested the hypothesis of an interaction effect between space available and the proportion of inmates identified as gang members. A regression analysis produced the following results:

INDEPENDENT VARIABLE	B	MEAN
Space available (square feet per inmate)	–0.25	10.0
Proportion gang members	0.77	0.77
Space available * Proportion gang members	–0.05	

a. What is the effect of space available at the mean proportion of gang members for the sample of cell blocks? Interpret this effect.

b. What is the effect of proportion of gang members at the mean level of space available for the sample of cell blocks? Interpret this effect.

c. What is the effect of space available at each of the following proportions of gang membership?

- 0.2
- 0.4
- 0.6
- 0.8

Describe how the effect of space available changes as proportion of gang membership increases.

d. What is the effect of proportion of gang membership at each of the following values of space available?

– 3
– 6
– 12
– 15

Describe how the effect of proportion of gang membership changes as the level of space available increases.

17.6 Rachel collects police data on a series of burglaries and wishes to determine the factors that influence the amount of property stolen in each case. She creates a multivariate regression model and runs a test of tolerance for each of the independent variables. Her results are as follows, where Y = Amount of property stolen ($):

INDEPENDENT VARIABLE	SCALE	TOLERANCE
X_1: Time of robbery (AM or PM)	Nominal	0.98
X_2: Accessibility of property	Ordinal	0.94
X_3: Number of rooms in house	Interval	0.12
X_4: Size of house	Interval	0.12
X_5: Joint income of family	Interval	0.46

Would you advise Rachel to make any changes to her model? Explain your answer.

17.7 A researcher examining neighborhood crime rates computes a regression model using the following variables:

Y = crime rate (per 100,000)

X_1 = percent living in poverty

X_2 = percent unemployed

X_3 = median income

X_4 = percent of homes being rented

The researcher finds the *F*-statistic for the overall model to be statistically significant (with $\alpha = 0.05$), but the results for each variable are as follows:

INDEPENDENT VARIABLE	B	SIG.	TOLERANCE
X_1: Percent living in poverty	52.13	0.17	0.15
X_2: Percent unemployed	39.95	0.23	0.07
X_3: Median income	22.64	0.12	0.19
X_4: Percent of homes being rented	27.89	0.33	0.05

a. Explain why the researcher found a statistically significant regression model, but no significant regression coefficients

b. What would you recommend the researcher do in this case?

Computer Exercises

In Chapter 16, we illustrated the use of the regression command in SPSS and Stata to estimate multivariate regression models. The analyses described in this chapter—nonlinear terms, interaction effects, and a test for multicollinearity—are accomplished with the same regression command in each program. The following exercises should help to illustrate how to perform these analyses, as will the sample syntax files for SPSS (Chapter_17.sps) and Stata (Chapter_17.do).

SPSS

Computing Nonlinear and Interaction Terms

To include a nonlinear or an interaction term in a multivariate regression model, it is necessary to first compute the nonlinear or the interaction term. This computation is done with the COMPUTE command that we briefly discussed in the Computer Exercises in Chapter 4. The general format for the COMPUTE command is

```
COMPUTE new_var_name = calculation
```

The calculation can be a function of one or more variables, which we illustrate below.

Nonlinear Terms

Suppose we wanted to compute a squared term for a variable AGE. We might name this new variable AGE_SQ. The COMPUTE command would look like

```
COMPUTE AGE_SQ = AGE**2.

EXECUTE.
```

In SPSS, the double asterisk (**) indicates that we want to take a variable to a power. In this example, we want to square AGE, so we add the value 2. If, for some reason, we had wanted to cube AGE, then we would have typed AGE**3. Also, recall that the addition of the EXECUTE command forces SPSS to perform this calculation immediately.

An alternative that accomplishes the same thing is

```
COMPUTE AGE_SQ = AGE * AGE.

EXECUTE.
```

where we simply multiply the variable AGE by itself.

In either case, once the COMPUTE command has been executed, the new variable will appear in the data file.

Interaction Terms

The computation of an interaction term is as direct as the equations given in this chapter. We again use the COMPUTE command, but our calculation involves multiplying the two variables of interest. We have found that it is often helpful to make the name of the new variable representing an interaction term a combination of fragments from both of the original variables being used in the calculation.

For example, suppose that we want to create an interaction term for two variables that we have named EDUCATION and INCOME. We might call the interaction variable EDUC_INC:

```
COMPUTE EDUC_INC = EDUCATION * INCOME.

EXECUTE.
```

Estimating the Regression Model

After computing the nonlinear or the interaction term, we then simply treat the created variable as an additional variable added to our multivariate regression model—identical to how we presented these terms in the discussion in this chapter. For situations where we are using nonlinear terms, we may need to drop the original variable. Prior research and theory indicating that a nonlinear term was appropriate will often be the best guide on what the regression equation should look like. In the case of an interaction term, keep in mind that we must include both of the original variables and the interaction term; otherwise, it will be nearly impossible to interpret the coefficients that we do obtain from a regression analysis.

Collinearity Diagnostics

SPSS's regression command will produce a wide assortment of collinearity statistics, including the tolerance statistic discussed above. To obtain the collinearity diagnostics, we include the /STATISTICS option in our REGRESSION command:

```
REGRESSION
  /STATISTICS COEFF R ANOVA COLLIN TOL
  /DEPENDENT dep_var_name
  /METHOD = ENTER list_of_indep_vars.
```

where TOL requests the tolerance statistics for each independent variable and COLLIN requests all other collinearity measures. The tolerance statistics are presented in the coefficients table, where you will also find the regression coefficients. Recall from the discussion in the chapter that a tolerance of less than about 0.20 is indicative of collinearity problems in a regression model.

Stata

Computing Nonlinear and Interaction Terms

To include a nonlinear or an interaction term in a multivariate regression model, it is necessary to first compute the nonlinear or the interaction term. This computation is done with the **gen** command (short for **generate**) that we briefly discussed in the Computer Exercises in Chapter 4. The general format for the **gen** command is

```
gen new_var_name = calculation
```

The calculation can be a function of one or more variables, which we illustrate below. (Note that the structure of the discussion is nearly identical to that in the SPSS section, with slight changes for the Stata syntax.)

Nonlinear Terms

Suppose we wanted to compute a squared term for a variable AGE. We might name this new variable AGE_SQ. The **gen** command would look like

```
gen AGE_SQ = AGE^2
```

In Stata, the upward pointing arrow (^) indicates that we want to take a variable to a power. In this example, we want to square AGE, so we add the value 2. If, for some reason, we had wanted to cube AGE, then we would have typed AGE^3.

An alternative that accomplishes the same thing is

```
gen AGE_SQ = AGE * AGE
```

where we simply multiply the variable AGE by itself.

In either case, once the **gen** command has been executed, the new variable will appear in the data file.

Interaction Terms

The computation of an interaction term is as direct as the equations given in this chapter. We again use the **gen** command, but our calculation involves multiplying the two variables of interest. We have found that it is often helpful to make the name of the new variable representing an interaction term a combination of fragments from both of the original variables being used in the calculation.

For example, suppose that we want to create an interaction term for two variables that we have named EDUCATION and INCOME. We might call the interaction variable EDUC_INC:

```
gen EDUC_INC = EDUCATION * INCOME
```

Estimating the Regression Model

After computing the nonlinear or the interaction term, we then simply treat the created variable as an additional variable added to our multivariate regression model—identical to how we presented these terms in the discussion in this chapter. For situations where we are using nonlinear terms, we may need to drop the original variable. Prior research and theory indicating that a nonlinear term was appropriate will often be the best guide on what the regression equation should look like. In the case of an interaction term, keep in mind that we must include both of the original variables and the interaction term; otherwise, it will be nearly impossible to interpret the coefficients that we do obtain from a regression analysis.

Collinearity Diagnostics

Within Stata, only the VIF is directly available, but there is a user-written package—**collin**—that will compute the full set of collinearity statistics, including the tolerance. To install collin on your system, run the following command (one time only—see Chapter_17.do):

```
net install collin, from(http://www.ats.ucla.edu/stat/stata/
ado/analysis)
```

This will make the **collin** command available for use.

In order to use the **collin** command, you must first run the **regress** command. Immediately following the regress command, run the **collin** using the same set of independent variables:

```
regress dep_var list_of_independent_variables

collin list_of_independent_variables
```

The output will be similar to that in SPSS and will produce a wide assortment of collinearity statistics, including the tolerance statistic discussed above. Each measure of collinearity is presented in a table that lists each independent variable (by row) and collinearity statistic across columns.

Problems

Open the NYS data file (nys_1.sav, nys_1_student.sav, or nys_1.dta) to complete Exercises 1 through 4.

1. Using one of the line graph commands (described in Chapter 3), generate a plot for the mean of a delinquency measure by age.

   In SPSS, this should look something like

   ```
   GRAPH /LINE(SIMPLE) = MEAN(delinquency_variable)
     BY age .
   ```

   In Stata, the corresponding syntax is a bit more complicated, but use the following lines to accomplish a similar graph as in SPSS:

   ```
   egen mean_delinq1 = mean(delinquency_variable), by(age)

   sort age

   twoway (line mn_delinq1 age, connect(ascending))
   ```

   The resulting line graphs from the syntax given above will plot the mean level of delinquency—for the variable you picked—for each age recorded in the NYS sample. Try this command with other measures of delinquency and see if you notice any variations in the shapes of the lines. (NOTE: If you are using Stata, you will need to change the output variable name in the egen command line, perhaps most simply by changing the number at the end.)

   Do they imply a linear relationship between age and delinquency? A nonlinear relationship? If nonlinear, what kind of nonlinear relationship? (You may want to refer back to Figure 17.2 for some approximations of different curves.)

2. Compute a squared term for the age variable as described above. Estimate a multivariate regression model using one of the measures of delinquency as the dependent variable. Use age and age squared as the independent variables. As noted earlier in the chapter, this will provide a quadratic equation that will test for a nonlinear relationship between age and the measure of delinquency that you have selected.

   Report the values of the regression coefficients for age and age squared and whether or not the coefficients are statistically significant (with $\alpha = 0.05$). Interpret the effect of age on this measure of delinquency.

3. Compute a multivariate regression model using number of times drunk as the dependent variable. Include the measures for age, sex, race (recoded as a dummy variable), and at least two other variables that you think are related to the number of times drunk. This will be the "baseline model" in the following questions.

   a. Compute the tolerance statistic for each of the independent variables in the baseline model. Does it appear that there is a problem with collinearity in the regression model? Explain why.

   b. Compute an interaction term for sex with age. Add this term to the baseline model, and rerun the regression command. Is the effect of age on number of times drunk significantly different for males and females (i.e., is the interaction effect statistically significant)? If so, interpret the effect of age on the number of times drunk for males and females.

      i. Compute the tolerance statistic for each of the independent variables in this model. Does it appear that there is a problem with collinearity in this model? Explain why.

   c. Compute an interaction term for race (which should be coded as a dummy variable) and age. Add this term to the baseline model, and rerun the regression command. (The interaction effect from part (a) should no longer be included in the analysis.) Is the effect of age on number of times drunk significantly different for these two race groups? If so, interpret the effect of age on the number of times drunk for each race group.

      i. Compute the tolerance statistic for each of the independent variables in this model. Does it appear that there is a problem with collinearity in this model? Explain why.

   d. If one of the additional variables that you have added to your regression model is measured at the interval level of measurement, compute an interaction term between this variable and either the sex or the race variable. Add this term to the baseline model (there should be no other interaction terms included in this analysis), and rerun the regression command. Is the effect of this variable on number of times drunk significantly different for the two groups? If so, interpret the effect of this variable for each group.

      i. Compute the tolerance statistic for each of the independent variables in this model. Does it appear that there is a problem with collinearity in this model? Explain why.

4. Compute a multivariate regression model using number of times cheated at school as the dependent variable. Include the measures for age, sex, race (recoded as a dummy variable), grade point average, and amount of time spent studying as the independent variables. This will be the baseline model in the following questions.

a. Compute an interaction term for grade point average and time spent studying and add this term to the regression model. Prior to rerunning the regression command, check the item for descriptive statistics available through the regression command window. Report the coefficient values and whether or not the coefficients are statistically significant (with $\alpha = 0.05$).

b. Compute the tolerance statistic for each of the independent variables in this model. Does it appear that there is a problem with collinearity in this model? Explain why.

c. What is the effect of grade point average on number of times cheated at the mean level of time spent studying? (You will need to use the mean reported in the results for part (a).) Interpret this effect.

   i. How does the effect of grade point average change as the value for time spent studying increases or decreases?

d. What is the effect of time spent studying on number of times cheated at the mean grade point average? Interpret this effect.

   i. How does the effect of time spent studying change as the value for grade point average increases or decreases?