10
Multiple Regression
Summary, Assumptions, Diagnostics, Power, and Problems

You should now have a reasonably complete, conceptual understanding of the basics of multiple regression analysis. This chapter will begin by summarizing the topics covered in Part 1. I will touch on some issues that you should investigate and understand more completely to become a sophisticated user of MR and will close the chapter with some nagging problems and inconsistencies that we have discussed off and on throughout Part 1 (and will try to resolve in Part 2).

SUMMARY

"Standard" Multiple Regression

For social scientists raised on statistical analyses appropriate for the analysis of experiments (ANOVA and its variations), multiple regression often seems like a different animal altogether. It is not. MR provides a close implementation of the general linear model, of which ANOVA is a part. In fact, MR subsumes ANOVA, and as shown in several places in this portion of the book, we can easily analyze experiments (ANOVA-type problems) using MR. The reverse is not the case, however, because MR can handle both categorical and continuous independent variables, whereas ANOVA requires categorical independent variables.

Those with such an experimental background may need to change their thinking about the nature of their analyses, but the underlying statistics are not fundamentally different. In my experience, this transition to MR tends to be more difficult for those with a background in psychology or education; in other social sciences, such as sociology and political science, experimentation (i.e., random assignment to treatment groups) is less common. Even in psychology and education the trend increasingly appears to be to focus on the general linear model, and multiple regression, early in students' research training, so the sometimes-difficult transition I mention here may not apply to you.

In early chapters we covered how to calculate the fundamental statistics associated with multiple regression. More practically, we discussed how to conduct, understand, and interpret MR using statistical analysis programs. R is the multiple correlation coefficient, and R^2 the squared multiple correlation. R^2 is an estimate of the variance explained in the dependent variable by all the multiple independent variables in combination; an R^2 of .2 means that the independent variables jointly explain 20% of the variance in the dependent variable. In applied social science research, R^2'*s* are often less than .5 (50% of the variance explained), unless some sort of pretest is included as a predictor of some posttest outcome, and R^2'*s* of .10 are not uncommon. A high R^2 does not necessarily mean a good model; it depends on the dependent variable to be explained. R^2 may be tested for statistical significance by comparing the variance explained (regression) to the variance unexplained (residual) using an F table, with degrees of freedom equal to the number of independent variables (k) and the sample size minus this number, minus 1 (Nc–k–1).

R^2 provides information about the regression as a whole. The MR also produces information about each independent variable alone, controlling for the other variables in the model. The unstandardized regression coefficients, generally symbolized as b (or sometimes as B), are in the original metric of the variables used, and the b can provide an estimate of the likely change in the dependent variable for each 1-unit change in the independent variable (controlling for the other variables in the regression). For example, Salary, in thousands of dollars a year, may be regressed on Educational Attainment, in years, along with several other variables. If the b associated with Educational Attainment is 3.5, this means that for each additional year of schooling salary would increase, on average, by 3.5 thousand dollars per year. The b is equal to the slope of the regression line. The b's may also be tested for statistical significance using a simple t test ($t = \frac{b}{SE_b}$), with the *df* equal to the *df* residual for the overall F test. This t simply tests whether the regression coefficient is statistically significantly different from zero. More interestingly, it is also possible to determine whether the b differs from values other than zero, either using a modification of the t test or by calculating the 95% (or 90%, or some other level) confidence interval around the b's. Suppose, for example, that previous research suggests that the effect of Educational Attainment on Salary is 5.8. If the 95% CI around our present estimate is 2.6 to 4.4, this means that our present estimate is statistically significantly lower than are estimates from previous research. The use of confidence intervals is increasingly required by journals (see, for example, American Psychological Association, 2010).

We can also examine the standardized regression coefficients associated with each independent variable, generally symbolized as β. β's are in standard deviation units, thus allowing the comparison of coefficients that have different scales. A β of .30 for the effect of Educational Attainment on Salary would be interpreted as meaning that each standard deviation increase in Educational Attainment should result in a .30 *SD* average increase in Salary.

The standardized and unstandardized regression coefficients serve different purposes and have different advantages. Unstandardized coefficients are useful when the scales of the independent and dependent variables are meaningful, when comparing results across samples and studies, when we wish to develop policy implications or interventions from our research, and when interpreting the results of interaction (moderation) analyses. Unstandardized coefficients are also the coefficients that are tested for statistical significance. Standardized coefficients are useful when the scales of the variables used in the regression are not meaningful or when we wish to compare the relative importance of variables in the same regression equation.

The regression analysis also produces an intercept or constant. The intercept represents the predicted score on the dependent variable when all the independent variables have a value of zero. The regression coefficients and the intercept can be combined into a regression equation (e.g., $Y_{\text{predicted}} = \text{intercept} + b_1X_1 + b_2X_2 + b_3X_3$), which can be used to predict someone's score on the outcome from the independent variables.

The regression equation, in essence, creates an optimally weighted composite of the independent variables to predict the outcome variable. This composite is weighted so as to maximize the prediction and minimize the errors of prediction. We can graph this prediction by plotting the outcome (Y-axis) against the predicted outcome (X-axis). The spread of data points around the regression line illustrates the accuracy of prediction and the errors of prediction. Errors of prediction are also known as residuals and may be calculated as outcome scores minus predicted outcome scores. The residuals may also be considered as the outcome variable with the effects of the independent variables statistically removed.

Explanation and Prediction

MR may serve a variety of purposes, but these generally fall under one of two broad categories: prediction or explanation. If our primary interest is in explanation, then we are interested in using MR to estimate the effects or influences of the independent variables on the dependent variable. Underlying this purpose, whether we admit it or not, is an interest in cause and effect. To estimate such effects validly, we need to choose carefully the variables included in the regression equation; it is particularly important that we include any common causes of our presumed cause and presumed effect. An understanding of relevant theory and previous research can help one choose variables wisely. Throughout this text, I have assumed that in most instances we are interested in using MR in the service of explanation, and most of the examples have had an explanatory focus.

In contrast, MR may also be used for the general purpose of prediction. If prediction is our goal, we are not necessarily interested in making statements about the effect of one variable on another; rather, we only want to be as accurate as possible in predicting some outcome. A predictive purpose is often related to selection; a college may be interested in predicting students' first-year GPAs as an aid in determining which students should be admitted. If prediction is the goal, the larger the R^2 the better. One does not need to worry about common causes, or even cause and effect, if one's interest is in prediction, and thus variable selection for prediction is less critical. It may even be perfectly acceptable to have an "effect" predicting a "cause" if prediction is the goal. Theory and previous research can certainly help you choose the variables that will predict your outcome successfully, but they are not critical to the interpretation of your findings as they are when MR is used for explanation. If your interest is in prediction, however, you must refrain from making statements or coming to conclusions about the effects of one variable on another (an explanatory purpose). It is unfortunately common to see research in which the purpose is supposedly prediction, but then when you read the discussion you find explanatory (causal) conclusions are being made. Any time you wish to use MR to make recommendations for intervention or change (if we increase X, Y will increase), your primary interest is in explanation, not prediction. Explanation subsumes prediction. If you can explain a phenomenon well, then you can generally predict it well. The reverse does not hold, however; being able to predict something does not mean you can explain it.

Three Types of Multiple Regression

There are several types, or varieties, of multiple regression. The type of MR used in the earlier chapters of this book is generally referred to as simultaneous, or forced entry, or standard multiple regression. In *simultaneous regression,* all independent variables are entered into the regression equation at the same time. The regression coefficients and their statistical significance are used to make inferences about the importance and relative importance of each variable. Simultaneous regression is useful for explanation or prediction. When used in an explanatory context, the regression coefficients from simultaneous regression provide estimates of the direct effects of each independent variable on the outcome (taking the other independent variables into account); this is one of this method's major advantages. Its chief disadvantage is that the regression coefficients may change depending on which variables are included in the regression equation; this disadvantage is related to the exclusion of relevant common causes or the presence of intervening or mediating variables.

In sequential, or hierarchical, regression, each variable [or group or block of variables] is entered separately into the regression equation, sequentially, in an order determined by the researcher. With *sequential regression,* we generally focus on ΔR^2 from each step to judge the statistical significance of each independent variable. ΔR^2 is a stingy and misleading estimate of the *importance* of variables, however; the square root of ΔR^2 provides a better estimate of the importance of each variable (*given* the order of entry). Order of entry is critical with sequential regression because variables entered early in the sequential regression will appear, other things being equal, more important than variables entered later. Time precedence and presumed causal ordering are common methods for deciding the order of entry. The regression coefficients for each variable from the block in which it enters a sequential regression may be interpreted as the *total* effect of the variable on the outcome, including any indirect or mediating effects through variables entered later in the regression. To interpret sequential regression results in this fashion, variables must be entered in their correct causal order. Causal, or path, models are useful for both sequential and simultaneous regression and have been used to illustrate regression models and results throughout Part 1 of this text; they will be explored in more depth in Part 2. Sequential regression may be used for explanation or prediction. An advantage is that it can provide estimates of the total effects of one variable on another, given the correct order of entry. A chief disadvantage is that the apparent importance of variables changes depending on the order in which they are entered in the sequential regression equation.

Simultaneous and sequential regression may be combined in various ways. One combination is a method sometimes referred to as *sequential unique regression.* It is commonly used to determine the "unique" variance accounted for by a variable or a group of variables, after other relevant variables are accounted for. In this method, the other variables are entered in a simultaneous block, and a variable or variables of interest are entered sequentially in a second block. If a single variable is of interest, simultaneous regression may be used for the same purpose; if the interest is in the variance accounted for by a block of variables, this combination of simultaneous and sequential regression should be used. We made extensive use of this sort of combination of methods when we tested for interactions and curves in the regression line.

A final general method of multiple regression is stepwise regression and its variations. *Stepwise regression* operates in a similar fashion to sequential regression, except that the computer program, rather than the researcher, chooses the order of entry of the variables; it does so based on which variable will lead to the greatest single increment in ΔR^2 at each step. Although this solution seems a blessing—it avoids lots of hard thinking and potentially embarrassing statements about causal ordering—it is not. Using ΔR^2 or $\sqrt{\Delta R^2}$ as a measure of the importance of variables is predicated on the assumption that the variables have been entered in the regression equation in the proper order. To also use ΔR^2 to determine the order of entry thus requires circular reasoning. For this reason, stepwise methods should be used only for prediction, not explanation. In the words of my friend Lee Wolfle, stepwise regression is "theoretical garbage" (1980, p. 206), meaning that its results will mislead rather than inform if you try to use it in explanatory research. And, in fact, stepwise regression may not be a particularly good choice even for prediction. If your interest is simply selecting a subset of variables for efficient prediction, stepwise regression may work (although I still wouldn't recommend it); large samples and cross-validation are recommended. Whatever method of MR you use, be sure you are clear on the primary purpose of your research and choose your regression method to fulfill that purpose.

Categorical Variables in MR

It is relatively easy to analyze categorical, or nominal, variables in multiple regression. One of the easiest ways is to convert the categorical variable into one or more *dummy variables.* With dummy variables, a person is assigned a score of 1 or 0, depending on whether the person is a member of a group or not a member. For example, the categorical variable sex can be coded so that males are scored 0 and females 1, essentially turning it into a "female" variable on which those who are members of the group (females) receive a score of 1 and those who are not members (males) receive a score of 0. For more complex categorical variables, multiple dummy codes are required. We need to create as many dummy variables as there are categories, minus 1 ($g - 1$). When a categorical variable has more than two categories, thus requiring more than one dummy variable, one group will be scored 0 on all the dummy variables; this is essentially the reference group, or often the control group. When dummy variables are analyzed in MR, the intercept is equal to the mean score on the dependent variable for the reference group, and the b's are equal to the mean deviations from that group for each of the other groups.

We demonstrated that MR results match those of ANOVA when the independent variables are all categorical: the F from the two procedures is the same, and the effect size η^2 from ANOVA is equal to the R^2 from MR. The coefficients from MR may be used to perform various post hoc procedures. There are other methods besides dummy coding for coding categorical variables for analysis in MR; we illustrated effect coding and criterion scaling. The different methods will provide the same overall results, but different contrasts from the regression coefficients.

Categorical and Continuous Variables, Interactions, and Curves

Our primary interest in discussing the analysis of categorical variables in MR was as preparation for combining categorical and continuous variables together in MR analyses. Analyses including both categorical and continuous variables are conceptually and analytically little different from those including only continuous variables. It is also possible to test for interactions between categorical and continuous variables. To do so, we centered the continuous variable and created a new variable that was the cross product of the dummy variable and the centered continuous variable. If there are multiple dummy variables, then there will also be multiple cross products. These cross products are then entered as the second, sequential step in a regression following the simultaneous regression with all other independent variables (including the categorical and continuous variables used to create the cross products). The statistical significance of the ΔR^2 associated with the cross products is the test of the statistical significance of the interaction. With multiple dummy variables, and thus multiple cross products, the ΔR^2 associated with the *block* of cross products is used to determine the statistical significance of the interaction.

Given the presence of a statistically significant interaction, the next step is to graph the interaction to provide an understanding of its nature, perhaps followed by additional regressions across the values of the categorical variable or other post hoc probing. Tests of predictive bias and attribute–treatment interactions are specific examples of analyses that should use this MR approach. ANCOVA can also be considered as MR with categorical and continuous variables, but researchers using MR can also test for possible interactions between the covariate and the treatment.

It is equally possible to test for interactions between two continuous variables in MR. The same basic procedure is used: the continuous variables are centered and multiplied, and this cross product is entered sequentially in a regression equation. Follow-up of this type of interaction may be a little more difficult, but the first step again is generally to graph the interaction. Several methods were discussed for graphing and exploring interactions between continuous variables. All types of interactions are often well described using the phrase "it depends."

A special type of interaction between continuous variables is when a variable interacts with itself, meaning that its effects depend on the *level* of the variable. For example, we found that the effect of homework depends on the amount of homework being discussed; homework has a stronger effect on achievement for fewer hours of homework than for higher levels of homework. This type of interaction shows up as curves in the regression line. We test for curves in the regression line by multiplying a variable times itself and then entering this squared variable last in a combined simultaneous–sequential regression. We can test for more than one curve by entering additional product terms (variable-cubed, to the fourth power, etc.). Again, graphs were recommended as a method for understanding the nature of these curvilinear effects.

Moderation, Mediation, and Common Cause

Interactions in multiple regression also go by the name of "moderation." To say that sex moderates the effect of self-concept on achievement means the same thing as saying that sex and self-concept interact in their effect on achievement, or that self-concept has differential effects on achievement by sex. Why do we use different terms to mean what is essentially the same thing? Thompson's contention that we do so to "confuse the graduate students" seems as plausible as any other (Thompson, 2006, p. 4). The term moderation is sometimes confused with that of mediation. Mediation describes the process by which one variable has an indirect effect on another variable through another mediating variable. If homework mediates the effect of motivation on achievement, this means that motivation affects homework, which in turn affects achievement. In Chapter 9 we discussed several methods for testing for mediation in multiple regression, but also noted that it is often easier to understand and test for mediation in the context of path analysis and SEM (as in Part 2). Indeed, we used path diagrams extensively to illustrate mediation. Although I tend to use the terms "mediation" and "indirect effect" fairly interchangeably, others suggest that the term mediation should be reserved for analyses involving longitudinal data (e.g., Kline, 2016, chap. 6). Fewer writers discuss the issue of common cause (and there are also several terms used to discuss this concept). A common cause is a variable that affects both our presumed influence and our presumed outcome; such variables *must* be included in multiple regression for the results to provide valid estimates of "effects." It is not unusual to see and hear this concept confused with that of moderation. When you hear researchers vaguely state that two variables likely interact in some way, pay attention. Do they really mean interaction/moderation? Or are they really talking about a potential common cause? Again, this is a topic that becomes clearer with the presentation of path diagrams (as used in Chapter 9) and is an important topic in Part 2 of this book.

ASSUMPTIONS AND REGRESSION DIAGNOSTICS

We have postponed discussion of several important topics until you had a more complete understanding of multiple regression and how to conduct and interpret results of multiple regression analyses. Now it is time to discuss assumptions underlying our multiple regressions, as well as how to diagnose various problems that can affect regression analyses and what to do about these problems. References are given to sources that provide more detail about these topics.

Assumptions Underlying Regression

What assumptions underlie our use of multiple regression? If we are to be able to trust our MR results and interpret the regression coefficients, we should be able to assume the following:

1. The dependent variable is a linear function of the independent variables.
2. Each person (or other observation) should be drawn independently from the population. Recall one general form of the regression equation: $Y = a + bX_1 + bX_2 + e$. This assumption means that the errors (*e*'s) for each person are independent from those of others.
3. The variance of the errors is not a function of any of the independent variables. The dispersion of values around the regression line should remain fairly constant for all values of *X*. This assumption is referred to as homoscedasticity.
4. The errors are normally distributed.

The first assumption (linearity) is the most important. If it is violated, then all of the estimates we get from regression—R^2, the regression coefficients, standard errors, tests of statistical significance—may be biased. To say the estimates are biased means that they will likely not reproduce the true population values. When assumptions 2, 3, and 4 are violated, regression coefficients are unbiased, but standard errors, and thus significance tests, will not be accurate. In other words, violation of assumption 1 threatens the meaning of the parameters we estimate, whereas violation of the other assumptions threatens interpretations from these parameters (Darlington, 1990, p. 110). Assumptions 3 and 4 are less critical, because regression is fairly robust to their violation (Kline, 1998). The violation of assumption 4 is only serious with small samples. We have already discussed methods of dealing with one form of nonlinearity (curvilinearity, in Chapter 8) and will discuss here and later methods for detecting and dealing with violations of the other assumptions.

In addition to these basic assumptions, to interpret regression coefficients as the *effects* of the independent variables on the dependent variable, we need to be able to assume that the errors are uncorrelated with the independent variables. This assumption further implies the following:

5. The dependent variable does not influence any of the independent variables. In other words, the variables we think of as causes must in fact be the causes, and those that we think of as the effects must be the effects.
6. The independent variables are measured without error, with perfect reliability and validity.
7. The regression must include all common causes of the presumed cause and the presumed effect (Kenny, 1979, p. 51).

We have already discussed assumptions 5 and 7 and will continue to develop them further in Part 2. Assumption 6 is a concern, because in the social sciences we rarely have perfect measurement. Again, we will discuss the implications of violation of this assumption in Part 2. There are a number of very readable, more detailed explanations of these seven assumptions. Allison (1999), Berry (1993), and Cohen and colleagues (2003) are particularly useful.

Regression Diagnostics

Here and in earlier chapters I noted that a good habit in any data analysis is to examine the data to make sure the values are plausible and reasonable. Always, always, always check your data. Regression diagnostics take this examination to another level and can be used to probe violations of assumptions and spot impossible or improbable values and other problems with data. In this section I will briefly describe regression diagnostics, illustrate their use for the data from previous chapters, and discuss what to do with regression diagnostic results. I will emphasize a graphic approach.

Diagnosing Violations of Assumptions

Nonlinearity

In Chapter 8, we examined how to deal with nonlinear data by adding powers of the independent variable to the regression equation. In essence, by adding both Homework and Homework^2 to the regression equation, we turned the nonlinear portion of the regression line into a linear one and were thus able to model the curve effectively using MR.

This approach thus hints at one method for determining whether we have violated the assumption of linearity: If you have a substantive reason to suspect that an independent variable may be related to the outcome in a curvilinear fashion, add a curve component (variable^2) to the regression equation to see whether this increases the explained variance.

The potential drawback to this approach is that the curve modeled by variable^2 may not adequately account for the departure from linearity. Therefore, it is useful to supplement this approach with a more in depth examination of the data using scatterplots. Rather than plotting the dependent variable of interest against the independent variable, however, we will plot the *residuals* against the independent variables; the residuals should magnify departures from linearity. Recall that the residuals represent the predicted values of the dependent variable minus the actual values of the dependent variable $(Y' - Y)$. They are the errors in prediction.

To illustrate, we will use the example from Chapter 8 that was used to illustrate testing for curves in MR: the regression of Grades on SES, previous Achievement, and time spent on Homework out of school. The addition of a Homework^2 variable was statistically significant, indicating (and correcting) a departure from linearity in the regression. Let's see if we can pick up this nonlinearity using scatterplots.

I reran the initial regression (without the Homework^2 variable and using the original uncentered metric) and saved the residuals (regression programs generally allow you to save unstandardized residuals as an option). Figure 10.1 shows the plot of the residuals against the original variable Homework. Note the two lines in the graph. The straight, horizontal line is the mean of the residuals. The line should also represent the regression line of the residuals on Homework. That line would be horizontal because the residuals represent Grades with the effects of Homework (and the other independent variables) *removed.* Because Homework has been removed, it is no longer related to the residuals. Recall that when two variables are unrelated our best prediction for Y is its mean for all values of X. The regression line is thus equal to the line drawn through the mean of the residuals. The other, almost straight line is what is called a *lowess* (or *loess*) fit line, which represents the *nonparametric* best fitting

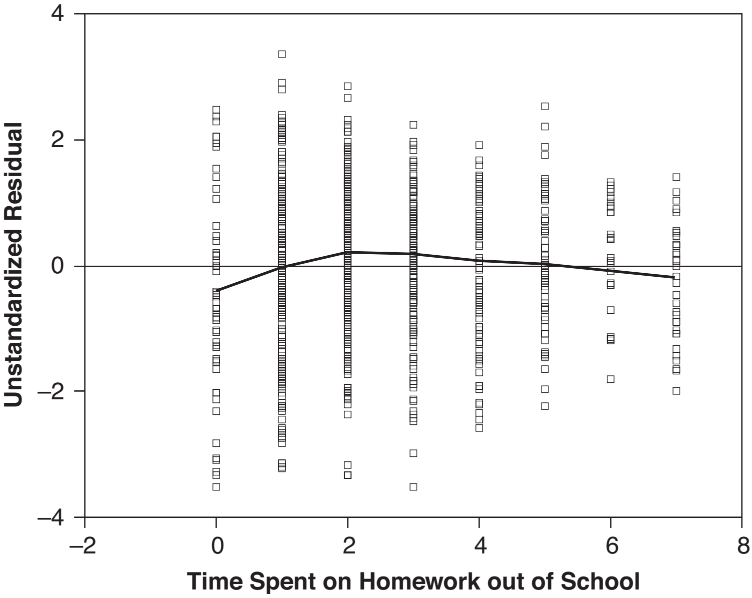


Figure 10.1 Plot of the unstandardized residuals against one independent variable (Homework). The lowess line is fairly straight.

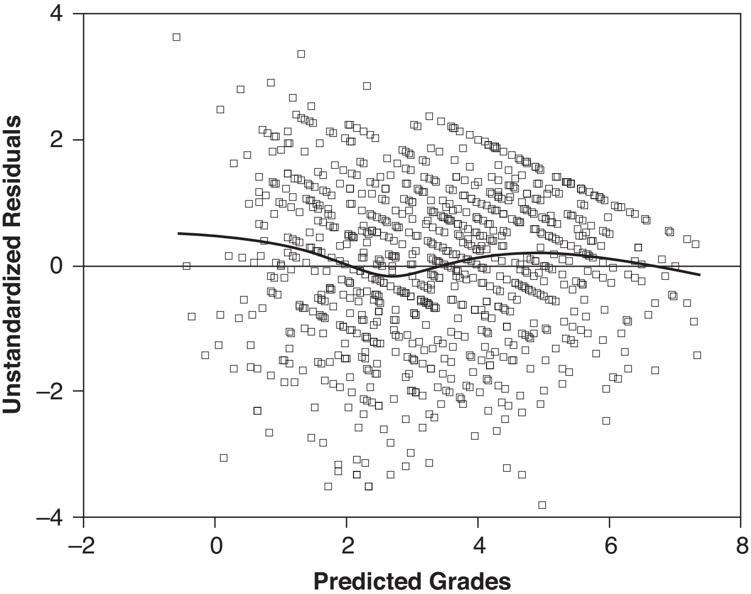


Figure 10.2 Plot of unstandardized residuals against the predicted Grades (a composite of the independent variables).

line, one that does not impose the requirement of linearity. Most computer programs can easily add this line to a regression scatterplot.

If there is no departure from linearity in the data, we would expect the lowess line to come close to the regression line; Cohen and colleagues note that the lowess line should look like "a young child's freehand drawing of a straight line" (2003, p. 111). With a significant departure from linearity, you would expect the lowess line to be curved, something more similar to the curvilinear regression lines shown in Chapter 8 (e.g., Figure 8.10) but without the upward slope. The lowess line in this plot indeed approaches the straight regression line. Figure 10.2 shows another useful plot: the residuals and the predicted values for the Grades dependent variable. Recall in Chapter 3 that we demonstrated that the predicted *Y* is an optimally weighted composite of the independent variables. It is, then, a variable that represents all independent variables in combination. Again, the lowess line comes close to the regression line and does not suggest a departure from linearity.

In this example, the test of the addition of a curve component (Chapter 8) was more successful in spotting a departure from linearity than was the inspection of data through scatterplots. This will not always be the case, and thus I recommend that you use both methods if you suspect a violation of this assumption. If theory or inspection suggests a departure from linearity, a primary method of correction is to build nonlinear terms into the regression (e.g., powers, logarithms). The method is discussed in Chapter 8; see also Cohen and colleagues (2003) and Darlington and Hayes (2017) for more depth.

Nonindependence of Errors

When data are not drawn independently from the population, we risk violating the assumption that errors (residuals) will be independent. As noted in the section on multilevel modeling in the next chapter, the NELS data, with students clustered within schools, risks violation of this assumption. Violation of this assumption does not affect regression coefficients but does affect standard errors. When clustered as described, we risk underestimating standard errors and thus labeling variables as statistically significant when they are not. This danger is obviated, to some degree, with large samples like the NELS data used here, especially when we are more concerned with the magnitude of effects than with statistical significance.

Are the residuals from the regression of Grades on SES, Previous Achievement, and Homework nonindependent? Is there substantial variation within schools? Unfortunately, this assumption is difficult to test with the NELS data included on the Web site because, with the subsample of 1000 cases, few of the schools had more than one or two students. Therefore, I used the original NELS data and selected out 414 cases from 13 schools. I conducted a similar regression analysis (Grades on SES, Previous Achievement, and Homework) and saved the residuals.

One way to probe for the violation of this assumption is through a graphing technique called *boxplots*. The boxplots of residuals, clustered by schools, are shown in Figure 10.3. The center through each boxplot shows the median, with the box representing the middle 50% of cases (from the 25th to the 75th percentile). The extended lines show the high and low values, excluding outliers and extreme values. For the purpose of exploring the assumption of independence of errors, our interest is in the variability of the boxplots. There is some variability up and down by school, and thus this clustering may indeed be worth taking into account. Another, quantitative test of the independence of observations uses the intraclass correlation coefficient, which compares the between-group (in this case, between-schools) variance to the total variance (for an example, see Stapleton, 2006). The intraclass correlation could be computed on the residuals or on a variable (e.g., Homework) that you suspect might vary across schools.

One option for dealing with a lack of independence of errors is to include categorical variables (e.g., using criterion scaling; see Chapter 6) that take the clustering variable into account. Another option is the use of multilevel or hierarchical linear modeling, discussed briefly in the next chapter. This assumption can also be violated in longitudinal designs in which the same tests or scales are administered repeatedly. We will deal with this issue briefly in Part 2.

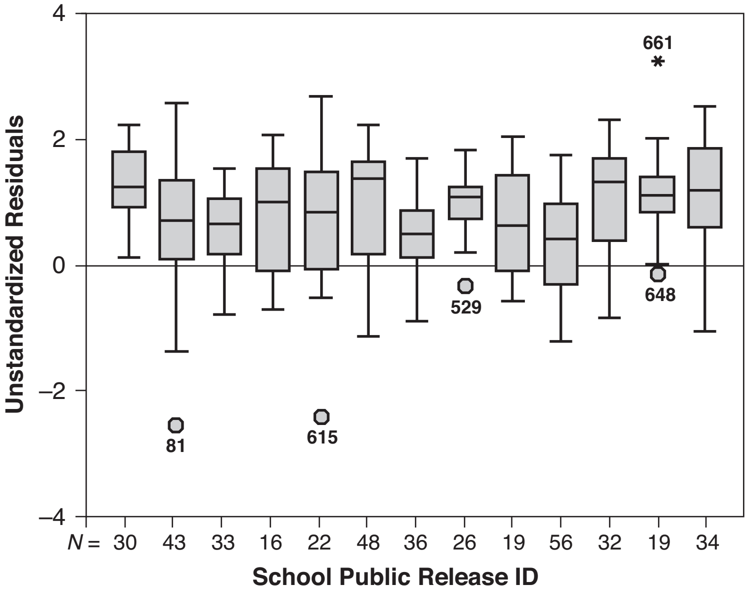


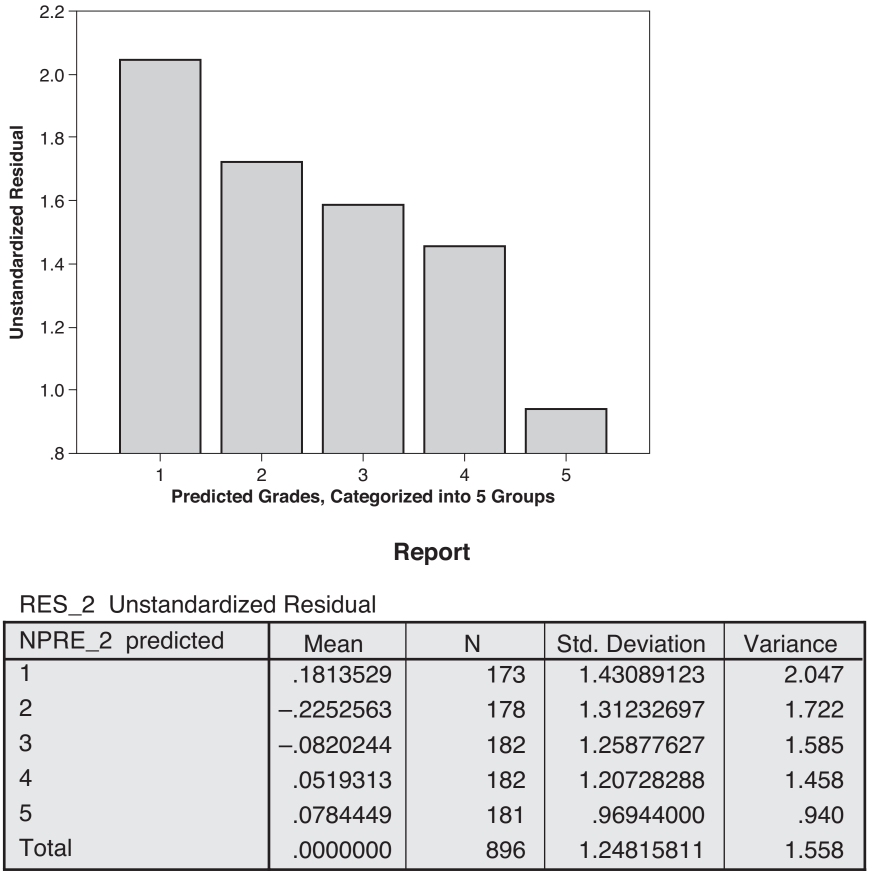
Figure 10.3 Boxplots of residuals, grouped by the school from which NELS students were sampled. The data are 414 cases from the full NELS data.

Homoscedasticity

We assume that the variance of errors around the regression line is fairly consistent across levels of the independent variable. In other words, the residuals should spread out consistently across levels of *X*. Violation of this assumption affects standard errors and thus statistical significance (not the regression coefficients), and regression is fairly robust to its violation. Scatterplots of residuals with independent variables or predicted values are also helpful for examining this assumption.

Return to Figure 10.1, the scatterplot of Homework with the Residuals from the regression of Grades on SES, Previous Achievement, and Homework. Although the residuals are spread out more at lower levels of homework than at upper levels, the difference is slight; visual inspection suggests that heteroscedasticity (the opposite of homoscedasticity) is not a problem. A common pattern of heteroscedasticity is a fan shape with, for example, little variability at lower levels of Homework and large variability at higher levels of Homework. Butterfly shapes are also possible (residuals constricted around the middle level of Homework), as is the opposite shape (a bulge in the middle).

Focus again on Figure 10.2. Notice how the residuals bunch up at higher levels of the Predicted *Y;* the plot has something of a fan shape, narrowing at upper levels of the predicted values. Do these data violate the assumption of homoscedasticity? To test this possibility, I collapsed the predicted Grades into five equal categories so that we can compare the variance of the residuals at each of these five levels. The data are displayed in Figure 10.4 as both a bar chart and table. As shown in the table, for the lowest category of predicted values, the variance of the residuals was 2.047, versus .940 for the highest category. There is a difference, but it is not excessive. One rule of thumb is that a ratio of high to low variance of less than 10 is not problematic. Statistical tests are also possible (Cohen et al., 2003).



Report

RES_2 Unstandardized Residual

NPRE_2 predicted	Mean	N	Std. Deviation	Variance
1	.1813529	173	1.43089123	2.047
2	–.2252563	178	1.31232697	1.722
3	–.0820244	182	1.25877627	1.585
4	.0519313	182	1.20728288	1.458
5	.0784449	181	.96944000	.940
Total	.0000000	896	1.24815811	1.558

Figure 10.4 Comparison of the variance of residuals for different levels of predicted Grades.

Normality of Residuals

The final assumption we will deal with is that the errors, or residuals, are normally distributed. What we are saying with this assumption is that if we plot the values of the residuals they will approximate a normal curve. This assumption is fairly easily explored because most MR software has tools built in to allow such testing.

Figure 10.5 shows such a plot: a bar graph of the residuals from the NELS regression of Grades on SES, Previous Achievement, and Homework (this graph was produced as one of the plot options in regression in SPSS). The superimposed normal curve suggests that the residuals from this regression are indeed normal. Another, more exacting, method is what is known as a q–q plot (or, alternatively, a p–p plot) of the residuals. A q–q plot of the residuals shows the value of the residuals on one axis and the expected value (if they are normally distributed) of the residuals on the other. Figure 10.6 shows the q–q plot of the residuals from the Grades on SES, Previous Achievement, Homework regression. If the residuals are normally distributed, the thick line (expected versus actual residuals) should come close to the diagonal straight line. As can be seen from the graph, the residuals conform fairly well to the superimposed straight line. The reason this method is more exact is that it is easier to spot a deviation from a straight line than a normal curve (Cohen et al., 2003). Some programs (e.g., SPSS) produce a p–p plot of the residuals as an option in multiple regression. A p–p plot

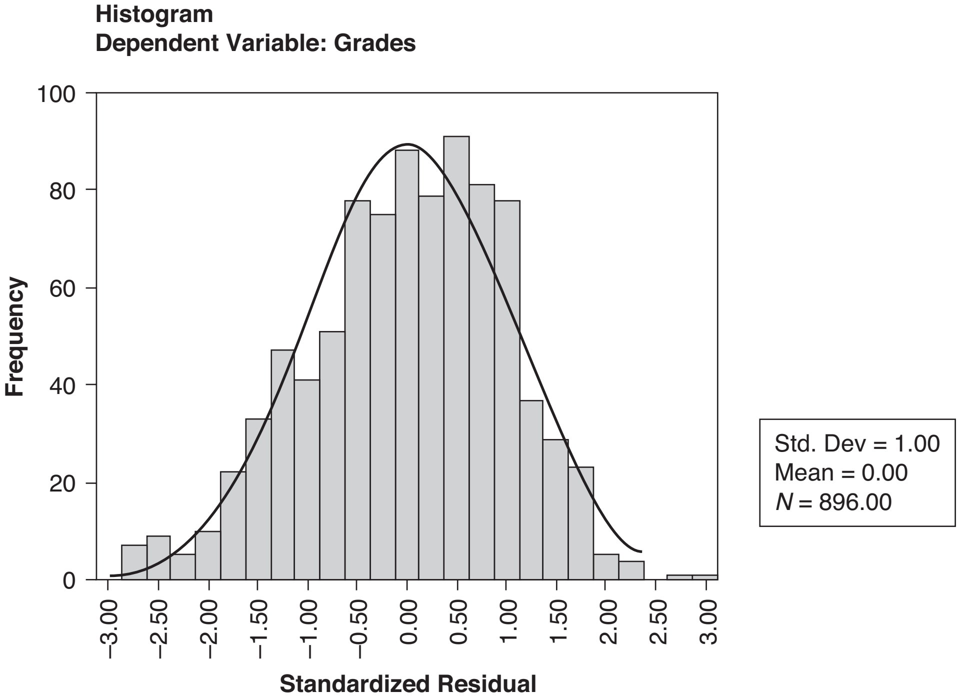


Figure 10.5 Testing for the normality of residuals. The residuals form a nearly normal curve.

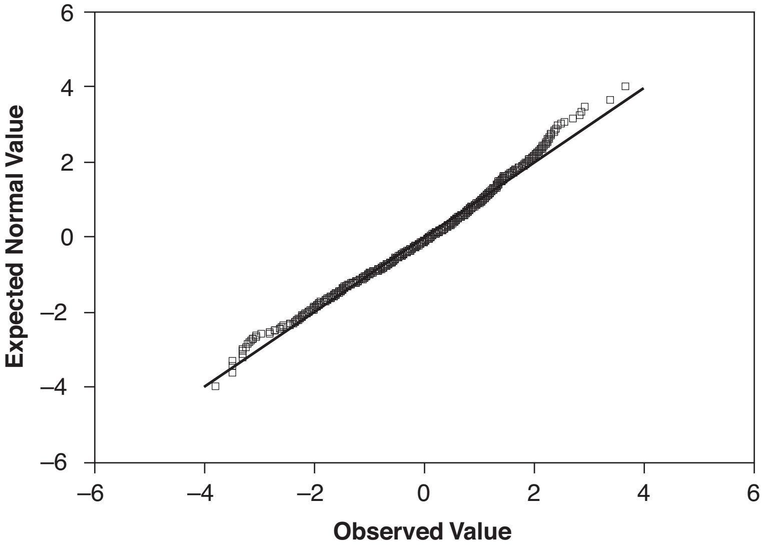


Figure 10.6 A q–q plot of the residuals. The residuals' adherence to a nearly straight line supports their normality

uses the cumulative frequency and is interpreted in the same fashion (looking for departures from a straight line).

Excessive heteroscedasticity and nonnormal residuals can sometimes be corrected through transformation of the dependent variable. Eliminating subgroups from the regression may also be useful. Finally, there are alternative regression methods (e.g., weighted least squares regression) that may be useful when these assumptions are seriously violated (see Cohen et al., 2003, and Darlington, 1990, for more information).

Diagnosing Data Problems

Regression diagnostics for spotting problematic data points focus on three general characteristics: distance, leverage, and influence. Conceptually, how would you spot unusual or problematic cases, commonly referred to as outliers or as extreme cases? Focus on Figure 10.7, a reprint of the earlier Figure 3.7. The figure is a byproduct of the regression of students' Grades on Parent Education and Homework. Recall that we saved the variable Predicted Grades, which I demonstrated was an optimally weighted composite of the two independent variables, weighted so as to best predict the outcome. The figure shows students' GPA plotted against their Predicted GPAs. Note the case circled in the lower right of the figure. This case is among the farthest from the regression line; this is one method of isolating an extreme case, called *distance. Leverage* refers to an unusual pattern on the independent variables and does not consider the dependent variable. If you were using homework in different academic areas to predict overall GPA, it would not be unusual to find a student who spent 1 hour per week on math homework nor would it be unusual to find a student who spent 8 hours per week on English homework. It would likely be unusual to find a student who combined these, who spent only 1 hour per week on math while spending 8 hours per week on English. This case would likely have high leverage. Because leverage is not calculated with respect to the dependent variable, the graph shown here may not be informative as to leverage; a graph of the two independent variables may be more useful (as we will soon see). The final characteristic of interest is

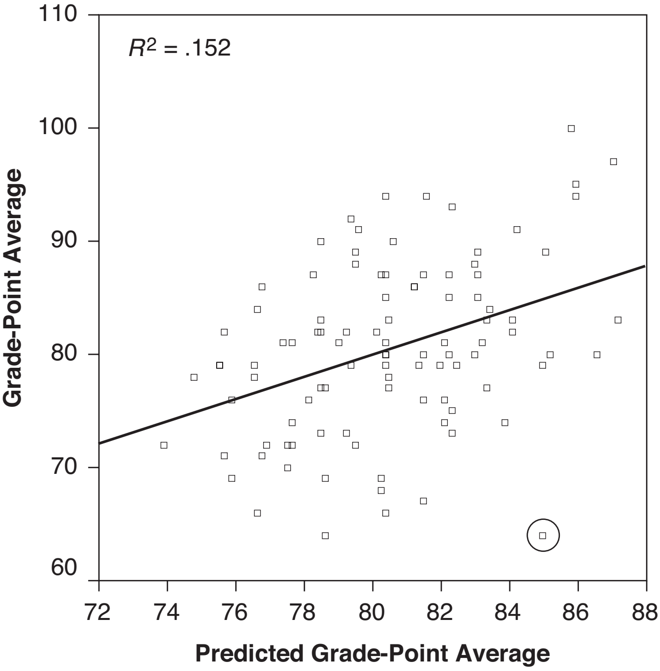


Figure 10.7 Predicted versus actual Grades plot from Chapter 3. The circled case is a potential extreme case, a long distance from the regression line.

influence. As the name implies, a case that has high influence is one that, if removed from the regression, results in a large change in the regression results. Cases with high influence are those that are high on both distance and leverage. The circled case would likely fit this description as well. If it were deleted from the regression, the regression line would likely be somewhat steeper than it is in the figure.

Distance

Common measures of distance are derived from the residuals. In Figure 10.7, the residual for the circled case is the point on the regression line above the case (approximately 85) minus the actual value of the case (64). This definition matches well the conceptual definition of distance given previously.

In practice, the unstandardized residuals are less useful than are standardized versions of residuals. Table 10.1 shows some of the cases from this data set. The first column shows the case number, followed by the dependent variable Grades and the two independent variables Parent Education and Homework. Column five shows the Predicted Grades used to create the graph in Figure 10.7. The remaining columns show various regression diagnostics. The first row of the table shows the names assigned these variables in SPSS, under which I have included a brief explanation. Column six, labeled ZRE_1, shows the standardized residuals, which are the residuals standardized to approximately a normal distribution. Think of them like z scores, with values ranging from 0 (very close to the regression line) to ±3 or more. The next column (SRE_1) represents the standardized residuals converted to a t distribution (the t distribution is also referred to as Student's t, hence the S), which are generally called the studentized or t residuals. The advantage of this conversion is that the t residuals may be tested for statistical significance (see Darlington, 1990, p. 358). In practice, however, researchers often simply examine large positive or negative standardized or studentized residuals or, with reasonable sample size, those greater than an absolute value of 2 (with very large samples, there may be many of these).

The cases shown in Table 10.1 were chosen for display because they have high values for distance, leverage, or influence. As shown in the table, cases 34 (–3.01) and 83 (2.06) show high values for studentized residuals.

Figure 10.8 shows the same plot of Predicted and actual Grades, with a few of the cases identified. Note the case that was originally circled is case number 34, the highest negative studentized (and standardized) residual. As can be seen, case 83, with a high positive standardized residual, is also far away from the regression line. It might be worth investigating these cases with high residuals further to make sure that they have been coded and entered correctly.

Leverage

Leverage gets at the unusualness of a pattern of independent variables, without respect to the dependent variable. The column in Table 10.1 labeled LEV_1 provides an estimate of leverage (this measure is also often referred to as h). Leverage ranges from 0 to 1, with an average value of $(k + 1)/ n$ (k = number of independent variables); twice this number has been suggested as a rule of thumb for high values of leverage (Pedhazur, 1997, p. 48). Case 16 in the table had the highest value for leverage (.098), followed by cases 36 (.088) and 32 (.084). Both these values are higher than the rule of thumb would suggest:

$$2\left(\frac{k+1}{N}\right)=2\left(\frac{3}{100}\right)=.06.$$

Table 10.1 Regression Diagnostics for the Regression of Grades on Parent Education and Homework (Data from Chapter 3).

Casenum	*Grades*	*Pared*	*Hwork*	*Predgrad*	*ZRE_1*	*SRE_1*	*SDR_1*	*COO_1*	*LEV_1*	*SDB0_1*	*SDB1_1*	*SDB2_1*
					standardized residual	*Studentized, t residual*	*tResid, deleted*	*Cook*	*Leverage*	*Standardized DF Beta*		
										intercept	*pared*	*hwork*
12.00	72.00	13.00	5.00	79.48435	−1.05539	−1.06231	−1.06302	0.00495	0.00299	−0.07044	0.05836	−0.01163
13.00	66.00	12.00	3.00	76.63804	−1.50010	−1.52071	−1.53122	0.02134	0.01693	−0.19095	0.12503	0.11783
14.00	79.00	14.00	4.00	79.36713	−0.05177	−0.05211	−0.05184	0.00001	0.00303	−0.00098	−0.00072	0.00287
15.00	76.00	10.00	4.00	75.88464	0.01627	0.01673	0.01664	0.00001	0.04405	0.00377	−0.00347	0.00009
16.00	80.00	20.00	6.00	86.56656	−0.92597	−0.98069	−0.98049	0.03901	0.09848	0.30209	−0.32258	0.04489
17.00	91.00	15.00	8.00	84.18914	0.96042	0.97535	0.97510	0.00994	0.02038	−0.04474	0.01145	0.13224
32.00	83.00	15.00	11.00	87.15267	−0.58558	−0.61536	−0.61338	0.01317	0.08446	0.03559	0.01979	−0.18448
33.00	78.00	13.00	6.00	80.47220	−0.34861	−0.35156	−0.34996	0.00070	0.00669	−0.02205	0.02422	−0.02180
34.00	64.00	17.00	7.00	84.94254	−2.95316	−3.00886	−3.14360	0.11492	0.02668	0.45737	−0.42923	−0.16864
35.00	82.00	13.00	4.00	78.49651	0.49404	0.49765	0.49571	0.00121	0.00448	0.03455	−0.02020	−0.01998
36.00	81.00	17.00	1.00	79.01546	0.27984	0.29462	0.29322	0.00313	0.08776	−0.03788	0.06746	−0.07800
37.00	73.00	13.00	4.00	78.49651	−0.77508	−0.78075	−0.77917	0.00298	0.00448	−0.05430	0.03175	0.03141
80.00	72.00	10.00	5.00	76.87248	−0.68708	−0.70760	−0.70576	0.01012	0.04714	−0.15793	0.15778	−0.04060
81.00	79.00	17.00	4.00	81.97900	−0.42008	−0.42961	−0.42780	0.00283	0.03391	0.05808	−0.07712	0.04376
82.00	93.00	14.00	7.00	82.33067	1.50451	1.51942	1.52989	0.01533	0.00954	0.01337	−0.04408	0.15087
83.00	100.00	18.00	7.00	85.81316	2.00052	2.05698	2.09249	0.08073	0.04414	−0.41586	0.40491	0.08101
84.00	90.00	13.00	4.00	78.49651	1.62214	1.63401	1.64841	0.01307	0.00448	0.11487	−0.06717	−0.06645
85.00	69.00	10.00	4.00	75.88464	−0.97082	−0.99817	−0.99815	0.01898	0.04405	−0.22643	0.20833	−0.00511

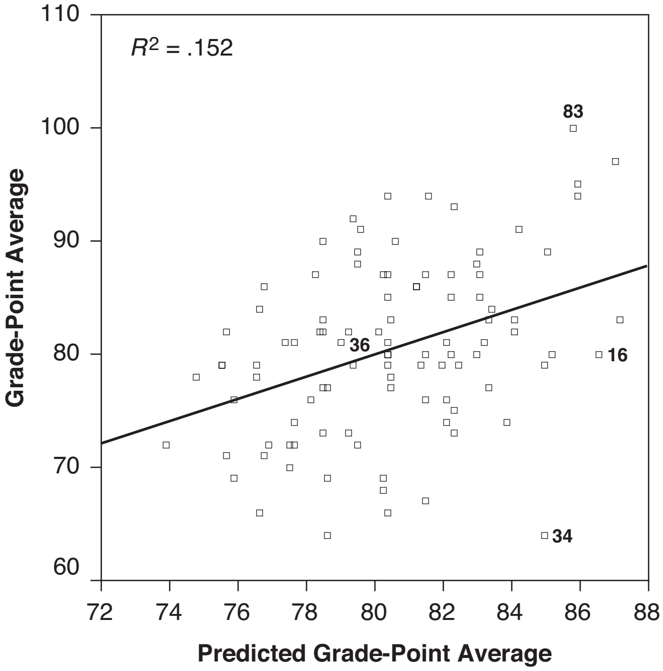


Figure 10.8 Plot from Figure 10.7 with several noteworthy cases highlighted.

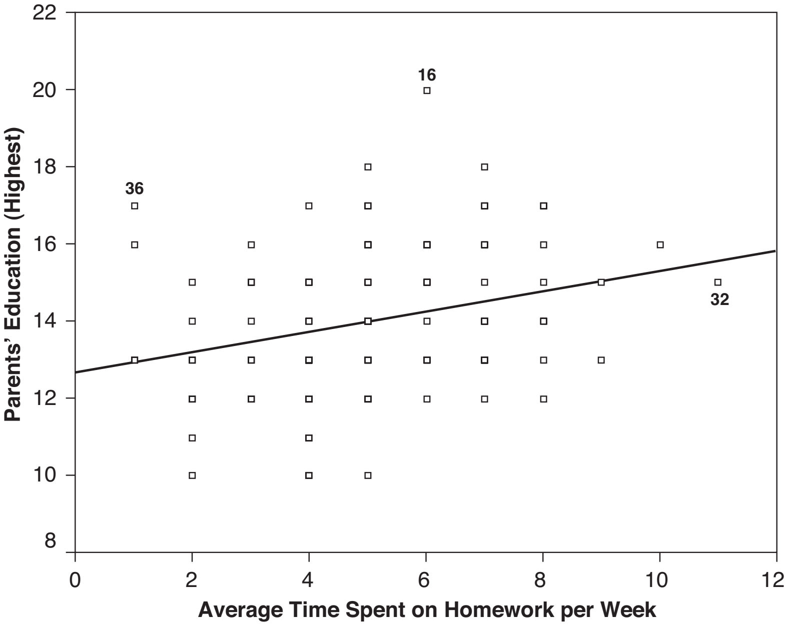


Figure 10.9 Leverage illustrated.

As can be seen in Figure 10.8, you might suspect that case 16 was unusual from a visual display (because it is on one edge of the graph), but case 36 is right in the middle of the graph. Recall, however, that leverage does not depend on the dependent variable. Figure 10.9 shows a plot of the two independent variables. Cases 16, 36, and 32 are outside the "swarm" of most of the cases; they indeed represent an unusual combination of independent variables. These cases may also be worth checking.

Influence

Influence means what the name suggests: a case that is highly influential on the intercept or the regression line. The column labeled Coo_1 (for Cook's Distance) in Table 10.1 provides values of an estimate of influence; cases with large values are worth inspecting. The cases with the largest Cook's D values were cases 34 (.115) and 83 (.081). The regression plane would move the most if these cases were omitted.

Most computer programs also compute estimates of *partial* influence (as in influence, with the effects of the other independent variables accounted for). The DF Betas, standardized, listed in the last three columns are estimates of partial influence. The first of these columns (SDB0_1) pertains to the regression intercept, the second (SDB1_1) to the first independent variable (Parent Education), and the third (SDB2_1) to the second independent variable (Homework). The values shown are the change in each parameter, if a particular case were removed. A negative value means that the particular case lowered the value of the parameter, whereas a positive value means that the case raised the parameter. So, for example, case 34 had standardized DF Beta values of .457, –.429, and –.169. Case 34 served to raise the intercept and lower the regression coefficient for Parent Education and Homework. Although the unstandardized DF Betas are not shown in Table 10.1, they were 2.29, –.158, and –.058. If you run the regression without case 34, you will find that the intercept reduces by 2.29, the Parent Education *b* increases by .158, and the Homework *b* increases by .058.

An inspection of the standardized DF Betas showed large negative values by case 83 for the intercept (–.416) and large positive value for case 34 (.457). These two cases were also very influential for the Parent Education regression coefficient, although reversed: case 34 (–.429), case 83 (.405). The partial influence values for the Homework variable were considerably smaller. Cases 21 and 29 had the highest values (.334 and .335).

Uses

What do these various regression diagnostics tell us? In the present example, cases 34 and 83 showed up across measures; it would certainly be worth inspecting them. But inspecting them for what? Sometimes these diagnostics can point out errors or misentered data. A simple slip of the finger may cause you to code 5 hours of homework as 50. This case will undoubtedly show up in the regression diagnostics, thus alerting you to the mistake. Of course, a simple careful inspection of the data will likely spot this case as well! Think about the example I used initially to illustrate leverage, however, someone who reports 1 hour of Math Homework and 8 hours of English Homework. This case will not show up in a simple inspection of the data, because these two values are reasonable and, taken by themselves, only become curious when taken together. The case will likely be spotted in an analysis of both leverage and influence; we might well discover that errors were made in entering this datum as well.

If there are not obvious errors for the variables spotted via regression diagnostics, then what? In our present example, cases 34 and 83, although outliers, are reasonable. A check of the raw data shows that case 34 had well-educated parents, higher than average homework, but poor grades. Case 83 simply had an excellent GPA and higher than average homework. On further investigation, I might discover that case 34 had a learning disability, and I might decide to delete this case and several other similar cases. Or I might decide that the variation is part of the phenomenon I am studying and leave case 34 in the analysis. Another option is additional analysis. If a number of outliers share characteristics in common and are systematically different from other cases, it may suggest that a different regression is needed for these participants or the advisability of including an interaction term in the analysis (e.g., Disability Status by Parent Education). It might also suggest the inclusion of an important common cause (e.g., disability status affecting both time spent on homework and subsequent grades).

Obviously, unless clear-cut errors are involved, considerable judgment is involved in the inspection of regression diagnostics. Note that deletion of case 34 will increase the regression weight for Homework; if I did delete this case, I will need to be sure that my deletion is based on a concern about its extremity rather than a desire to inflate the apparent importance of my findings. If you do delete cases based on regression diagnostics, you should note this in the research write-up and the reasons for doing so. With the present example and after examining cases with high values on all the regression diagnostics, I would first double-check each of these values against the raw data but would likely conclude in the end that all the cases simply represented normal variation. I would then leave the data in their present form.

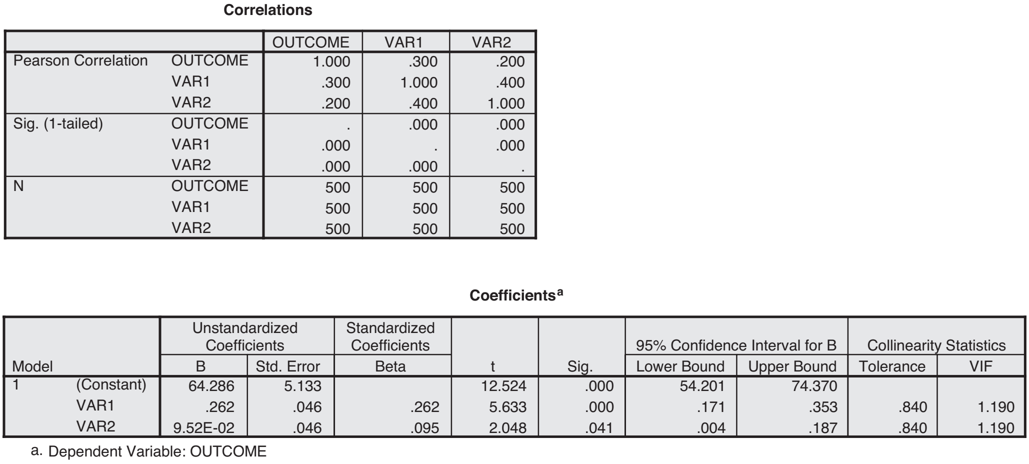
Again, I have barely scratched the surface of an important topic; it is worth additional study. Darlington (1990, chap. 14), Darlington and Hayes (2017), Fox (2008), and Pedhazur (1997) each devote chapters to regression diagnostics and are worth reading.

Multicollinearity

I mentioned briefly when discussing interactions the potential problem of multicollinearity (also called collinearity). Briefly, multicollinearity occurs when several independent variables correlate at an excessively high level with one another or when one independent variable is a near linear combination of other independent variables. Multicollinearity can result in misleading and sometimes bizarre regression results.

Figure 10.10 shows some results of the regression of a variable named Outcome on two independent variables, Var1 and Var2. The correlations among the three variables are also shown. The results are not unusual and suggest that both variables have positive and statistically significant effects on Outcome.

Now focus on Figure 10.11. For this analysis, the two independent variables correlated at the same level with the dependent variable as in the previous example (.3 and .2). However, in this example, Var1 and Var2 correlate .9 with each other (versus .4 in the previous example). Notice the regression coefficients. Even though all variables correlate positively with



Correlations

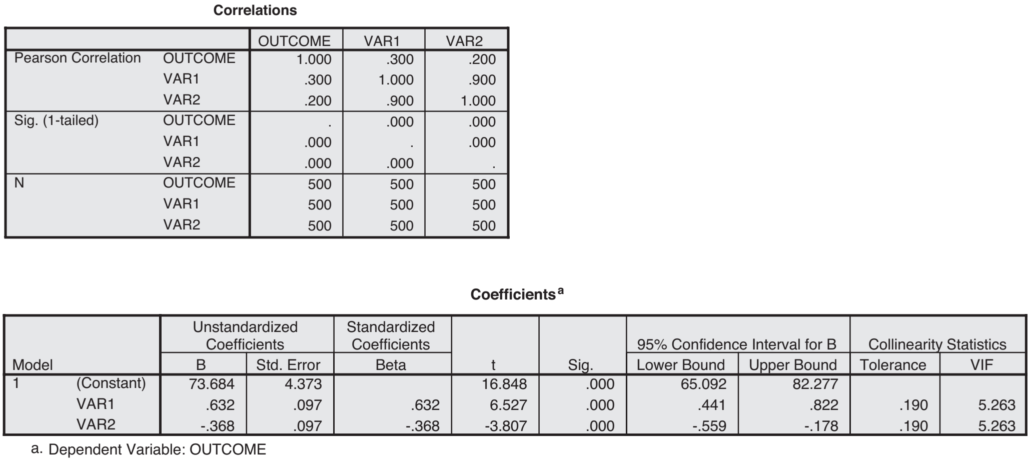
		OUTCOME	VAR1	VAR2
Pearson Correlation	OUTCOME	1.000	.300	.200
	VAR1	.300	1.000	.400
	VAR2	.200	.400	1.000
Sig. (1-tailed)	OUTCOME	.	.000	.000
	VAR1	.000	.	.000
	VAR2	.000	.000	.
N	OUTCOME	500	500	500
	VAR1	500	500	500
	VAR2	500	500	500

Coefficients[a]

		Unstandardized Coefficients		Standardized Coefficients			95% Confidence Interval for B		Collinearity Statistics	
Model		B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound	Tolerance	VIF
1	(Constant)	64.286	5.133		12.524	.000	54.201	74.370		
	VAR1	.262	.046	.262	5.633	.000	.171	.353	.840	1.190
	VAR2	9.52E-02	.046	.095	2.048	.041	.004	.187	.840	1.190

a. Dependent Variable: OUTCOME

Figure 10.10 Regression of Outcome on Var1 and Var2. The results are reasonable.



Correlations

		OUTCOME	VAR1	VAR2
Pearson Correlation	OUTCOME	1.000	.300	.200
	VAR1	.300	1.000	.900
	VAR2	.200	.900	1.000
Sig. (1-tailed)	OUTCOME	.	.000	.000
	VAR1	.000	.	.000
	VAR2	.000	.000	.
N	OUTCOME	500	500	500
	VAR1	500	500	500
	VAR2	500	500	500

Coefficients[a]

		Unstandardized Coefficients		Standardized Coefficients			95% Confidence Interval for B		Collinearity Statistics	
Model		B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound	Tolerance	VIF
1	(Constant)	73.684	4.373		16.848	.000	65.092	82.277		
	VAR1	.632	.097	.632	6.527	.000	.441	.822	.190	5.263
	VAR2	-.368	.097	-.368	-3.807	.000	-.559	-.178	.190	5.263

a. Dependent Variable: OUTCOME

Figure 10.11 Regression of Outcome on Var1 and Var2 when Var1 and Var2 are very highly correlated (collinear). The results are puzzling, and the interpretation will likely be misleading.

one another, Var1 seems to have a positive effect on Outcome, whereas Var2 has a negative effect. As noted previously, multicollinearity can produce strange results such as these; standardized regression coefficients greater than 1 are also common. Notice also that the standard errors of the *b*'s are also considerably larger for the second example than for the first. Multicollinearity also inflates standard errors; sometimes two variables will correlate at similar levels with an outcome, but one will be a statistically significant predictor of the outcome, while the other will not, as a result of multicollinearity.

Conceptually, multicollinearity suggests that you are trying to use two variables in a prediction that overlap completely or almost completely with one another. Given this definition, it makes intuitive sense that multicollinearity should affect standard errors: the more that variables overlap, the less we can separate accurately the effects of one versus the other. Multicollinearity is often a result of a researcher including multiple measures of the same construct in a regression. If this is the case, one way to avoid the problem is to combine the overlapping variables in some way, either as a composite or, as is done in Part 2, using the variables as indicators of a latent variable. Multicollinearity is also often a problem when researchers use a kitchen-sink approach: throwing a bunch of predictors into regression and using stepwise regression, thinking it will sort out which are important and which are not.

Given the example, you may think you can spot multicollinearity easily by examining the zero-order correlations among the variables, with high correlations alerting you to potential problems. Yet multicollinearity can occur even when the correlations among variables are not excessive. A common example of such an occurrence is when a researcher, often inadvertently, uses both a composite and the components of this composite in the same regression. For example, in Figure 10.12 I regressed BYTests on grades in each academic area, in addition to a composite Grades variable (BYGrads). Notice the results: the overall R^2 is statistically significant, but none of the predictors is statistically significant. In this example, the largest individual correlation was .801, however, not overly large. The zero-order correlations are not always useful in spotting collinearity.

How can you avoid the effects of multicollinearity? Computer programs provide, on request, collinearity diagnostics. Such statistics are shown in Figures 10.10 through 10.12. Tolerance is a measure of the degree to which each variable is independent of (does not

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.558[a]	.311	.307	7.10940

a. Predictors: (Constant), BYGRADS GRADES COMPOSITE, BYS81B math88-grades, BYS81A English88-grade, BYS81D sstudies88-grades, BYS81C science88-grades

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	20119.51	5	4023.903	79.612	.000[a]
	Residual	44579.48	882	50.544		
	Total	64699.00	887			

a. Predictors: (Constant), BYGRADS GRADES COMPOSITE, BYS81B math88-grades, BYS81A English88-grade, BYS81D sstudies88-grades, BYS81C science88-grades

b. Dependent Variable: BYTESTS 8th-grade achievement tests (mean)

Coefficients[a]

		Unstandardized Coefficients		Standardized Coefficients			Collinearity Statistics	
Model		B	Std. Error	Beta	t	Sig.	Tolerance	VIF
1	(Constant)	33.123	5.370		6.168	.000		
	BYS81A English88-grade	8.19E-02	1.501	.009	.055	.956	.027	36.744
	BYS81B math88-grades	-.698	1.490	-.077	-.469	.639	.029	34.153
	BYS81C science88-grades	.767	1.499	.090	.511	.609	.025	39.871
	BYS81D sstudies88-grades	-.125	1.487	-.015	-.084	.933	.026	38.350
	BYGRADS GRADES COMPOSITE	6.241	6.008	.538	1.039	.299	.003	343.374

a. Dependent Variable: BYTESTS 8th-grade achievement tests (mean)

Figure 10.12 Another cause of multicollinearity. A composite and its components are both used in the regression.

overlap with) the other independent variables (Darlington & Hayes, 2017). Tolerance can range from 0 (no independence from other variables) to 1 (complete independence); larger values are desired. The variance inflation factor (VIF) is the reciprocal of tolerance and is "an index of the amount that the variance of each regression coefficient is increased" over that with uncorrelated independent variables (Cohen et al., 2003, p. 423). Small values for tolerance and large values for VIF signal the presence of multicollinearity. Cohen and colleagues (2003, p. 423) note that a common rule of thumb for a large value of VIF is 10, which means that the standard errors of *b* are more than three times as large as with uncorrelated variables $\left(\sqrt{10} = 3.16\right)$, but that this value is probably too high. Note that use of this value will lead toan inspection and questioning of the results in Figure 10.12, but not those in Figure 10.11. Values for the VIF of 6 or 7 may be more reasonable as flags for excessive multicollinearity (cf. Cohen et al., 2003). These values of the VIF correspond to tolerances of .10 (for a VIF of 10), .14 (VIF of 7), and .17 (VIF of 6), respectively.

Factor analysis of independent variables and "all subsets" regression can also be useful for diagnosing problems. When you get strange regression results, you should consider and investigate multicollinearity as a possible problem. Indeed, it is a good idea to routinely examine these statistics. A method known as ridge regression can be used when data are excessively collinear.

Obviously, I have just touched the surface of this important topic; it is worth additional study. Pedhazur (1997) presents a readable, more detailed discussion of the topic, as does Darlington (1990, chaps. 5, 8). Darlington and Hayes (2017, chap. 4) offer useful suggestions for dealing with collinearity.

SAMPLE SIZE AND POWER

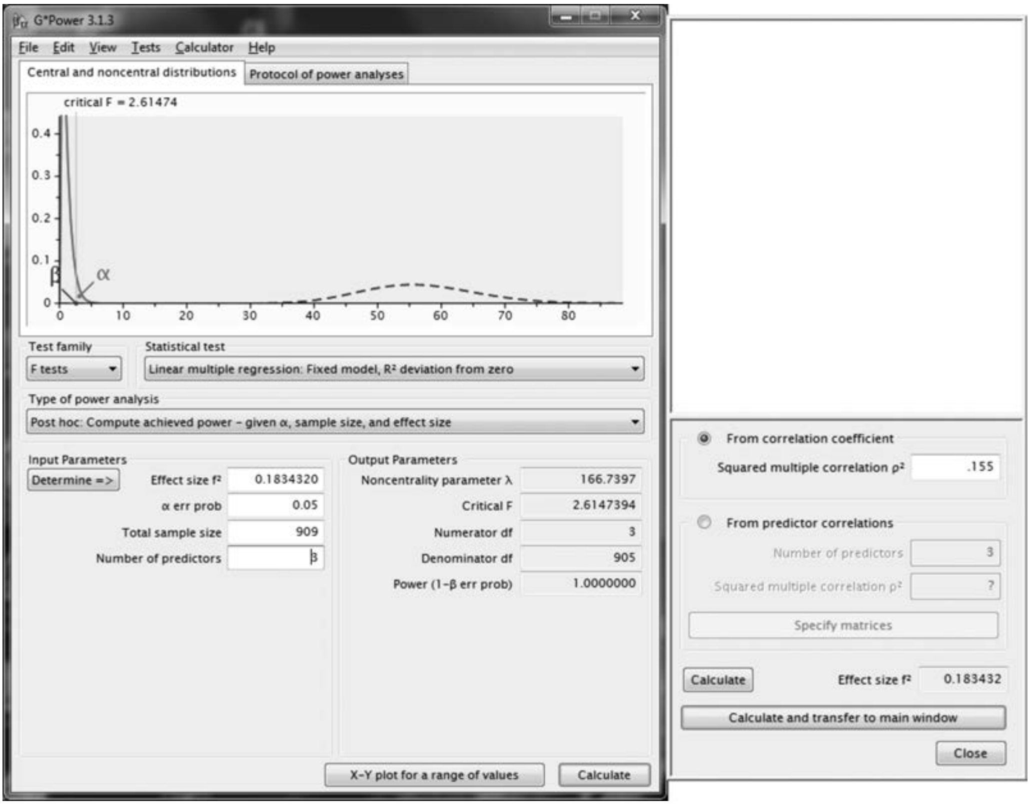
"How large a sample do I need?" Anyone who has advised others on the use of multiple regression (or any other statistical method) has heard this question more times than he or she can count. This question may mean several things. Some who ask it are really asking, "Is there some minimum sample size that I can't go below in MR?" Others are looking for a rule of thumb, and there is a common one: 10 to 20 participants for each independent variable. Using this rule, if your MR includes 5 independent variables, you need at least 50 (or 100) participants. I've heard this rule of thumb many times but have no idea where it comes from. We will examine it shortly to see if it has any validity for the types of MR problems we have been studying. Finally, more sophisticated researchers will ask questions about what sample size they need to have a reasonable chance of finding statistical significance.

I hope you recognize this final version of the question as one of the *power* of MR. I have alluded to power at several points in this text (e.g., in the discussion of interactions in MR, testing for mediation), but, as you will see, we have really sidestepped the issue until this point by our use of the NELS data. With a sample size of 1000, we had adequate power for all the analyses conducted. You can't always count on sample sizes in the thousands, however, so let us briefly turn to the issue of power and sample size.

Briefly, power generally refers to the ability correctly to reject a false null hypothesis. It is a function of the magnitude of the effect (e.g., whether Homework has a small or a large effect on Grades); the alpha, or probability level chosen for statistical significance (e.g., .05, .01, or some other level); and the sample size used in the research. Likewise, the necessary sample size depends on effect size, chosen alpha, and desired power. The needed sample size increases as desired power increases, effect size decreases, and alpha gets more stringent (i.e., as the probability chosen gets smaller). Common values for power are .8 or .9, meaning that given a particular effect size one would like to have an 80% or 90% chance of rejecting a false null hypothesis of no effect. Like alpha, and despite conventions, power levels should be chosen based on the needs of a particular study.

This short section is, of course, no treatise on power analysis. What I do plan to do here is to examine power and sample size for the rule of thumb given previously, as well as some of the examples we have used in this book, to give you some sense of what sorts of sample sizes are needed with the kinds of problems used in this book. Fortunately, there are some excellent books on power analysis, including Cohen's classic book on the topic (1988). The Darlington and Hayes (2017) and Cohen and colleagues (2003) text is useful on this topic as well and many others; for experimental research, I found Howell's (2013) introduction to power especially clear. If you intend to conduct research using MR (or other methods), I recommend that you read further on this important issue. You should and can also have access to a program for conducting power analysis. The examples that follow use G*Power 3.1 (Faul, Erdfelder, Buchner, & Lang, 2009; Faul, Erdfelder, Lang, & Buchner, 2007), a free power analysis program available for download (www.gpower.hhu.de/, or just search for "GPower"). I have also used SamplePower from SPSS, and the PASS (Power Analysis and Sample Size) program from NCSS (www.ncss.com); they also are easy to use and work well.

First, let's examine several of the examples in this text. In Chapter 4, we regressed GPA in 10th grade on Parent Education, In School Homework, and Out of School Homework in a simultaneous regression. The R^2 for the overall regression was .155, with a sample size of 909.


G*Power 3.1.3
File Edit View Tests Calculator Help
Central and noncentral distributions
Protocol of power analyses
critical F = 2.61474
0.4
0.3
0.2
0.1
0
β
α
0
10
20
30
40
50
60
70
80
X-Y plot for a range of values
Calculate
Specify matrices
Calculate
Calculate and transfer to main window
Close

Test family: F tests

Statistical test: Linear multiple regression: Fixed model, R² deviation from zero

Type of power analysis: Post hoc: Compute achieved power – given α, sample size, and effect size

Input Parameters: Determine =>

Input Parameters		Output Parameters	
Effect size f²	0.1834320	Noncentrality parameter λ	166.7397
α err prob	0.05	Critical F	2.6147394
Total sample size	909	Numerator df	3
Number of predictors	3	Denominator df	905
		Power (1–β err prob)	1.0000000

● From correlation coefficient

Squared multiple correlation ρ² .155

○ From predictor correlations

Number of predictors 3

Squared multiple correlation ρ² ?

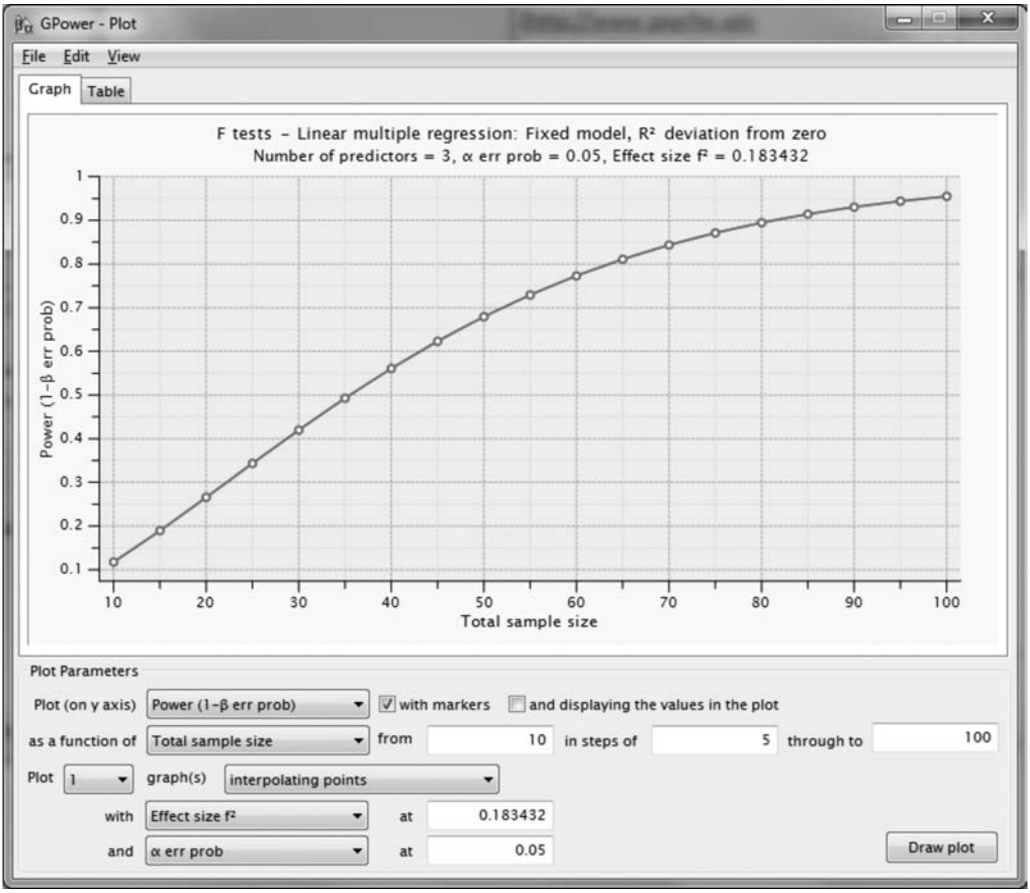
Effect size f² 0.183432

Figure 10.13 Power analysis for the overall regression of GPA on Parent Education, In-School Homework, and Out-of-School Homework from Chapter 4.

What sort of power did we have with this simultaneous regression? According to G*Power, this example had a power of 1.0 (for this and the other examples, I will assume an alpha of .05) for the overall regression. In other words, given the information previously, we had a 100% chance of correctly rejecting a false null hypothesis. Figure 10.13 shows the relevant screen shot. We are interested in an *F* test (Test Family), and are interested in the overall regression (e.g., the statistical significance of the overall R^2), so choose "Fixed model . . . R^2 deviation from zero." G*Power uses f^2 as its measure of effect size, but it is easy to convert R^2 and ΔR^2 into f^2 (see chapters 4 and 5); indeed, G*Power will do these calculations for you, as shown on the smaller right-hand screen (to get this screen, click on the "Determine" button under "Input Parameters."). The figure also shows the results.

These findings are for a post-hoc power analysis; that is, we conducted the regression and then wondered what the power was. Much more useful for most researchers is an a priori power analysis, in which we plan the research and then calculate the needed sample size to have a good chance of rejecting a false null hypothesis. With these three variables and an R^2 of .155, we will have a power of .8 with 64 participants and a power of .9 with 82 participants. Figure 10.14 shows a graph of power (Y-axis) as a function of sample size, given an alpha of .05 and an R^2 of .155.

We often are interested in the power of the addition of one variable or a block of variables to the regression equation, with other variables (background variables or covariates) controlled. For example, in Chapter 5 we considered the sequential regression in which we added Locus of Control and Self-Esteem to the regression, with SES and Previous Grades already


GPower - Plot
File Edit View
Graph Table
F tests - Linear multiple regression: Fixed model, R² deviation from zero
Number of predictors = 3, α err prob = 0.05, Effect size f² = 0.183432
Power (1-β err prob)
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
10
20
30
40
50
60
70
80
90
100
Total sample size
Draw plot

Plot Parameters

Plot (on y axis) Power (1-β err prob) ☑ with markers □ and displaying the values in the plot

as a function of Total sample size from 10 in steps of 5 through to 100

Plot 1 graph(s) interpolating points

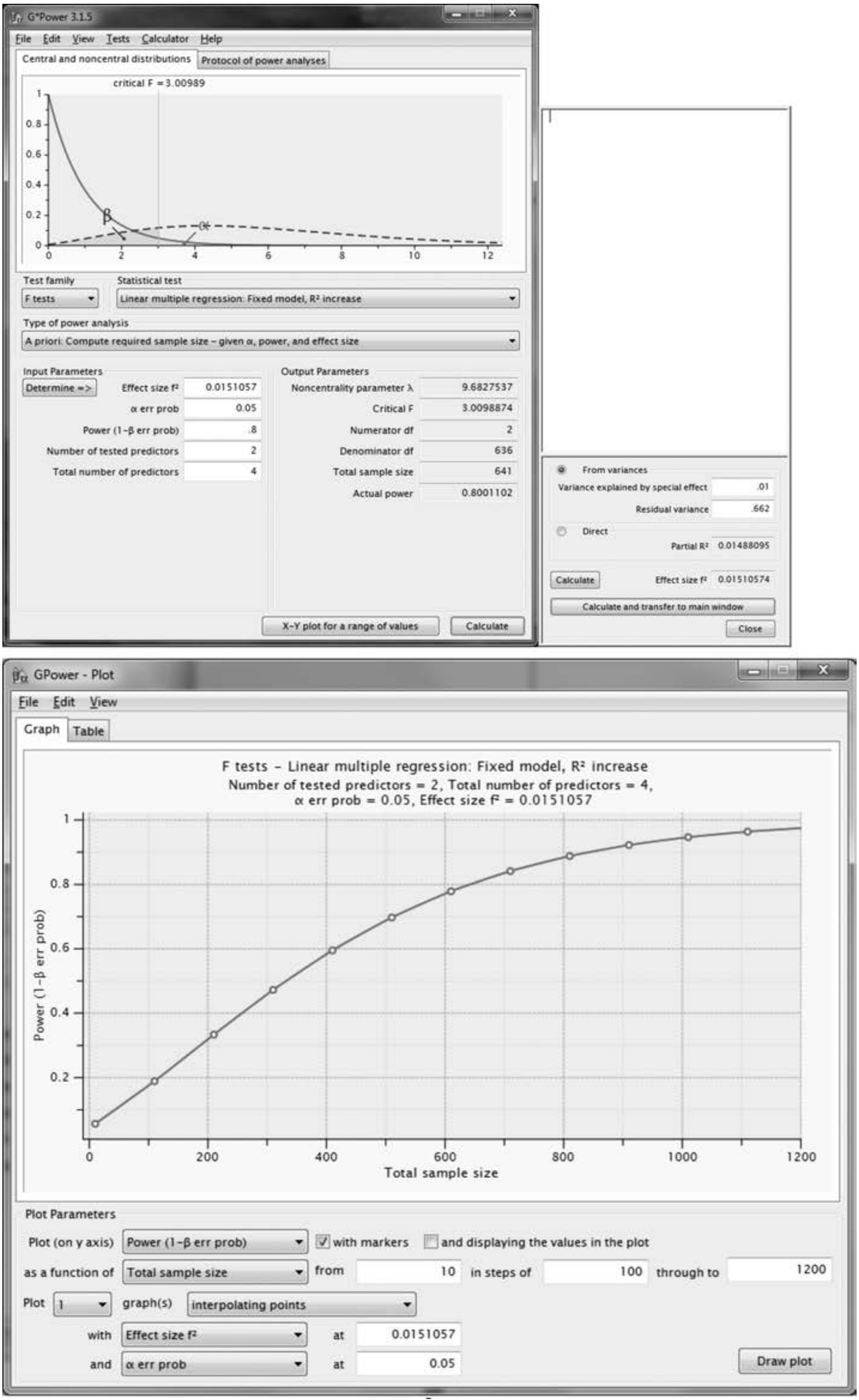
with Effect size f² at 0.183432

and α err prob at 0.05

Figure 10.14 Power to detect a statistically significant R^2 as a function of sample size. This figure refers to the same regression as Figure 10.13, both from Chapter 4.

in the equation. The R^2 with two variables in the equation was .328, and the psychological variables added another .010 to the R^2. What sort of power was associated with this block? Given the sample size of 887, this final block in the regression had a power of .92; given this information, we had a 92% chance to reject correctly a false null hypothesis of no effect for the psychological variables. Given these same numbers, a sample size of 641 (see Figure 10.15) would be needed for a power .80 and sample size of 841 for a power of .90 for this block. The top of Figure 10.15 shows the input values for G*Power; the lower portion shows the sample size graph.

Consider the regressions in which we added interaction terms to the regression. In Chapter 7 we tested the interaction of Previous Achievement and Ethnic origin in their possible effect on Self-Esteem. The categorical and continuous variable accounted for 2% of the variance in Self-Esteem, and the cross product added another .8% (which I will round off to 1%) to the variance explained, with a sample size of approximately 900. In this example, the test of the interaction term had a power of .86 (post hoc) and .80 power would be achieved with a sample size of 764 (a priori). Although the test of the interaction has lower power than the initial variables, with this sample size we still had adequate power to examine the statistical significance of the interaction.



Test family: F tests

Statistical test: Linear multiple regression: Fixed model, R² increase

Type of power analysis: A priori: Compute required sample size – given α, power, and effect size

Input Parameters

Effect size f²: 0.0151057

α err prob: 0.05

Power (1-β err prob): .8

Number of tested predictors: 2

Total number of predictors: 4

Output Parameters

Noncentrality parameter λ: 9.6827537

Critical F: 3.0098874

Numerator df: 2

Denominator df: 636

Total sample size: 641

Actual power: 0.8001102

● From variances

Variance explained by special effect: .01

Residual variance: .662

○ Direct

Partial R²: 0.01488095

Effect size f²: 0.01510574

F tests – Linear multiple regression: Fixed model, R² increase

Number of tested predictors = 2, Total number of predictors = 4,

α err prob = 0.05, Effect size f² = 0.0151057

Plot Parameters

Plot (on y axis): Power (1-β err prob) ☑ with markers □ and displaying the values in the plot

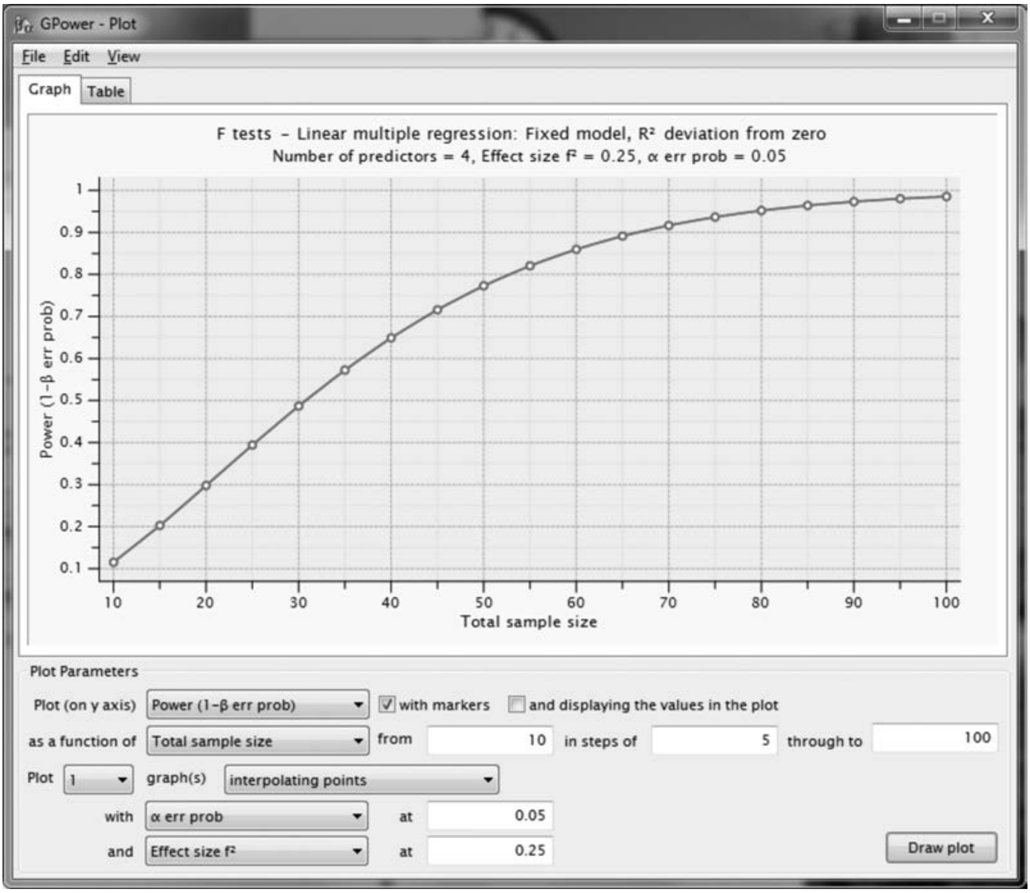
as a function of: Total sample size from 10 in steps of 100 through to 1200

Plot 1 graph(s) interpolating points

with Effect size f² at 0.0151057

and α err prob at 0.05

Figure 10.15 Power analysis for ΔR^2 for one of the sequential regressions from Chapter 5.


GPower - Plot
File Edit View
Graph Table
F tests - Linear multiple regression: Fixed model, R² deviation from zero
Number of predictors = 4, Effect size f² = 0.25, α err prob = 0.05
Power (1-β err prob)
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
10
20
30
40
50
60
70
80
90
100
Total sample size
Draw plot

Plot Parameters

Plot (on y axis) Power (1-β err prob) ☑ with markers □ and displaying the values in the plot

as a function of Total sample size from 10 in steps of 5 through to 100

Plot 1 graph(s) interpolating points

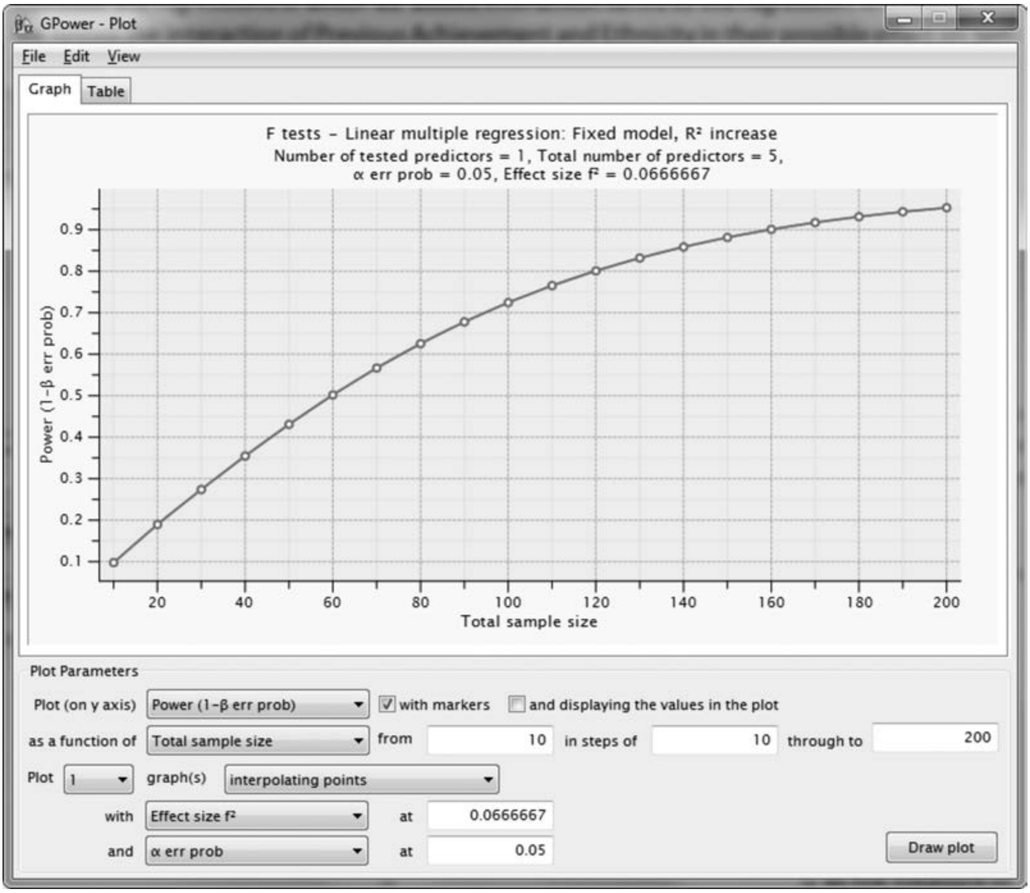
with α err prob at 0.05

and Effect size f² at 0.25

Figure 10.16 Power as a function of sample size for $R^2 = .20$ ($f^2 = .25$). The example illustrates potential problems with a common rule of thumb for sample size in multiple regression.

Finally, consider the 10 to 20 participants per independent variable rule of thumb. Let's model this on some of the other regressions discussed here. Suppose four independent variables account for 20% of the variance in the outcome ($f^2 = .25$), a value that seems reasonable given our examples. Will a sample size of 40 to 80 produce adequate power? Forty cases will produce a total power of only .65, but 80 cases will result in a power of .95. The relevant graph is shown in Figure 10.16. If the R^2 for these four variables was .30 ($f^2 = .43$) instead of .20, then the power associated with 40 cases is .89 (no graph shown). Suppose instead that you were interested in the power associated with one variable that increased the R^2 by .05 above an $R^2 = .20$ ($\Delta f^2 = .067$) from the first four variables in the regression. You will need a sample size of 120 to have a power of .80 for this final variable (see Figure 10.17). It appears that this rule of thumb, although sometimes accurate, will produce low power in many real-world research problems.[1]

In real-world research, you should, of course, conduct these power calculations prior to the research to make sure you collect data on the needed number of participants. You will not know the exact effect size but can generally estimate effect sizes from previous research and your knowledge of relevant theory in the area. Most programs use R^2 or ΔR^2 as the measure of effect size, or the easily calculable f^2 or Δf^2 (as in the previous examples). You can, of course, get estimates of ΔR^2 if researchers have used sequential regression or by squaring


GPower - Plot
File Edit View
Graph Table
F tests - Linear multiple regression: Fixed model, R² increase
Number of tested predictors = 1, Total number of predictors = 5,
α err prob = 0.05, Effect size f² = 0.0666667
Power (1−β err prob)
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
20
40
60
80
100
120
140
160
180
200
Total sample size
Draw plot

Plot Parameters

Plot (on y axis) Power (1−β err prob) ☑ with markers □ and displaying the values in the plot

as a function of Total sample size from 10 in steps of 10 through to 200

Plot 1 graph(s) interpolating points

with Effect size f² at 0.0666667

and α err prob at 0.05

Figure 10.17 Power as a function of sample size for $\Delta R^2 = .05$ (with $1 - R^2 = .75$).

the semipartial correlations (which you can calculate using *t* values, if necessary). If you have no previous research to go on, you can use common rules of thumb (e.g., R^2's of .01, .09, and .25; f^2's of .02, .13, and .35 represent small, medium, and large effects in the social sciences; Cohen et al., 2003). A medium effect size is generally recognized as one noticeable to a knowledgeable observer (Howell, 2013).

As you plan your own research, I encourage you to investigate power more completely and spend some time estimating the sample size you will need in your research (assuming you are not using a large data set like NELS). You don't want to be filled with regrets after having conducted the research and finding nothing of statistical significance and then wishing that you had collected data from 10, or 100, additional participants!

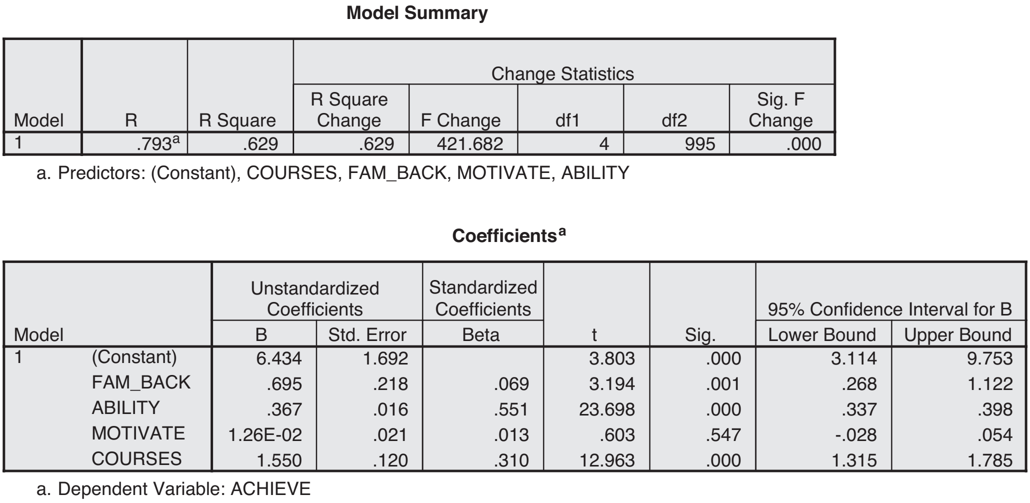
PROBLEMS WITH MR?

Let's revisit some of the interpretive problems we've dealt with throughout this part of the book. I conducted three multiple regressions of high school Achievement on Family Background (SES), Intellectual Ability, Academic Motivation, and Academic Coursework in high school. Our interest is in the effects of these variables on students' high school achievement. We will briefly examine the results of a simultaneous, a sequential, and a stepwise multiple regression, with a focus on the different conclusions we can reach using the different methods. Because our primary interest is in the differences across methods, I won't define the variables in any more detail. The data are taken from Keith and Cool (1992), however, if you are interested in learning more. For this example, rather than simulating the data, I have conducted the regressions using a portion of the correlation matrix as presented in the article. The file "problems w MR 3.sps" illustrates how to conduct a MR using a correlation matrix in SPSS. You may want to save or print this file; it's a useful method and one you can use to reanalyze any published correlation matrix.

Figure 10.18 shows the primary results from a simultaneous MR of Achievement on the four explanatory variables. The regression is statistically significant, and over 60% of the variance in Achievement is explained by these four variables (R^2 = .629). The table of coefficients in the figure provides information about the relative influence of the variables. All the variables appear important, with the exception of Academic Motivation. The effects of Motivation appear very small (β = .013) and are not statistically significant. Motivation, it seems, has no effect on high school Achievement. Turning to the other variables and based on the β's, Ability appears the most important influence, followed by high school Coursework; both effects were large. Family Background, in contrast, had a small but statistically significant effect on Achievement.

Figure 10.19 shows the same data analyzed via sequential MR. For this problem, the explanatory variables were entered in the order of presumed time precedence. Parents' background characteristics generally come prior to their children's characteristics; Ability, a relatively stable characteristic from an early age, comes prior to the other student characteristics; Motivation determines in part the courses students take in high school; and these courses, in turn, determine in part a high school student's Achievement. Thus, achievement was regressed on Family Background, then Ability, then Motivation, and finally Coursework. Relevant results of this regression are shown in Figure 10.19.

There are several differences in these results and those from the simultaneous MR. What is more disturbing is that we will likely come to different conclusions depending on which printout we examine. First, with the sequential regression and focusing on the statistical significance of ΔR^2 for each step, it now appears that Academic Motivation *does* have a statistically significant effect on Achievement (ΔR^2 = .009, $F[1, 996]$ = 19.708, $p < .001$). Second,



Model Summary

			Change Statistics				
Model	R	R Square	R Square Change	F Change	df1	df2	Sig. F Change
1	.793[a]	.629	.629	421.682	4	995	.000

a. Predictors: (Constant), COURSES, FAM_BACK, MOTIVATE, ABILITY

Coefficients[a]

		Unstandardized Coefficients		Standardized Coefficients			95% Confidence Interval for B	
Model		B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound
1	(Constant)	6.434	1.692		3.803	.000	3.114	9.753
	FAM_BACK	.695	.218	.069	3.194	.001	.268	1.122
	ABILITY	.367	.016	.551	23.698	.000	.337	.398
	MOTIVATE	1.26E-02	.021	.013	.603	.547	-.028	.054
	COURSES	1.550	.120	.310	12.963	.000	1.315	1.785

a. Dependent Variable: ACHIEVE

Figure 10.18 Simultaneous regression of Achievement on Family Background, Ability, Motivation, and Academic Coursework.

Model Summary

Model	R	R Square	Change Statistics				
			R Square Change	F Change	df1	df2	Sig. F Change
1	.417[a]	.174	.174	210.070	1	998	.000
2	.747[b]	.558	.384	865.278	1	997	.000
3	.753[c]	.566	.009	19.708	1	996	.000
4	.793[d]	.629	.063	168.039	1	995	.000

a. Predictors: (Constant), FAM_BACK

b. Predictors: (Constant), FAM_BACK, ABILITY

c. Predictors: (Constant), FAM_BACK, ABILITY, MOTIVATE

d. Predictors: (Constant), FAM_BACK, ABILITY, MOTIVATE, COURSES

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	50.000	.288		173.873	.000	49.436	50.564
	FAM_BACK	***4.170***	.288	***.417***	14.494	.000	3.605	4.735
2	(Constant)	4.557	1.559		2.923	.004	1.498	7.617
	FAM_BACK	1.328	.232	.133	5.729	.000	.873	1.782
	ABILITY	***.454***	.015	***.682***	29.416	.000	.424	.485
3	(Constant)	.759	1.766		.430	.667	-2.706	4.224
	FAM_BACK	1.207	.231	.121	5.221	.000	.753	1.661
	ABILITY	.445	.015	.667	28.768	.000	.414	.475
	MOTIVATE	***9.53E-02***	.021	***.095***	4.439	.000	.053	.137
4	(Constant)	6.434	1.692		3.803	.000	3.114	9.753
	FAM_BACK	.695	.218	.069	3.194	.001	.268	1.122
	ABILITY	.367	.016	.551	23.698	.000	.337	.398
	MOTIVATE	1.26E-02	.021	.013	.603	.547	-.028	.054
	COURSES	***1.550***	.120	***.310***	12.963	.000	1.315	1.785

a. Dependent Variable: ACHIEVE

Figure 10.19 Sequential regression results for the same data.

Model Summary

Model	R	R Square	Adjusted R Square	Change Statistics				
				R Square Change	F Change	df1	df2	Sig. F Change
1	.737[a]	.543	.543	.543	1186.615	1	998	.000
2	.791[b]	.625	.624	.082	217.366	1	997	.000
3	.793[c]	.629	.628	.004	10.453	1	996	.001

a. Predictors: (Constant), ABILITY

b. Predictors: (Constant), ABILITY, COURSES

c. Predictors: (Constant), ABILITY, COURSES, FAM_BACK

Figure 10.20 Stepwise regression of Achievement on the same four school learning variables.

although we still conclude that Ability was the most important variable, we now conclude that Family Background was second in importance ($\sqrt{\Delta R^2}$ = .620, .417, .251, .095, for Ability, Family Background, Coursework, and Motivation, respectively; of course this rank order would stay the same if we were to focus on ΔR^2 instead).

Figure 10.20 shows the results from a stepwise regression of these same variables. Again, Academic Motivation appears unimportant, because it never entered the regression equation.

Table 10.2 Regression Coefficients from the Simultaneous versus Sequential Regression of Achievement on Family Background, Ability, Academic Motivation, and Academic Coursework.

Variable	*Simultaneous Regression*	*Sequential Regression*
Family Background	.695 (.218)	4.170 (.288)
	.069	.417
Ability	.367 (.016)	.454 (.015)
	.551	.682
Academic Motivation	.013 (.021)	.095 (.021)
	.013	.095
Academic Coursework	1.550 (.120)	1.550 (.120)
	.310	.310

Note. The first row for each variable shows the unstandardized coefficient followed by the standard error (in parentheses). The second row shows the standardized coefficient.

And again, the order of "importance" changed. In the stepwise regression, Ability entered the equation first, followed by Coursework, followed by Family Background. The stepwise regression thus seems to paint yet another picture of the importance of these variables for Achievement.

How do we resolve these differences? First, we can ignore the results of the stepwise regression, because this is an explanatory problem and stepwise regression is not appropriate for explanatory research. But we still have the differences between the simultaneous and the sequential regressions, both of which are appropriate for explanation.

We have touched on these differences in previous chapters. As noted primarily in Chapter 5, simultaneous regression focuses the *direct* effects of variables on an outcome, whereas sequential regression focuses on *total* effects. Thus, the two approaches may well produce different estimates, even when they are based on the same underlying model and even when one interprets the same statistics. Table 10.2 shows the relevant regression coefficients from Figures 10.18 (simultaneous regression) and 10.19 (sequential regression). For the sequential regressions, the coefficients are from the step at which each variable was entered (shown in italic boldface in the table of coefficients in Figure 10.19). Note the differences in the coefficients; many of the differences are large. Family Background, for example, has an effect of .069 (standardized) in the simultaneous regression versus .417 in the sequential regression.

Again, these differences are not so startling if we know that the simultaneous regression focuses on direct effects versus total effects for sequential regression. But many users of multiple regression seem unaware of this difference. Likewise, many users of MR seem unaware that their regression, when used for explanatory purposes, implies a model and that this model should guide the analysis. The model that underlies these regressions is shown in Figure 10.21, and it can be used to illustrate the differences in coefficients between simultaneous and sequential regression. The simultaneous regression estimates the direct effects, labeled a, b, c, and d in the figure. The sequential regression estimates aspects of the total effects. Thus for the variable motivation, the coefficient for Motivation is the direct effect of Motivation on Achievement (path b) plus the indirect effect of Motivation on Achievement through Academic Coursework (path e times path a).

In Part 2 of this book we will develop such models in considerably more detail and, along the way, gain a deeper understanding of MR and our current difficulties in interpretation. Even if you are using this book for a class in MR only and focusing on Part 1 only, I urge you

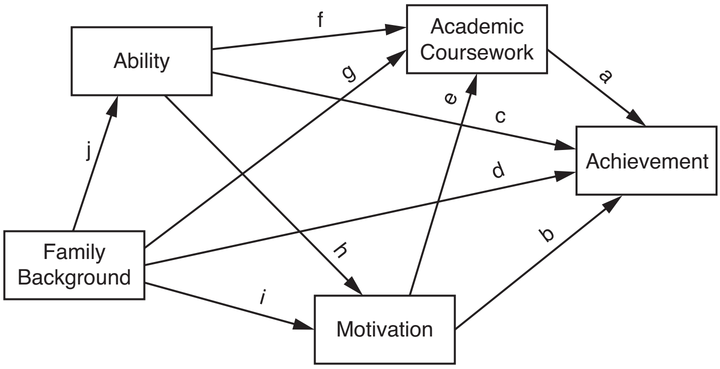


Figure 10.21 Model underlying the simultaneous and sequential regressions of Achievement on Family Background, Ability, Academic Motivation, and Academic Coursework.

to read Part 2 (at least the first two chapters). I think you will find they help you resolve many of the issues that have vexed us—and apparently others—in the use and interpretation of multiple regression. If nothing else, these chapters will give you a more complete heuristic aid in understanding MR results.

EXERCISES

1. Return to the first regression we did with the NELS data. Regress 10th-grade GPA (FFUGrad) on Parent Education (BYParEd) and Time Spent on Homework Out of School (F1S36A2) (see the exercises in Chapter 2). Save the unstandardized residuals and predicted values. Use the residuals to test for linearity in the Homework variable and for the overall regression. Are the residuals normally distributed? Is the variance of the errors consistent across levels of the independent variables (to conduct this final analysis, I suggest you reduce the Predicted Grades variable into a smaller number of categories)?
2. Rerun the regression; save standardized and studentized residuals, leverage, Cook's Distance, and standardized DF Betas. Check any outliers and unusually influential cases. Do these cases look okay on these and other variables? What do you propose to do? Discuss your options and decisions in class. (To do this analysis, you may want to create a new variable equal to the case number [e.g., COMPUTE CASENUM=$CASENUM in SPSS]. You can then sort the cases based on each regression diagnostic to find high values, but still return the data to their original order.)
3. Do the same regression, adding the variable BYSES to the independent variables (BYParEd is a component of BYSES). Compute collinearity diagnostics for this example. Do you note any problems?

Note

[1] Two slightly more sophisticated rules of thumb are N > 50 + 8k for calculating the *N* needed for adequate power in an overall regression and > 104 + k for the testing the statistical significance of a single variable (with k representing the number of independent variables). Green (1991) evaluated these and other rules of thumb and, although they work somewhat better than the simple N > 10k rule mentioned in this chapter, they also fall short, because they do not take effect sizes into account. Indeed, the second rule would underestimate the sample size needed for the final example given here. Green also developed several additional rules of thumb that take effect size into account and are therefore more useful. I recommend you use a power analysis program rather than rules of thumb, but this article is still worth reading.