
13

INTRODUCTION TO MULTIPLE LINEAR REGRESSION

I return to a topic in this chapter that I introduced in several earlier chapters, the connection between health and income. There is a common assumption that the greater the income, the better the health in the United States. Although this assumption is widely held, it is not strictly accurate among all indicators. A well-known example of this discrepancy is that the United States is among the wealthiest nations, but it is not the top in terms of life expectancy, among other indexes of health.

If wealth does not produce good health, what does? This is a question explored by many social science and health researchers. One such researcher, Marmot (2004) suggests that social standing might be a variable primarily responsible for health status, beyond income. Since social status consists of a number of elements (e.g., extent of education, nature of work, achievements, etc.), income or wealth may not be the best predictor of health. In the example that follows, I explore the impact of work and status on health, along with income.

THE ELEMENTS OF MULTIPLE LINEAR REGRESSION

At the end of Chapter 12, I mentioned that multiple linear regression (MLR) is an extension of bivariate linear regression in which more than one predictor variable is added to an equation that predicts values of an outcome variable. MLR is a complex procedure and well beyond the scope of this book. However, I believe it is important, and necessary, for all researchers to understand at least the basics of the procedure.

Using Statistics in the Social and Health Sciences with SPSS® and Excel®, First Edition.
Martin Lee Abbott.
© 2017 John Wiley & Sons, Inc. Published 2017 by John Wiley & Sons, Inc.

I will identify some of the primary elements of MLR and provide a simple example to highlight how to analyze and interpret MLR findings. I will not discuss the hand calculations but will instead examine the SPSS® and Excel output for the research example.

You will need to explore more advanced treatments of MLR in order to gain a depth of understanding of this versatile process. I encourage you to explore my more comprehensive treatment of MLR in Abbott (2010) as a starting point.

SAME PROCESS AS BIVARIATE REGRESSION

MLR like bivariate regression is a process that creates a model for *predicting* values of an outcome variable from predictor variables and a way of *explaining* the variance in an outcome variable. Although the hand calculations are more complex, the overall procedure is the same. *I discuss the linear relationship among variables (predictors and outcome) that will improve on the prediction of an outcome variable knowing the correlation between the outcome and predictors.*

Statistical Significance

Like bivariate regression, MLR can be examined for statistical significance by examining the *omnibus* test and the *individual predictor* test(s). Both procedures make use of the analysis of variance (ANOVA) table since both examine how the various components of variance relate to one another. Bivariate regression has only one individual predictor test, while MLR has several but the statistical significance of each individual predictor can be tested with an individual *T* test.

Effect Size

Effect size is measured the same way in MLR as in bivariate with the omnibus test of the model. In both cases, R^2 is an indicator of how much variance is explained in the outcome variable as a result of the predictor(s). Effect size in MLR measures the contribution of a set of predictors upon the explained variance of the outcome variable. There are more specific effect size indicators for individual predictors in MLR, but they are based on the same principle.

Coefficients

Like bivariate regression, there is a regression equation with coefficients that explain the relationship among the study variables. In MLR, of course, there are additional coefficients because there are additional predictors. The beta coefficient(s) express the same thing: the changes in Y_{pred} when the beta for a specific predictor changes one unit. Since MLR has multiple beta coefficients, this interpretation changes to “the changes in Y_{pred} when the beta changes one unit and the effects of the other predictors are controlled.” This last restriction allows the researcher to understand how a particular predictor affects the outcome *at similar levels of all the other predictors*. This is the way in which the other predictors are understood as being controlled.

Scatter Diagrams

I have shown how scattergrams visually express the relationship between a predictor and an outcome. MLR can use scattergrams for models that include a second predictor by creating a 3D graph that includes a “Z” axis as well as X and Y axes. More complex MLR designs use graphical methods that capture the additional influences of other variables, but these are beyond the scope of this book.

Fixed- versus Random-Effects Modeling

Many statisticians make a distinction between “fixed effects” and “random effects” modeling, which can have an impact on the findings of a study and in particular on the generalizability of the findings. Fixed-effects modeling uses the same categories of variables across studies. The predictor variable (x) has a set number of categories, like gender or experimental method, where the same categories are used in all situations. In this situation, generalizability of findings is limited to the fixed category values.

On the other hand, random-effects modeling is most common in non-experimental social science research analyses; it is used in situations where the evaluator takes a random set of observations from a larger population of values, with the categories or levels of the predictor variable (x) “emerging” from and determined by the data sampled. The categories and values of x are thus random.

According to many authorities, when the main assumptions for regression are met, fixed-effects and random-effects models are interchangeable. This is particularly the case when errors are not correlated to predicted values.

SOME DIFFERENCES BETWEEN BIVARIATE LINEAR REGRESSION AND MULTIPLE LINEAR REGRESSION

There are some differences between bivariate and multiple regression due to the added complexity of additional predictors. These differences are not related to the structure of the process but are due to the added complexity.

Multiple Coefficients

I mentioned this difference earlier. Added predictors do not change the essential nature of regression but rather call for additional methods for interpretation of output.

Explanation of R^2

The explanation of components of explained variance in the outcome is related to the complexity of multiple coefficients and multicollinearity. MLR can be used to pinpoint the contribution to the R^2 of individual predictors. This is one of the chief contributions of MLR to research.

Although I cannot pursue this matter very far in this book, I will mention that interpretation of the components of R^2 rests with some procedural elements (“order

of entry schemes”) as well as with the very important correlation procedures of *partial* correlation and *semipartial* (or “part”) correlation. I will demonstrate how to use semipartial correlation for interpreting MLR results.

Entry Schemes

MLR results can be affected by the order in which predictors are “added” to the equation by statistical software. Some software procedures allow the user to add (and remove) the predictors to the overall model in “stepwise” fashion using a set of predetermined statistical guidelines that judge whether the predictors meet certain numerical thresholds. Other schemes allow the researcher to add predictors to an MLR model in the order they determine based on a priori theoretical considerations.

The researcher can use several approaches to identify the influence of predictors on an outcome variable. (I will show a couple of these in the example in the following.) “Hierarchical” regression is simply an MLR study in which the researcher adds predictors according to an indication of importance to the study and observes the changes to the overall R^2 as a result of adding the predictors.

STUFF NOT COVERED

Because of the complexity of MLR, I cannot hope to discuss all the dimensions in this book. I mentioned some of these areas earlier (e.g., partial and semipartial correlation). Here are some other areas of MLR I cannot cover here. You might consult another of my works (Abbott, 2010) for further consideration of MLR processes.

Using MLR with Categorical Data

MLR is quite versatile and allows the researcher to use categorical as well as continuous predictors. It can even be used in experimental studies; there is a statistical connection between MLR and ANOVA, the latter of which is typically used for experimental study.

MLR uses categorical predictors through procedures in which the researcher “transforms” each of the categories of a predictor into separate “vectors” (or “subvariables”) which together comprise the predictor. Each of these vectors can then be understood in their relationship to the outcome variable. (An example might be creating four vectors from a treatment variable “noise” in predicting human learning in the fictitious study I introduced in an earlier chapter.) Transforming predictors to “sets” of vectors can be done using “dummy,” “effect,” or “contrast” coding according to the needs of the researcher.

MLR also uses categorical outcome variables (e.g., where values are “0” or “1”) in a process called “logistic regression.” In this procedure, predictor variables can be either or both continuous or/and categorical. The overall attempt is to identify the likelihood of an outcome taking place or not. An example might be to determine the likelihood of contracting a certain disease given inputs of certain enzyme levels, age, diet, etc.

ASSUMPTIONS OF MULTIPLE LINEAR REGRESSION

421

Curvilinear Regression

Like bivariate regression, the relationship of the predictors to the outcome variable in MLR may be nonlinear as well as linear. Statistical software like SPSS have ways of identifying these trends, so the researcher can express the most efficient model for predicting the outcome variable.

Multilevel Analysis

Multilevel analyses are procedures that recognize different “levels of data” in a regression study. For example, I might want to understand the relationship between patient survival rates of a certain illness and treatment protocols for treating the illnesses. Typically, researchers simply want to see if one variable (i.e., protocols) can predict outcomes (i.e., survival rates). This approach “understands” the relationship between two variables that are measured at the same (individual) level of analysis. However, might the hospitals within which the treatments are applied also affect the outcomes? Viewed thus, individual protocols used are “nested” within hospitals which aggregate the incidents in which the protocols were used. This is a different level from the individual level of analysis.

Viewing hospitals as a different variable in such a study suggests several questions: Are all hospitals equal, with respect to this research question? Would the unequal funding levels for newer technology, newer understandings of the disease process, more and easier analytical processes available, and other such considerations likely affect the relationship between given individual treatments and survival rates for this particular disease?

Multilevel analysis is a way of understanding what the relationships are at one level of analysis (treatment) by recognizing the influence of another level of analysis (hospital). These are very powerful tools for a researcher, but they are not easy to learn. You might explore hierarchical linear modeling approaches if you are interested in this topic.

ASSUMPTIONS OF MULTIPLE LINEAR REGRESSION

Generally speaking, MLR shares the assumptions of Bivariate regression, but I added several later due to the complexity of the data where there is more than one predictor.

Normal Distribution

The values of the variables are normally distributed by themselves and are normally distributed at each combination of levels of the other variables. This is multivariate normal distribution, and is very difficult to achieve in practice. You can use a scattergram to view the normal distribution of each variable (especially the outcome variable), but you will rely on the tests for the other assumptions that I describe later. Multivariate normality results in a linear relationship.

Homoscedasticity

As I discussed with correlation, homoscedasticity refers to equal variance among the variables in an MLR analysis. Since more than one predictor defines MLR, you cannot rely on scattergrams to test this assumption. You still need to examine visual patterns of relationships, but in MLR you rely on the analysis of residuals since these are an index of the set of predictor variables in relationship to the outcome variable. I discuss these procedures in the following.

Linear

Linearity is an important assumption to meet for most research uses of MLR. I discussed the procedure for checking linearity with bivariate regression in Chapter 12; MLR assumptions can be checked with a similar process in addition to using residual analysis discussed later.

Multicollinearity

Multicollinearity is a facet of MLR that relates to the *relationship among the predictors* as well as the relationship between predictors and outcome. MLR takes into account that predictors probably have some relationship *to one another* as well as to the outcome variable. For example, if I explain health as a combination of inputs like education, income, and work, each of these predictors probably is at least somewhat related to the others. This overlap among predictors makes it difficult to determine the extent of the relationship of each predictor with the outcome variable when the other predictors are controlled. If the overlap is too great, the set of relationships is too murky to make useful conclusions.

The MLR output can be examined to understand the size and nature of the components of variance produced in a model with several predictors. Statisticians have devised guidelines for detecting problems in which “too much” of the predictor-to-predictor variance clouds the understanding of explained variance in the outcome variable and makes it difficult to understand the contribution of individual predictors. I will show in the following example how SPSS reports information that is helpful for understanding the interrelationship among the predictors.

ANALYZING RESIDUALS TO CHECK MLR ASSUMPTIONS

Residual analyses can help to determine whether assumptions are met for normality, equal variance, and linearity in an MLR analysis. Recall that residuals are errors in predicting true values of an outcome variable in a regression procedure. In an analysis in which the variables are multivariate normal, have equal variance, and are related in linear fashion, the residuals will display certain patterns that can be assessed visually and by other tests.

Standardized Residuals Normally Distributed

When the assumptions are met, the residuals in an MLR analysis should be normally distributed. This would indicate no “biases” among the data, especially in relationship to the outcome variable. SPSS provides a residual analysis program that includes a histogram of the residuals to determine whether the residuals fit a normal distribution. The “*P-P* Plot” is also produced that allows the researcher to determine the extent to which residuals adhere to a straight line representing expected cumulative probabilities of the normal distribution.

Random Relationship Between Residuals and Predicted Values

When residuals are plotted with predicted values in a scattergram, the dispersion of the residual values can reveal nonlinear relationships and heteroscedasticity. When assumptions are met for normality, linearity and equal variance, there should be no pattern in the residuals when plotted with predicted values.

- Residuals should show a zero slope and even distribution of values when plotted with predicted values. This would indicate the assumptions are met and there is no additional predictability of the outcome by the errors, or randomness.
- When the pattern of residuals around the zero slope line is curved, there is likely a nonlinear relationship among the variables. This might appear like the bottom left panel of Figure 11.3 which showed a curvilinear pattern between two variables. When the residuals appear thus, the assumption for linearity is violated.
- When the pattern of residuals gets increasingly wider along the zero slope line (i.e., outward like a shotgun pattern), homogeneity of variance may be compromised. This pattern would look similar to the scattergram shown in the bottom panel of Figure 11.14, although this figure explained the general condition of heteroscedasticity.

DIAGNOSTICS FOR MLR: CLEANING AND CHECKING DATA

For an extended discussion of regression diagnostics, see Abbott (2010). In this section, I present some of the most important considerations for understanding the nature of the data you use in an MLR analysis. Diagnostic procedures such as these are used to detect problematic patterns within an MLR analysis that may have great influence on the outcome. I discuss the primary diagnostic categories here, and then I discuss the diagnostics of the MLR example later.

In the following, I will use a small and brief example to demonstrate the steps for the diagnostic procedures. Table 12.2 shows the data for the fictitious study relating Healthy Days to Income Class. Table 13.1 shows the same data except I changed one value, case #7, from (0, 45) to (0, 80), to help illustrate the diagnostic procedures.

TABLE 13.1 The Diagnostic Study Values

Income Class (X)	Healthy Days (Y)
5	76
1	58
6	81
8	86
9	90
3	62
0	80

IncomeClass	HealthyDays	ZHealthyDays	ZIncomeClass	ZRE_1	COO_1	LEV_1	DFB0_1	DFB1_1
5	76	-0.012	0.126	-0.128	0.002	0.003	-0.161	-0.009
1	58	-1.518	-1.048	-1.003	0.361	0.183	-5.235	0.711
6	81	0.406	0.419	0.144	0.003	0.029	0.080	0.033
8	86	0.825	1.006	0.153	0.008	0.169	-0.170	0.102
9	90	1.160	1.299	0.318	0.065	0.281	-0.760	0.327
3	62	-1.183	-0.461	-1.102	0.160	0.035	-3.076	0.282
0	80	0.323	-1.341	1.618	1.865	0.300	11.990	-1.776

Figure 13.1 The SPSS[®] report of diagnostic values for the study data.

Figure 13.1 shows the results of several diagnostic procedures that I will use as a reference in the following sections.

These diagnostic measures result from choosing the “Regression – Linear” menus in SPSS as shown in Figure 12.10. When you use this menu, the specification menu shown in Figure 13.2 results.

You can specify the regression analysis using the submenu selections (top right buttons) as I demonstrated in Chapter 12 (see Figures 12.11 and following). I did not discuss diagnostics in those examples, so I introduce them here. Choosing the “Save” button in the Regression menu in Figure 13.2 results in the further specification window shown in Figure 13.3.

As you can see in this menu, I checked some of the boxes in the “Distances,” “Residuals,” and “Influence Statistics” sections to obtain the diagnostic values I reported in Figure 13.1. I will discuss these in the following sections as I encounter each of the values.

EXTREME SCORES

I have alluded several times thus far that “outlier” scores, or those scores that fall outside the group of other scores in a correlation or regression analysis, can dramatically affect results. Extreme scores can be a matter of examining the study variables

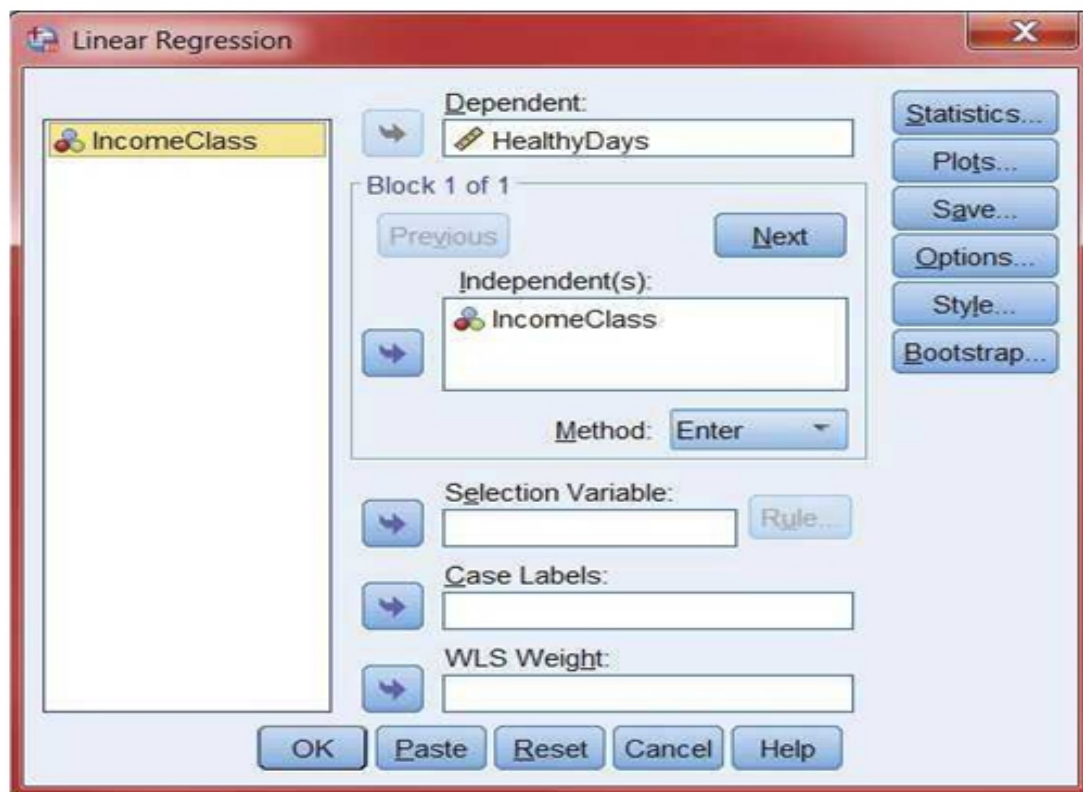


Figure 13.2 The SPSS® specification menu for linear regression.

individually (“univariate extreme scores”) or taken together in a regression study (“multivariate extreme scores”). Therefore, the first diagnostic tool for understanding potential problems with study data is to find any such scores.

Univariate Extreme Scores

One of the ways to detect extreme scores of the study variables is simply by running a descriptives program such as I have described for analyzing skewness and kurtosis. These measures may identify values of a variable that might inflate one or the other of these measures.

Another way to detect possible extreme scores of individual variables is to examine standard scores (z scores) for each of the values. SPSS creates these by you simply checking the box that requests saving standard scores when you use the “Descriptives” command (“Save standardized values as variables”). You can examine the standard scores (at both positive and negative ends) and look at scores of ± 3.00 as extreme scores.



Figure 13.3 The SPSS® specification menu for diagnostic values.

Standard scores for both variables are listed in columns 3 and 4 of Figure 13.1 as “ZIncomeClass” and “ZHealthyDays.” As you can see, there are no z scores listed above 2.0, so it appears that none of the values are extreme within their own value distributions.

A scattergram of the bivariate relationship might provide additional visual clues of possible extreme values. Figure 13.4 shows the scattergram of these variables. I highlighted the value I changed for this demonstration (0, 80). Compare this graph to the scattergram in Figure 12.4 that shows the unchanged case (0, 45). The scattergram for the unchanged data shows the study values deviating closely to the regression line. The scattergram in Figure 13.4 shows the changed values separated from the other values and the regression line by quite a distance. Remember, this is a change in value of only one study variable (Healthy Days).

The impact of this change is quite large on the study findings. The effect size changes from 0.96 in the original analysis to 0.492 in the study that includes the one

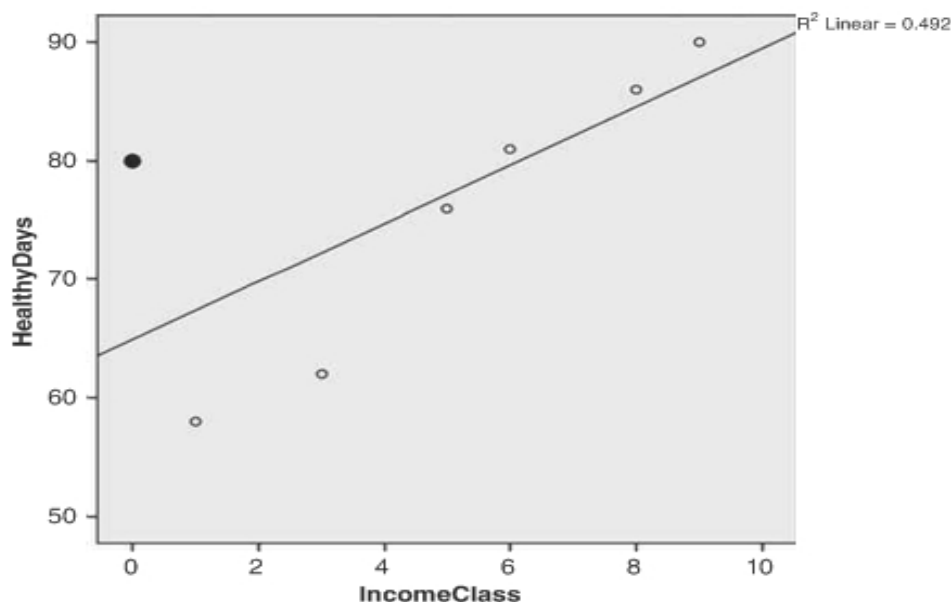


Figure 13.4 The scattergram of the study variables.

changed value. This is a reduction in the effect size of 0.47 ($0.96 - 0.49$), a very large impact from one extreme value.

When you encounter extreme values using this procedure, you need to think about how to manage the data and how to account for the extreme values: How many extreme scores are there? Are these “real” scores (e.g., are the extreme values accurate representations of respondent attitudes, or are they clerical errors)? How would it impact the analysis to delete or change them? These are considerations researchers must face as they model the data from the study.

Multivariate Extreme Scores

Often, extreme scores are identified by examining the analyses of the variables taken together rather than individually. The procedures I demonstrate are available for regression problems with more than one independent variable, but I am using an example from a bivariate regression procedure to illustrate how they work.

As you recall, residual scores (the difference between actual and predicted scores) provide a way of understanding how scores disperse around the regression line. Residual scores can be an excellent way of looking for extreme scores since large residual values indicate large distances between predictions from a regression equation and the actual value used to create it. Obviously, this has a great deal to do with the size of the correlation between the variables, but residuals can indicate extreme values and whether these might be problematic to an analysis.

You can view the residuals as unstandardized values that retain the scale of the study variables, but it is important to remember that each may have different

variances. Thus, SPSS provides a list of standardized residuals where each residual (the difference between the actual and predicted y values) has been divided by the standard deviation of the residuals (which is also the standard error of estimate).

Column 5 in Figure 13.1 lists the standardized residuals in the regression analysis as “ZRE_1.” As you can see, none of the values are greater than $(\pm) 3.00$ which I discussed previously as a reference for z scores. The standardized residual for the changed case is elevated a bit (1.61813) above the others but cannot be considered extreme by this measure.

DISTANCE STATISTICS

Figure 13.3 indicates that I chose two measures to assess extreme scores. These calculate how each score lies in reference to all the other scores in a regression analysis. Figure 13.1 shows the results of these two measures in columns 6 and 7 as “COO_1” and “LEV_1,” respectively.

- “COO_1” are Cook’s distance values (CDV).
- “LEV_1” are centered leverage values (CLV).

Centered Leverage Values (CVL)

These values indicate distances for the independent variable. The greater the distance between each x value and the mean of x , the larger the CLV, whereas values closer to the mean of x are increasingly smaller. The maximum leverage score is 1.0. Statisticians differ in their opinions of cutoff criteria based on these values, and some suggest that cutoff values may be misleading and somewhat artificial.¹ It may be helpful first to look for extreme values by sorting the CLV and then examining the scores for the highest values.

In this example, the highest values are 0.28 and 0.299 for cases 5 and 7, respectively. The mean CLV (reported in the SPSS output) for these data is 0.143 (SD = 0.123). Taking a broad view, these values are near but do not approach high values (e.g., mean + 2SD) of the reported residual statistics for CLV scores (i.e., $0.143 + 0.246 = 0.389$). Remember, the CLV measures extreme independent variable values, so you can examine the IV values in the data and find that these two cases represent the lowest and highest IV values but do not appear from the scattergram to be inordinately removed from the other values.

Cook’s Distance Values (CDV)

These values measure the effect on all the residuals if a particular case were dropped from the analysis. It therefore is a measure of both independent and dependent

¹Pedhazur (1997) discusses high leverage values noting $(k + 1)/N$ as average leverage: multiplying $(k + 1)$ by 2 or 3 might generally identify a high value. Thus, in this case, $3(2)/15 = 0.40$, compared to the 0.389 calculated above.

variables acting in concert in the analysis, with large values indicating more impact on the regression coefficients. Here again, cutoff criteria are debated by various researchers. One way to examine CDV visually is to sort the database by the values and to identify how extreme they become. They typically range from small values of “0” to larger values where the extreme scores may be several times larger than the immediately preceding scores.

Using this method, you can identify one CDV score that appears to be quite distant from the others in the regression analysis. Case 7 (the case I changed) shows a CDV of 1.865. With the database sorted according to these values, the next nearest high CDV is 0.36073 (case 2) and then 0.16020 (case 6). Given these range of values, Case 7 appears to be extreme.

The scattergram provides a helpful visual clue to whether this value might be extreme and therefore problematic to the analysis. As you can see in Figure 13.4, the highlighted case (0, 80) is clearly “outside the pack” of the other values. Since the CDV measures the joint influence of independent and dependent variables in the analysis, this case may be one to flag for further investigation. Recall that this case produced a larger but no inordinately large residual value.

INFLUENCE STATISTICS

As I noted, it is important to examine the data for extreme scores since they may have a large impact on the regression equation. Examining the residuals and distance statistics is helpful in locating potential extreme scores, but SPSS provides other measures that identify the impact on regression coefficients if individual scores are deleted from the analysis.

The last two columns (8 and 9) of Figure 13.1 show values for two measures that represent the change in the regression coefficients when a particular score is deleted from the analysis. These are “DfBeta(s)” that indicate changes on the intercept value (DFB0_1) and slope value (DFB1_1) of a regression equation when a specific case value is dropped from the database.

Thus, for example, if I dropped the changed case 7, what would be the impact on the regression equation? Essentially, these measures represent the amount of change in the intercept and slope of the regression line if you were to run the regression, delete the scores, and then rerun the regression line. If you subtract the values from the original regression, you will observe the change is equal to the DfBeta measures.

To take the example from my discussion earlier, I can show how the original regression equation is changed by dropping case 7. The overall (unstandardized) regression equation based on the regression analysis (including case 7) is

$$\text{Healthy Days} = 2.459 \text{ Income Class} + 64.902$$

Adjusted intercept value: Subtracting the DFB0 value (11.98987) score from the intercept in the original equation (64.902) yields the value of the intercept that would result if the extreme score (case 7) were deleted:

$$(64.902) - (11.990) = \underline{\underline{52.912}}$$

Adjusted slope value: Subtracting the DFB1 score (-1.776) from the slope value in the original equation (2.459) yields a value of the slope value that would result if the extreme score (case 7) were deleted:

$$(2.459) - (-1.77628) = \underline{4.235}$$

The resulting (unstandardized) regression equation with case 7 deleted is

$$\text{Healthy Days} = 4.235\text{Income Class} + 52.912$$

The impact of deleting an extreme score (in this example case #7) is registered in the regression equation as shown before. But you can also measure these changes by the change in effect size:

$$\text{Initial database (including extreme case \#7)} : R^2 = 0.492$$

$$\text{Revised database (deleting extreme case \#7)} : R^2 = 0.969$$

The overall impact of deleting the extreme case is therefore increasing the effect size by 0.477 ($0.969 - 0.492$).

MLR EXTENDED EXAMPLE DATA

In the introduction to this chapter, I noted the questions that have arisen among social science researchers about the connections between income and health. As I noted there, a common assumption states that the greater the income, the better the health in the United States, although this is not so simple a relationship as one might expect.

If wealth does not always produce good health, what other factors may help to elaborate an understanding of this relationship? In the example that follows, I explore the impact of social status and work factors that might help to shed light on the relationship between health and income.

Table 11.2 shows the data I used to explain correlation procedures. Although hypothetical, these are the kinds of data that researchers might use to explore the “income–health” relationship. In this chapter, I extend those considerations by adding two predictors of health in addition to income. In this section, I use an MLR procedure to show some of the ways to interpret the findings from SPSS and Excel. Both perform MLR analyses, but the Excel capacity is more limited than SPSS. In addition, because SPSS is specifically designed for these kinds of complex analytical procedures, there are more custom features provided in the menus.

In the study database, I use variables modeled after those in GSS as I did in the correlation study. I took a small random sample from the overall (GSS, 2014) database to demonstrate the procedures a researcher might use in a larger, more complete, MLR analysis. My primary concern in this example is to demonstrate the MLR procedures used in any study, not to produce a primary research finding. Therefore, as in the

ASSUMPTIONS MET?

431

correlation study, the conclusions I reach in this demonstration cannot be taken as representative of larger, more comprehensive studies.

Ordinarily, larger samples produce more representative findings. I use a small sample in this demonstration ($N = 78$) that is lower than most researchers would use in actual studies.² The data do illustrate the MLR procedures, however, and the findings help to illustrate how the study variables might be understood in terms of the income–health question.

The following are the variables I use. These are survey-based questions that record respondent opinions on each of the following:

- HeathOP – Opinion of health (1 = poor, 2 = fair, 3 = good, 4 = very good, 5 = excellent)
- IncomeOP – Opinion of family income (1 = far below average, 2 = below average, 3 = average, 4 = above average, 5 = far above average)
- RankOP – Opinion of social position (1 = bottom rank through 10 = top rank)
- StressOP – How often the respondent finds work stressful (1 = always, 2 = often, 3 = sometimes, 4 = hardly ever, 5 = never)

Before I proceed to the MLR procedure, I address whether the assumptions are met for the study. Analyzing the residuals requires that I run the procedure, so I will look at this section independently of the primary MLR findings.

ASSUMPTIONS MET?

Normal Distribution

As I noted earlier, the study variables can be examined individually as well as through residual analyses that incorporate all the study variables taken together. Visually, according to Figure 13.5, the outcome variable (HealthOP) appears to be slightly skewed negatively (indicating that slightly more of the respondent sample rated their health positively).

Using the Compare Means procedure, I gain additional numerical indications of normality for the other study variables. Figure 13.6 shows the descriptives report for all the study variables showing skewness and kurtosis figures that are in bounds of a normal distribution.

The optimal way to assess normality for the MLR with multiple predictors is to examine the residuals. This analysis will help to assess several assumptions, so I provide these findings in the following.

²There are numerous suggestions among researchers about the number of cases per number of predictors used in MLR studies. For smaller studies, less than 100 cases with three predictors would typically not be viewed as acceptable.

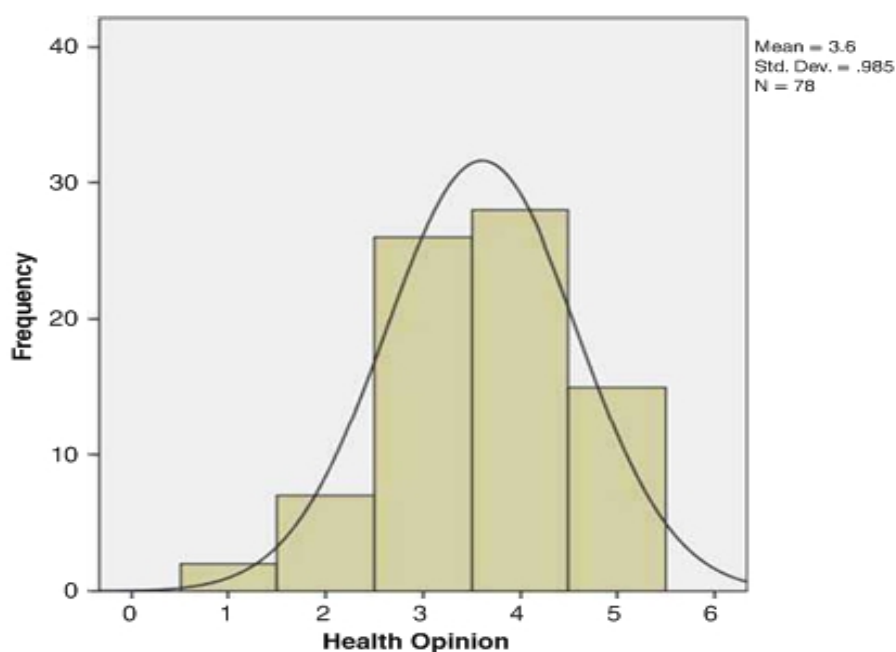


Figure 13.5 The histogram for the dependent variable in the study.

Report				
	OPINION OF FAMILY INCOME	HOW OFTEN DOES R FIND WORK STRESSFUL	Health Opinion	Rank Opinion
Mean	2.91	3.00	3.60	6.04
N	78	78	78	78
Std. Deviation	.871	1.162	.985	1.848
Kurtosis	.223	-.565	-.146	.754
Std. Error of Kurtosis	.538	.538	.538	.538
Skewness	-.186	-.051	-.378	-.058
Std. Error of Skewness	.272	.272	.272	.272

Figure 13.6 The SPSS® descriptives report showing skewness and kurtosis findings.

Homoscedasticity

As I discussed with correlation, homoscedasticity refers to equal variance among the variables in an MLR analysis. Since more than one predictor defines MLR, you cannot rely on scattergrams to test this assumption. You still need to examine visual patterns of relationships, but in MLR you rely on the analysis of residuals since these are an index of the set of predictor variables in relationship to the outcome variable. I discuss these procedures and findings later.

Linear

Linearity is an important assumption to meet for most research uses of MLR. I discussed the procedure for checking linearity with bivariate regression in Chapter 12. MLR assumptions can be checked with a similar process, but a straightforward procedure is to examine the residuals which I discuss later.

Multicollinearity

Multicollinearity is a facet of MLR that relates to the *relationship among the predictors* as well as the relationship between predictors and outcome. MLR takes into account that predictors probably have some *relationship to one another* as well as to the outcome variable. For example, if I explain health as a combination of inputs like education, income, and work, each of these predictors probably is at least somewhat related to the others. This overlap among predictors makes it difficult to determine the extent of the relationship of each predictor with the outcome variable when the other predictors are controlled. If the overlap is too great, the set of relationships is too murky to make useful conclusions.

The MLR output can be examined to understand the size and nature of the components of variance produced in a model with several predictors. Statisticians have devised guidelines for detecting problems in which “too much” of the predictor-to-predictor variance clouds the understanding of explained variance in the outcome variable and makes it difficult to understand the contribution of individual predictors. I will show in the example later how SPSS reports information that is helpful for understanding the interrelationship among the predictors.

ANALYZING RESIDUALS: ARE ASSUMPTIONS MET?

As I noted previously, residual analyses can help to determine whether assumptions are met for normality, equal variance, and linearity in an MLR analysis. Recall that residuals are errors in predicting true values of an outcome variable in a regression procedure. In an analysis in which the variables are multivariate normal, have equal variance, and are related in linear fashion, the residuals will display certain patterns that can be assessed visually and by other tests.

Standardized Residuals Normally Distributed

When the assumptions are met, the residuals in an MLR analysis should be normally distributed. This would indicate no “biases” among the data, especially in relationship to the outcome variable. SPSS provides a residual analysis program that includes a histogram of the residuals to determine whether the residuals fit a normal distribution.

I will demonstrate how to produce the residual findings when I explain the SPSS procedure in the following. For now, examine Figure 13.7 showing a histogram of the (standardized) residual values from the study. As you can see, the residuals appear to be normally distributed, which is an indication the primary assumptions for the data are met.

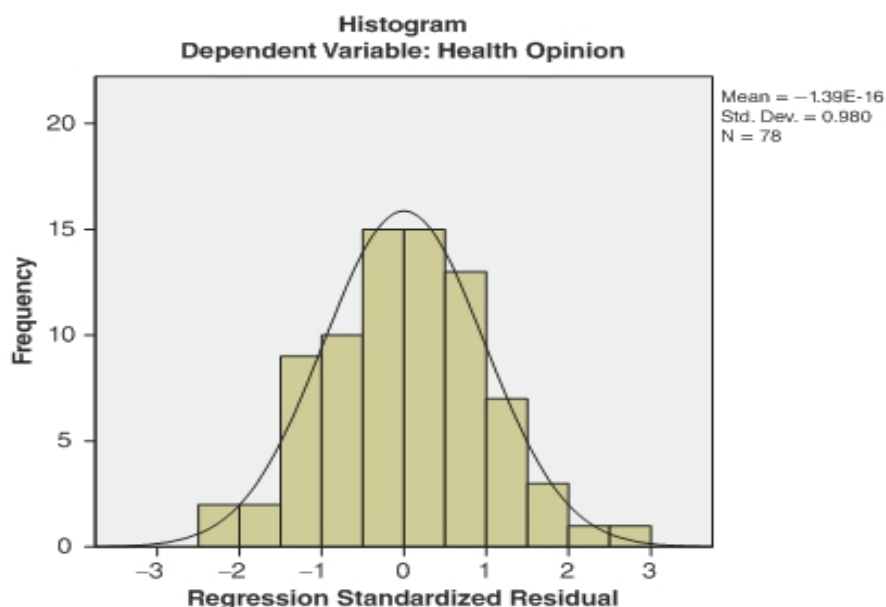


Figure 13.7 The histogram of standardized residuals from the study.

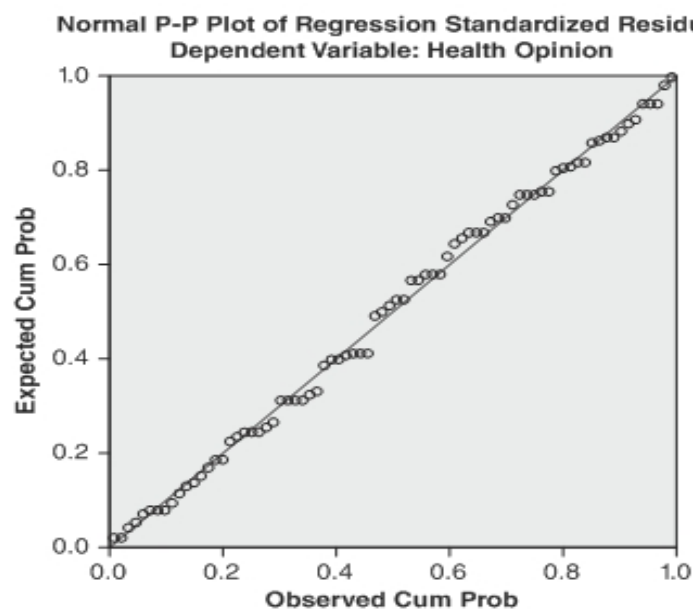


Figure 13.8 The *P-P* plot of standardized residuals from the study.

Figure 13.8 demonstrates a “*P-P* Plot of Regression Standardized Residual” values that also indicate assumptions are met. As you can see, the residual values conform to expected probabilities of a normal distribution by closely “hugging” the straight line.

Random Relationship between Residuals and Predicted Values

When residuals are plotted with predicted values in a scattergram, the dispersion of the residual values can reveal nonlinear relationships and heteroscedasticity. When assumptions are met for normality, linearity, and equal variance, there should be no pattern in the residuals when plotted with predicted values. Examine Figure 13.9 as a reference to the following elements of the assumptions:

- Residuals should show a zero slope and even distribution of values when plotted with predicted values. This would indicate the assumptions are met and there is no additional predictability of the outcome by the errors, or randomness. As you can see in Figure 13.9, the slope is “0” and the cases are evenly distributed around the regression line. There are a few extreme values that you might examine, but overall, the figure shows the pattern of residuals indicating assumptions are met.
- When the pattern of residuals around the zero slope line is curved, there is likely a nonlinear relationship among the variables. This might appear like the bottom left panel of Figure 11.3 which showed a curvilinear pattern between two variables. When the residuals appear thus, the assumption for linearity is violated. The curved pattern of residuals is not present in the study report as indicated in Figure 13.9.
- When the pattern of residuals gets increasingly wider along the zero slope line (i.e., outward like a shotgun pattern), homogeneity of variance may be compromised. This pattern would look similar to the scattergram shown in the bottom panel of Figure 11.14, although this figure explained the general condition of heteroscedasticity. No such pattern is shown in Figure 13.9.

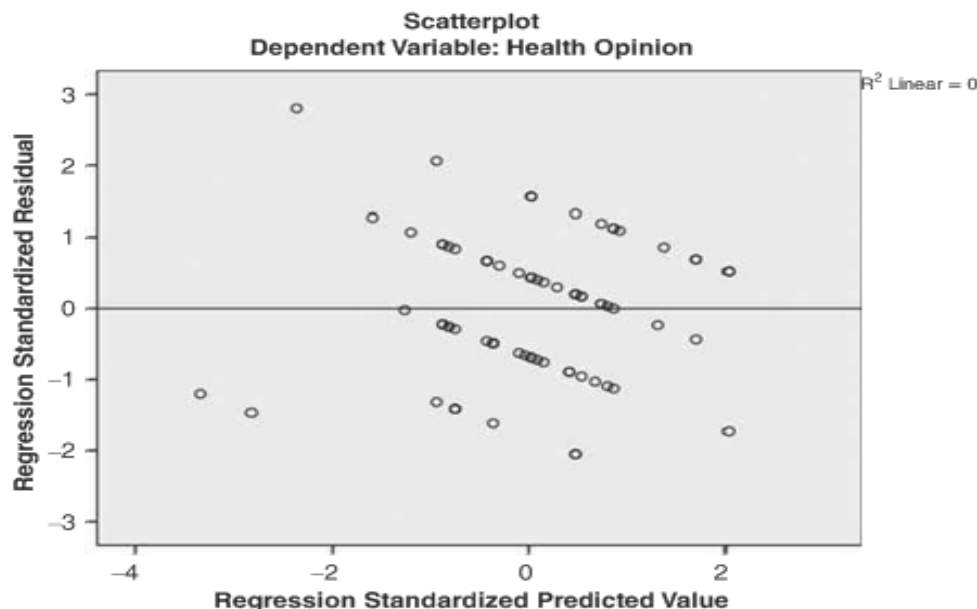


Figure 13.9 The scatterplot between standardized residuals and predicted values.

INTERPRETING THE SPSS® FINDINGS FOR MLR

In what follows, I present the SPSS findings for the study. The Excel program also produces a summary of findings for MLR, but the SPSS program produces a great deal more detail for this type of study. In what follows, I will point out some of the unique features of SPSS that help to explain some of the nuances of the model.

I will discuss several ways to analyze the impact of a set of predictors upon an outcome variable. Although this topic goes far beyond the space I have in this chapter to explain, I will cover the most important procedures and discuss the merits of each. I start with adding the predictors as a set (or block) together so that I can determine the overall impact on the outcome variable. The second method is to add predictors to the model sequentially so that the user can see what happens to the MLR model as each predictor is added (which I refer to as “hierarchical regression”). I will mention further methods that some researchers use to add predictors to the model.

Figure 13.10 shows the result of specifying linear regression through the “Analyze – Regression – Linear” instruction. I presented similar procedures for bivariate regression in Chapter 12 in Figures 12.10 and 12.11. Recall that this series of menus is available through the same Analyze menus. This selection yields the option window in Figure 13.10. I reproduce it here because I want to point out a couple of very important specifications for MLR analyses that go beyond bivariate regression.

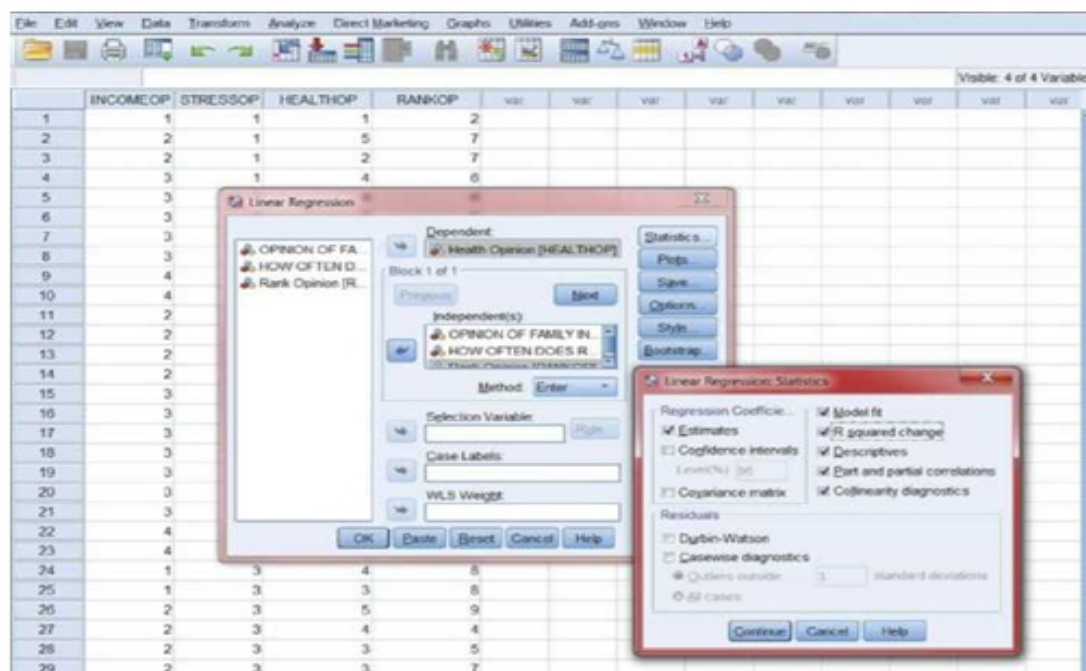


Figure 13.10 The specification menus for SPSS® MLR.

ENTERING PREDICTORS TOGETHER AS A BLOCK

In Figure 13.10, you see that I placed Health Opinion in the “Dependent” window, and the three predictor variables (IncomeOP, StressOP, and RankOP) in the “Independent(s)” window. This allows the researcher to determine the impact of all the predictors together on the outcome variable. The “Statistics” menu button produces the further specification menu in Figure 13.10 and allows me to call for a range of procedures.

- The “Model fit” provides the omnibus test result.
- The “R squared change” provides information relevant to effect size and to further analysis of the impact of each predictor (under a different model for adding predictors to the model, discussed later).
- “Part and partial correlations” are procedures critical for interpreting MLR findings. One very helpful use is to identify and square the part correlation for each variable in the analysis to show the contribution to R^2 for each variable added to the model. I will demonstrate this in my further explanation as follows.
- “Collinearity diagnostics” show the extent to which the predictors correlate to one another as well as to the outcome variable. This “intercorrelation” among the predictors can present difficulties for the MLR analysis. Choosing collinearity diagnostics helps to ensure that the study avoids serious violations of this assumption.

Figure 13.11 shows the further specification of the MLR by calling for several outputs helpful for understanding individual outcomes as well as for checking whether the assumptions have been met. Making these selections results in the output shown in Figures 13.7–13.9.

Partial Regression Plots

This specification also provides scatterplots for each predictor (“partial plots”) with the outcome variable, when the remaining predictors are “held constant.” Holding predictors constant is like saying that the model is assessing the relationship between a single predictor and the outcome variable *at all the different levels of the other predictors*. Thus, the model does not allow the values of the other predictors to interfere with the understanding of how the single predictor uniquely relates to the outcome. Figure 13.12 shows the “Partial Regression Plot” of Rank Opinion with the outcome Health Opinion while the other predictors are held constant.

As you can see in Figure 13.12, there is a positive relationship between rank and health opinion when the other predictors are held constant. In the upper right corner of the figure, you will see that “ R^2 linear = 0.124” which indicates the effect size of this single variable. The further output for the model (which I present later) shows that this effect size is the squared partial correlation for this predictor. (Refer to the “Partial Correlation” section of Chapter 12 to understand this concept.)

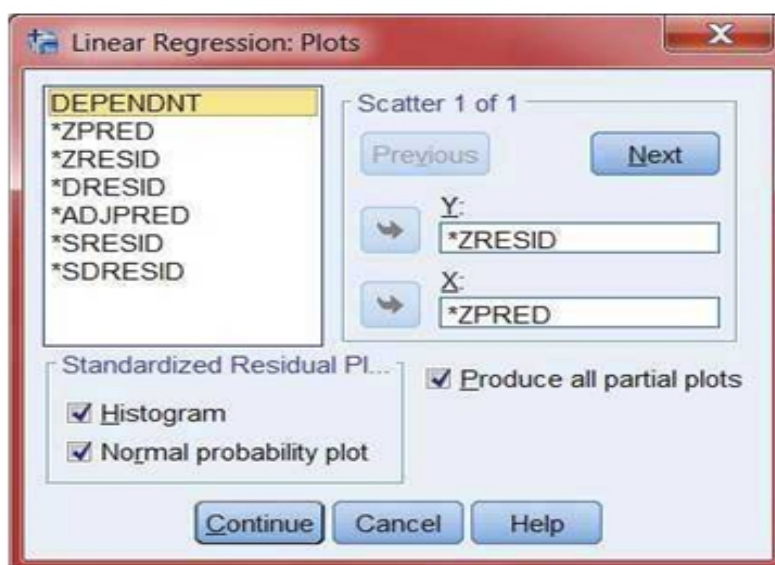


Figure 13.11 The SPSS® MLR specification for assumptions and individual predictors.

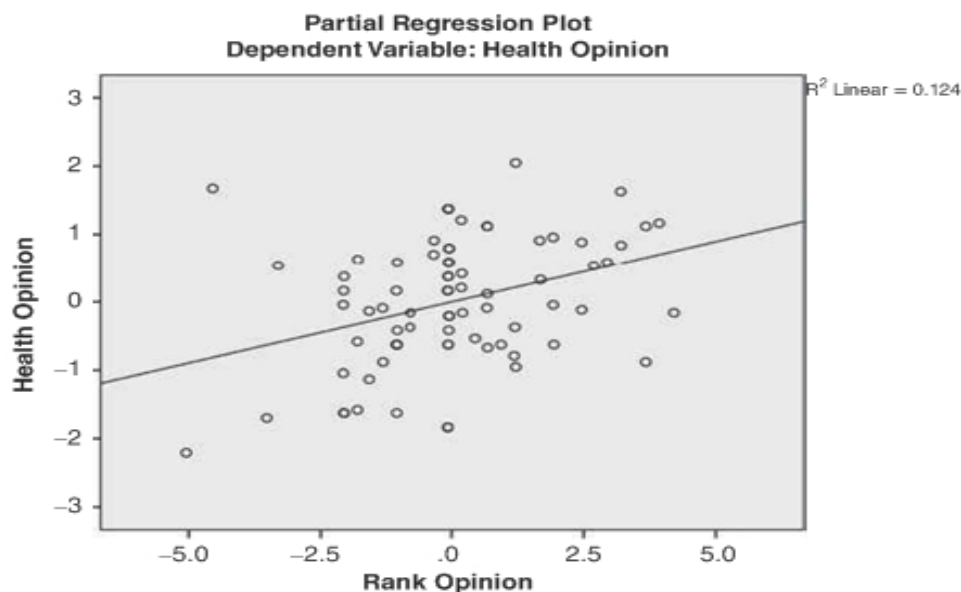


Figure 13.12 The Partial Regression Plot showing Rank Opinion predicting Health Opinion.

Assessing the MLR Findings

When you run the regression analysis (by hitting “OK” on the Linear Regression menu), you will be able to assess the primary elements of the MLR analysis: overall model significance, effect size, and significance of individual predictors. I will address these in the sections that follow.

ENTERING PREDICTORS TOGETHER AS A BLOCK

439

Overall Model Significance

Figure 13.13 shows the ANOVA table that indicates the three predictors significantly predict values of Health Opinion. The F value (6.884) indicates the significance of the entire model, or omnibus test.

Effect Size

Figure 13.14 shows the overall effect size of this model. The adjusted R Square (0.186) indicates that about 18.6% of the variance in Health Opinion is due to the predictors in this model, for this sample group. As I noted in Chapter 12, the Adjusted R Square values take into account sample size (where smaller samples produce greater adjustment typically) and the number of predictors.³

Figure 13.14 also reports “ R Square Change.” (I discussed this value in Chapter 12 with “Bivariate Regression.”) With this method of entry (i.e., entering all predictors as a block), the R Square Change value is equal to the R Square since the model does not highlight the impact of adding predictors separately, as I will show in the section in the following section for the hierarchical regression method of adding predictors. The “ F Change” and “Sig. F Change” values likewise report global figures, while the hierarchical regression method will show these values for each variable added to the regression model.

ANOVA ^a						
Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	16.294	3	5.431	6.884	.000 ^b
	Residual	58.385	74	.789		
	Total	74.679	77			

a. Dependent Variable: Health Opinion

b. Predictors: (Constant), Rank Opinion, HOW OFTEN DOES R FIND WORK STRESSFUL, OPINION OF FAMILY INCOME

Figure 13.13 The MLR omnibus test result.

Model Summary ^b									
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	.467 ^a	.218	.186	.888	.218	6.884	3	74	.000

a. Predictors: (Constant), Rank Opinion, HOW OFTEN DOES R FIND WORK STRESSFUL, OPINION OF FAMILY INCOME

b. Dependent Variable: Health Opinion

Figure 13.14 The effect size summary for the MLR procedure.

³You can reproduce R^2 (0.218) by calculating the proportion of total variance explained as reported in the ANOVA table (Figure 13.13) by calculating the proportion of regression to total variance ($\frac{16.294}{74.679}$) as I discussed in Chapter 9.

Significance of Individual Predictors

The next output shown in Figure 13.15 reports information about each single predictor in the analysis, among other findings. The regression coefficients are reported as “Unstandardized” and “Standardized” according to whether the researcher seeks to use values in their own raw scales or standardized (*Z*) scale values.

The researcher uses these coefficients to show the regression equation as I discussed in Chapter 12. The following is the MLR equation using the unstandardized coefficients:

$$Y_{\text{HealthOP}} = 0.207X_{\text{IncomeOP}} + 0.208X_{\text{StressOP}} + 0.179X_{\text{RankOP}} + 1.295X_{\text{Constant}}$$

Each of the unstandardized “*B*” coefficients for the predictor variables represents the individual impact on the outcome with the other predictors held constant. The constant is the value of *Y* when the predictors = 0. Researchers can use this equation to predict values of HealthOP using values of the predictor variables, much as I discussed with a single predictor (see Chapter 12). (You would follow the same process of substituting values and then constructing confidence intervals around the predictions using the (multiple) standard error of estimate shown in Figure 13.14.)

Figure 13.15 shows the significance of individual predictors using the *t* tests for the various predictors. As you can see, the StressOP ($t = 2.385$, $p < 0.02$) and RankOP ($t = 3.240$, $p < 0.002$) are significant predictors of HealthOP, while IncomeOP ($t = 1.770$, $p < 0.08$) is not.

Collinearity Statistics

The “Collinearity Statistics” columns (last two columns on the right side of Figure 13.15) show the results researchers can use to determine if the intercorrelation among the predictors is too high. “Tolerance” represents the extent of the

Coefficients ^a										
Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Correlations			Collinearity Statistics	
	B	Std. Error	Beta			Zero-order	Partial	Part	Tolerance	VIF
1 (Constant)	1.295	.533		2.430	.018					
OPINION OF FAMILY INCOME	.207	.117	.183	1.770	.081	.215	.201	.182	.983	1.017
HOW OFTEN DOES R FIND WORK STRESSFUL	.208	.087	.245	2.385	.020	.238	.267	.245	.998	1.002
Rank Opinion	.179	.055	.336	3.240	.002	.358	.352	.333	.965	1.015

a. Dependent Variable: Health Opinion

Figure 13.15 The SPSS[®] output for individual predictors.

intercorrelation of the predictors.⁴ Generally, the lower the Tolerance value the greater the intercorrelation complications. Thus, if the Tolerance value is too low (e.g., <0.10), there is not much variance left to explain in the outcome variable once the intercorrelation of the predictors is accounted for.

The “VIF” (variance inflation factor) values are derived from the Tolerance values, so higher numbers are problematic (e.g., $VIF > 10$ may indicate problems of multicollinearity).

The Squared Part Correlation

Part (or Semipartial) correlations are very important values for understanding the impact of individual predictors upon the outcome variable. The Part correlation is not to be confused with the Partial correlation that I discussed in Chapter 12. As you recall, I defined the Partial correlation as the correlation of a predictor variable with the outcome *after you remove the influence of the remaining predictors from both of the other variables (target predictor with the outcome variable)*.

Part correlations show the relationship of a predictor to an outcome when the effects of the remaining predictors are taken out of the predictor variable but not the outcome variable. Consider Figure 13.16 (this is a reproduction of Figure 12.31) that shows these relationships in the study with two predictors (see the discussion of this figure and “Partial Correlation” in Chapter 12).

The Part or Semipartial correlation is shown as the relationship of the “A” section to the entire variance of the outcome variable (i.e., $\frac{A}{A+B+C+D}$).⁵ Thus, in this example, the Part correlation of IncomeOP with HealthOP is represented by the unique correlation between IncomeOP and HealthOP (i.e., section A) as a proportion of the entire variance of HealthOP (i.e., sections $A + B + C + D$).

The importance of this measure is that, when squared, it represents the unique contribution to the outcome variable by one single predictor (when the other predictors are held constant). In the current study with three predictors, you can use the SPSS output for individual predictors to determine the unique impact of each predictor on the outcome variable.

Thus, if you look at Figure 13.15 you can see that squaring the part correlation for RankOP ($0.333^2 = 0.11$) results in the percent of variance in HealthOP contributed uniquely by RankOP, (i.e., 11% of the variance in HealthOP is contributed uniquely by RankOP.) In similar fashion, you can square the part correlations for the remaining predictors to show their unique contributions to the variance in HealthOP. Thus StressOP accounts for 6% (or 0.245^2) unique variance in HealthOP while IncomeOP accounts for 3% (0.182^2). Comparing all the predictors, it appears that RankOP is a better predictor of HealthOP than either StressOP or IncomeOP for this set of data.

⁴Tolerance is calculated by subtracting the squared multiple correlation of the additional predictors with the predictor in question from 1.

⁵This is similar to the Partial correlation as referenced in Chapter 12, but that correlation is represented by the relationship of the “A” section to the $(A + D)$ sections.

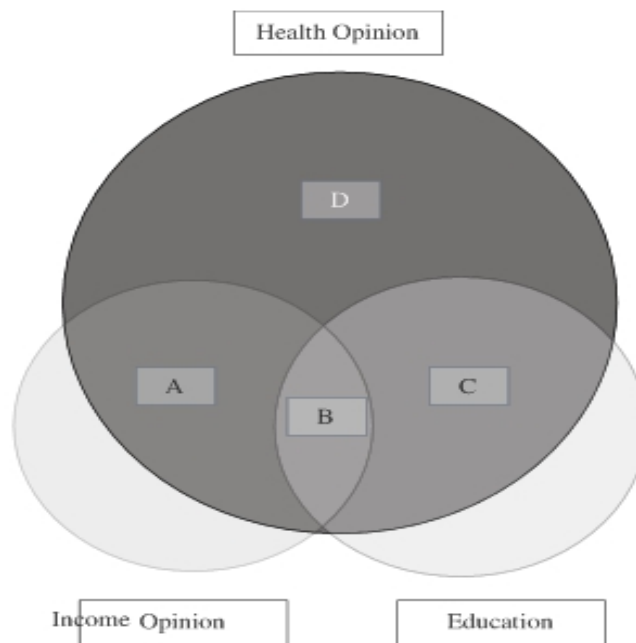


Figure 13.16 Isolating the effects of a predictor variable on an outcome variable through part or semipartial correlation.

ENTERING PREDICTORS SEPARATELY

This second method of entering predictors to an MLR analysis allows the researcher to specify changes in model values when successive predictors are added to the analysis. This constitutes hierarchical regression analysis since each predictor contribution is added sequentially to the others in “building block” fashion. The other facet of this method that sets it apart from other entry methods is that the predictor variables are added to the model according to the researcher’s theoretical understanding of the relationships among the variables, or some other a priori rationale.

Figure 13.17 shows the beginning SPSS specification menu for MLR as we discussed with the entry method earlier (i.e., “Entering Predictors Together as a Block” section). As you can see, I entered the IncomeOP as the first predictor because of the theoretical notions I described earlier in the chapter about the widespread assumptions between income and health.

I can add the other predictors separately, as separate “Blocks” by choosing the “Next” button after I add each predictor. Thus, in Figure 13.17, I entered IncomeOP first, and then can hit the Next button to add work stress (StressOP) as a predictor because of the emphasis in some research literature on the importance of work to health. Last, I can hit the Next button again to add the RankOP predictor given the potential importance of social status to health (as I noted in the research reported by Marmot, 2004).

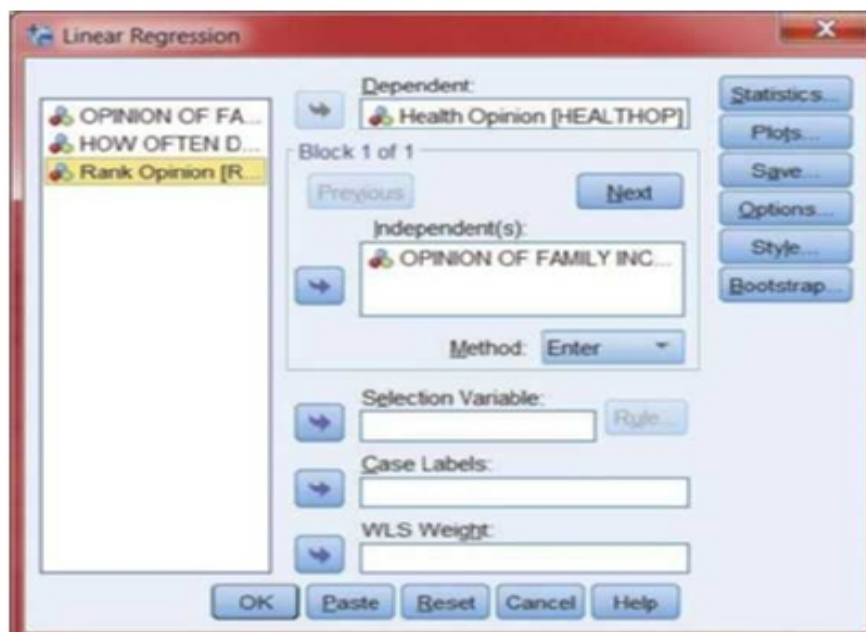


Figure 13.17 The MLR entry method for the first predictor.

The resulting MLR output enables the researcher to determine omnibus test finding, effect size, and individual predictor significance as in the former entry scheme. However, when I enter the predictors in separate blocks, the MLR results produce the results in hierarchical fashion created by adding predictors.

Notice that in Figure 13.18 Model 3 results will be the same as the results for the omnibus results in the first entry method (combined entry) since it includes all the variables specified in the study. This will be the same for the following output, which shows that results for Model 3 are the same in both entry schemes for effect size and individual predictor analyses.

Overall Model Significance

Figure 13.18 shows that the model with only the first predictor (HealthOP) as not being significant ($F = 3.696, p > 0.05$) but nearly so (see "Model 1" results). When the second predictor (StressOP) is added, the model is significant ($F = 4.506, p < 0.014$) (Model 2 results). When RankOP is added, the model is also significant ($F = 6.884, p < .000$) (Model 3 results).

Effect Size (Multiple R^2)

The R Square is likewise reported for each model representing the sequential addition of predictors. As you can see in Figure 13.19, the R Square value changes from 0.046 in Model 1 to 0.107 in Model 2, and to 0.218 in Model 3. Taken together (Model 3),

ANOVA ^a						
Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	3.463	1	3.463	3.696	.058 ^b
	Residual	71.216	76	.937		
	Total	74.679	77			
2	Regression	8.011	2	4.005	4.508	.014 ^c
	Residual	66.669	75	.889		
	Total	74.679	77			
3	Regression	16.294	3	5.431	6.884	.000 ^d
	Residual	58.385	74	.789		
	Total	74.679	77			

a. Dependent Variable: Health Opinion

b. Predictors: (Constant), OPINION OF FAMILY INCOME

c. Predictors: (Constant), OPINION OF FAMILY INCOME, HOW OFTEN DOES R FIND WORK STRESSFUL

d. Predictors: (Constant), OPINION OF FAMILY INCOME, HOW OFTEN DOES R FIND WORK STRESSFUL, Rank Opinion

Figure 13.18 The omnibus test results for hierarchical MLR.

Model Summary ^d									
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	.215 ^a	.046	.034	.968	.046	3.696	1	76	.058
2	.328 ^b	.107	.083	.943	.061	5.116	1	75	.027
3	.467 ^c	.218	.186	.888	.111	10.499	1	74	.002

a. Predictors: (Constant), OPINION OF FAMILY INCOME

b. Predictors: (Constant), OPINION OF FAMILY INCOME, HOW OFTEN DOES R FIND WORK STRESSFUL

c. Predictors: (Constant), OPINION OF FAMILY INCOME, HOW OFTEN DOES R FIND WORK STRESSFUL, Rank Opinion

d. Dependent Variable: Health Opinion

Figure 13.19 The hierarchical results for effect size.

the three predictors account for 21.8% of the variance in HealthOP (or 18.6% using the Adjusted R Square) which agrees with the combined entry method as shown in Figure 13.14.

By using separate entry, you can see how the model results change when you add predictors sequentially. Thus, Model 2 shows that adding StressOP to IncomeOP creates a significant omnibus model, and it adds a significant “amount” of R Square. This addition to R Square is represented in the “Change Statistics” columns as the “ R Square Change (0.061)” which shows that StressOP adds 0.061 to the previous overall R^2 (.046) to yield an accumulated value of 0.107.

This change also registers in the “Sig. F Change” column where F Change (5.116) is considered significant ($p < 0.027$). The F Change references the change in the omnibus test, but it is a referent for the R Square Change as well.

Individual Predictors analyses

Adding predictors separately is also very helpful because it allows the researcher to see how the effect sizes change as each is added. Figure 13.20 shows the significance tests and effect size values (i.e., part correlations) for each predictor as they are introduced. You interpret these values similarly to the interpretations produced by the separate entry method. The difference is that you can see the changes to significance and effect sizes as predictors are separately entered.

Note, for instance, that IncomeOP changes in significance depending on the variables in the equation. In Model 1, it is close to significance ($p = 0.058$), in Model 2 it is significant ($p = 0.043$), while in Model 3 it is not significant ($p = 0.081$). The changes are small, but they show that adding predictors changes the interpretation of the findings.

This dynamic is similar for examining part correlations. As with significance numbers, the output shows part correlations for each variable within the block it is introduced. For example, the part correlation for IncomeOP is 0.215 in Model 1 where it is the only predictor. By Model 3, the part correlation has dropped to 0.182 when the two other predictors have been added. Apparently this variable shares variance with variables that are added, as can be understood by squaring the part correlations.⁶ (In this particular case, it is understandable since income and social status are quite related variables.)

I mentioned previously that squaring the part correlations yields the unique contribution to understanding the variance in the outcome variable of each predictor.

Coefficients ^a									
		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Correlations		
		B	Std. Error	Beta			Zero-order	Partial	Part
1	(Constant)	2.894	.385		7.522	.000			
	OPINION OF FAMILY INCOME	.244	.127	.215	1.922	.058	.215	.215	.215
2	(Constant)	2.235	.475		4.708	.000			
	OPINION OF FAMILY INCOME	.254	.123	.225	2.059	.043	.215	.231	.225
	HOW OFTEN DOES R FIND WORK STRESSFUL	.209	.093	.247	2.262	.027	.238	.253	.247
3	(Constant)	1.295	.533		2.430	.018			
	OPINION OF FAMILY INCOME	.207	.117	.183	1.770	.081	.215	.201	.182
	HOW OFTEN DOES R FIND WORK STRESSFUL	.208	.087	.245	2.385	.020	.238	.267	.245
	Rank Opinion	.179	.055	.336	3.240	.002	.358	.352	.333

a. Dependent Variable: Health Opinion

Figure 13.20 The individual predictor summary for separate entry.

⁶This feature of separate entry is helpful for determining spuriousness. If, for example, bivariate correlations are high but drop considerably when a separate predictor is added, the bivariate relationship may be spurious to the last predictor added.

TABLE 13.2 The Change Values Resulting from Adding Predictors

		R Square	R Square Change	Part Corr.	Part Corr. Squared for last variable added
Model 1	IncomeOP	0.046	0.046	0.215	0.046
Model 2	ADD StressOP	0.107	0.061	0.247	0.061

Another feature of the output for separate entry is that the *R* Square Change values (noted in Figure 13.19) correspond to squared part correlations. The *R* Square Change value reflects the changes represented by the last variable entered into the analysis. These changes are mirrored by the squared part correlations. Table 13.2 shows an example of this.

In Table 13.2, IncomeOP is the first variable entered. Therefore, the *R* Square and *R* Square Change for IncomeOP will be the same values since the model changes from “no *R* Square” to an *R* Square that results from adding one predictor to the equation. In this example, the Part Correlation Squared is also the same value (0.046) since it is the last variable entered into the equation. Thus, the *R* Square changes from 0 to 0.046.

Table 13.2 shows the changes to the initial results when a second variable (StressOP) is added to the equation. In this case, the *R* Square Change (0.061) and “Part Corr. Squared for Last Variable Added” value (0.061) for StressOP are the same. The overall *R* Square value increases by a value of 0.061 (0.107 – 0.046) as a result of adding StressOP to the equation.

You can see similar results for adding the third predictor, RankOP. The *R* Square Change value (0.111) reported in Figure 13.20 is equal to its squared part correlation ($0.111 = 0.333^2$) when it is the last variable entered. Thus, the “*R* Square Change” and squared part correlation of a variable represent the increase in variance of the outcome variable (HealthOP) explained by adding the predictor.

In order to observe the “final” accounting of variance by the individual predictors, you can square the part correlations shown in the final model. Despite how they were entered, the final squared part correlations indicate the unique magnitude of impact of each predictor on the outcome variable.

One important point to remember is that *no matter how the variables are entered into the regression equation, the squared part correlation will indicate the unique contribution of a predictor to the variance of the outcome variable*. If, for example, we reran the MLR analysis but added RankOP in the first model and added IncomeOP in the third model, the *R* Square and *R* Square Change reported values would change (along with many others). However, the part correlations in the final model will be the same as in the first MLR analysis. This is a very important feature for interpretation since there are other ways to enter predictors to a model that are contentious among researchers.

ADDITIONAL ENTRY METHODS FOR MLR ANALYSES

I presented two methods for entering predictors in an MLR analysis (adding predictors together and adding predictors separately). Other choices for entry include “Stepwise” which allows SPSS pre-set entry and removal criteria to enter and retain each variable in the analysis. For example, if I had a study with five predictors, the stepwise procedure would add and retain (or delete) predictors depending on the size of their impact on the model. It might be the case that only two predictors would be retained and the others excluded even though the researcher has good reason to include them.

The stepwise procedure can be useful at times, but most researchers avoid the method since it takes the choice for building the model out of the researcher's control. If the researcher is including each predictor on some a priori grounds for inclusion, then it is important to see the impact on the results despite the variable not “making the cut” established by the program.

If you would like to examine the stepwise (or other) method for entry, you can specify it by choosing the “Entry” button on the main MLR specification window (see Figure 13.10). Figure 13.21 shows this window in which I chose “Stepwise” under “Method” which produces a separate menu of methods to enter predictors other than

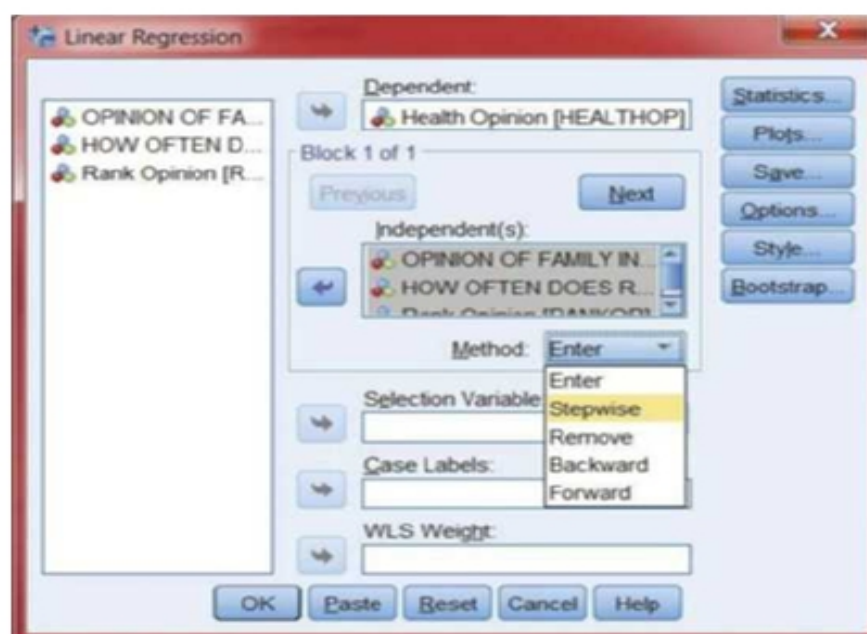


Figure 13.21 The specification menu for the Stepwise (and other) entry method(s).

“Enter”. If you were to rerun the MLR I have discussed before, the result would be that two variables would be retained in rank order of their contribution to the outcome (RankOP and then StressOP) according to the criteria used in the program to include and/or exclude variables.⁷ In practice, this means that IncomeOP would not be retained despite its theoretical importance and its near significance ($p = 0.081$) in the previous models.

The complete menu of entry schemes includes other methods that exceed the boundaries of this treatment of MLR. However, most researchers can profitably use hierarchical regression (“Enter” method) for their analyses.

EXAMPLE STUDY CONCLUSION

The example study concludes that the two predictors, RankOP and StressOP are significant predictors of HealthOP, while IncomeOP is nearly significant. According to the squared part correlations, RankOP is the best predictor among these variables accounting for 11% of the variance in HealthOP. The other predictors also contribute to an understanding of the variance in HealthOP, but by lesser values (StressOP = 6% and IncomeOP = 3%).

If this were an actual study, we might suggest that social status may be more important than income or work stress for explaining health outcomes. Those with higher social statuses enjoy better health despite the levels of their work stress and/or incomes. These findings would support Marmot’s (2004) notions as I discussed above.

TERMS AND CONCEPTS

Centered leverage values (CVL):	Diagnostic values that indicate distances for the independent variable. The greater the distance between each x value and the mean of x , the larger the centered leverage value, whereas values closer to the mean of x are increasingly smaller.
Collinearity statistics:	Statistical values (“Tolerance” and “VIF”) reported that indicate whether the intercorrelation among the predictors is too high to obtain meaningful MLR results.
Cook’s distance values (CDV):	Diagnostic values that measure the effect on all the residuals if a particular case were dropped from the analysis. It therefore is <u>a measure of both independent and dependent variables</u> acting in concert in the analysis, with large values indicating more impact on the regression coefficients.

⁷ You can find these criteria under the Options button of the Linear Regression specification menu like that shown in Figure 13.21.

Distance statistics:	A diagnostic measurement that assesses how each score lies in reference to all the other scores in a regression analysis. (See Centered leverage values and Cook's distance values.)
Fixed-effects modeling:	An MLR analysis that uses the same categories of variables across studies. The predictor variable has a set number of categories, like gender or experimental method, where the same categories are used in all situations. In this situation, generalizability of findings is limited to the fixed category values.
Hierarchical regression:	An interpretive technique in multiple regression studies in which the sequential impact of predictor variables on the outcome variable is measured.
Influence statistics:	Diagnostic values that identify the impact on regression coefficients <u>if individual scores are deleted from the analysis.</u>
Multicollinearity:	A measure of the "overlap" (or intercorrelation) among predictors in a multiple regression study. This overlap can obfuscate the overall relationship of the set of predictors to the outcome variable.
Part Correlation:	Part correlations show the relationship of a predictor to an outcome when the effects of the remaining predictors are taken out of the predictor variable but not the outcome variable. (Also called Semipartial correlations.)
Partial Regression Plots:	These are scatterplots for each predictor ("partial plots") with the outcome variable, when the remaining predictors are "held constant."
Random effects modeling:	The most common MLR analysis used primarily in non-experimental social science research analyses; it is used in situations where the evaluator takes a random set of observations from a larger population of values, with the categories or levels of the predictor variable "emerging" from and determined by the data sampled. The categories and values of x are thus random.
Residuals:	Residuals are errors in predicting true values of an outcome variable in a regression procedure
Semipartial correlations:	(See Part correlation.)
Squared part correlation:	Values that indicate the percent of variance in the outcome variable uniquely explained by each individual predictor (when other predictors are held constant).
Stepwise entry method:	This method allows statistical programs to enter/remove and retain predictors in an MLR analysis according to pre-set entry and removal criteria.

TABLE 13.3 The Study Variables and Labels

SPSS Variable	Full Variable Label	Values
JobOpp	Job Opportunities	Low (1) -- High (7)
HlthServices	Health Services	Low (1) -- High (7)
RecOpp	Recreational Opportunities	Low (1) -- High (7)
CommInvolv	Community Involvement	Low (1) -- High (7)
FearCrime	Fear of Crime	Low (1) -- High (7)

DATA LAB AND EXAMPLE (WITH SOLUTION)

In this lab, I present several cases of the community survey I have used in past chapters as an example for MLR.⁸ For this example study, investigate the data to determine whether, and to what extent, Community Involvement is predicted by the other four (predictor) variables. Thus, to what extent is community involvement predicted by respondent perceptions of job opportunities, recreational opportunities, health services, and fear of crime? Table 13.3 shows the study variables, their labels, and their values and Table 13.4 shows the study data.

Lab questions: (Enter predictors together as a single block.)

1. Are assumptions met for MLR?
2. What are the findings: omnibus test, omnibus effect size, and individual predictor significance?
3. What are the unique explanations of variance in community involvement by each of the predictor variables?
4. What conclusions can you draw about the research question from the findings of this example study?
5. Briefly examine diagnostics (standardized residuals, Cook's, CLV). Observations?

DATA LAB: SOLUTION**1. Are assumptions met for MLR?**

Normal distribution According to Figure 13.22, the outcome variable (community involvement) appears to be fairly balanced with a very slight positive skew. Figure 13.23 shows the Sk/Ku findings that indicate the study variables are within the boundaries of a normal distribution.

⁸Once again, due to the small sample size and other features of the data, no conclusions should be drawn from the analyses. The data are shown here for illustration of MLR procedures only.

TABLE 13.4 Problem 1 MLR Data

JobOpp	HlthServices	RecOpp	CommInvolv	FearCrime
2	4	2	2	4
1	1	1	1	1
7	6	6	6	1
4	4	5	4	3
6	5	6	5	5
1	3	1	1	4
7	6	6	6	2
4	5	4	4	3
5	7	7	5	7
5	6	5	5	2
2	5	5	3	2
4	3	4	3	3
4	6	6	5	1
3	2	6	3	5
4	4	7	4	4
4	4	4	4	5
4	5	4	3	7
4	5	5	5	4
3	4	4	2	4
2	2	4	3	2
4	7	6	4	6
3	5	2	2	4
2	5	4	2	4
6	5	6	6	1
4	4	4	3	1
4	4	5	5	4
3	5	4	4	6
1	1	1	1	7
2	5	4	4	3
2	4	2	1	3
6	6	6	4	3
3	5	7	3	4
3	3	3	1	5
4	6	4	2	4
4	4	7	6	3
2	2	6	4	4
1	7	6	4	2
5	2	7	6	1
7	7	5	7	1
3	3	7	2	1
1	6	1	3	7

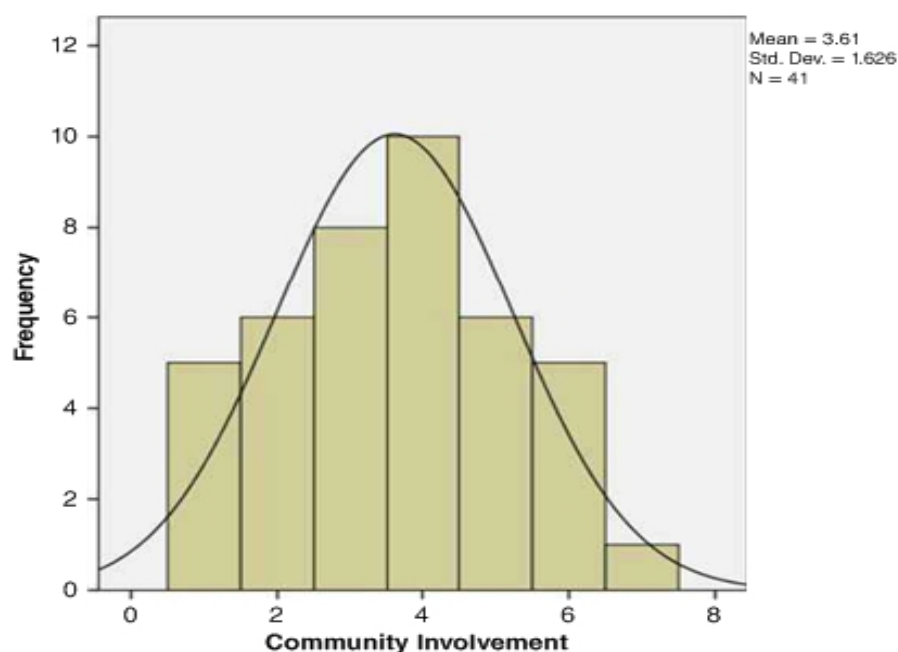


Figure 13.22 The histogram for the outcome variable.

Report					
	Community Involvement	Job Opportunities	Recreational Opportunities	Health Services	Fear of Crime
Mean	3.61	3.56	4.61	4.46	3.49
N	41	41	41	41	41
Std. Deviation	1.626	1.689	1.829	1.614	1.846
Kurtosis	-.782	-.395	-.485	-.416	-.632
Std. Error of Kurtosis	.724	.724	.724	.724	.724
Skewness	.050	.348	-.597	-.400	.334
Std. Error of Skewness	.369	.369	.369	.369	.369

Figure 13.23 The findings indicating all study variables are normally distributed.

Multicollinearity Multicollinearity is controlled for this set of data according to Tolerance and VIF values.

Multivariate normality, linearity, and equal variance (using residual analysis)

Standardized residuals normally distributed. Figure 13.24 shows the residuals are fairly normally distributed for the study variables.

Random relationship between residuals and predicted values. Figure 13.25 shows a random array of residuals plotted with predicted values.

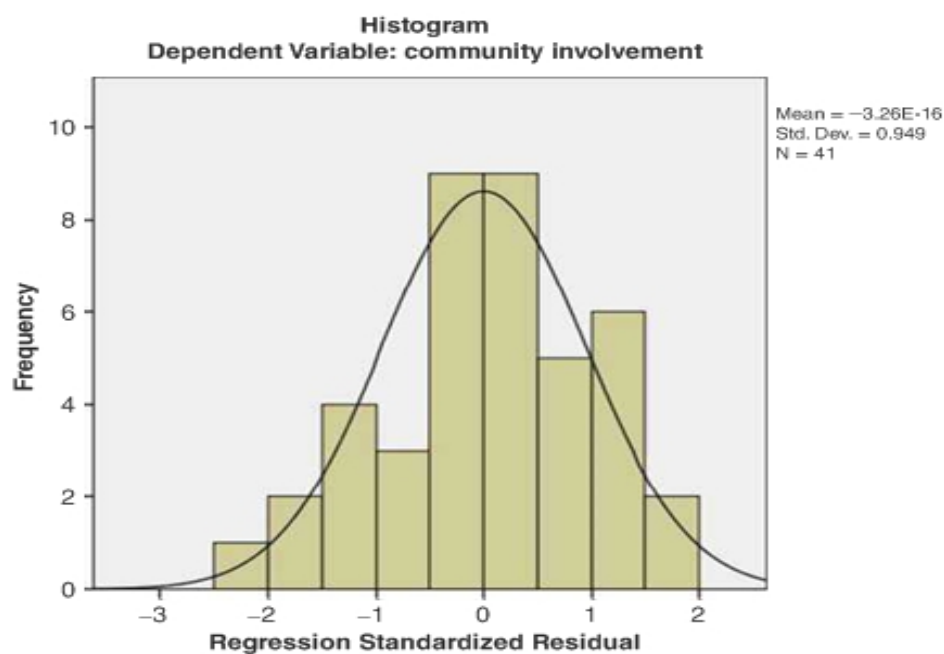


Figure 13.24 The residuals plot for normality assessment.

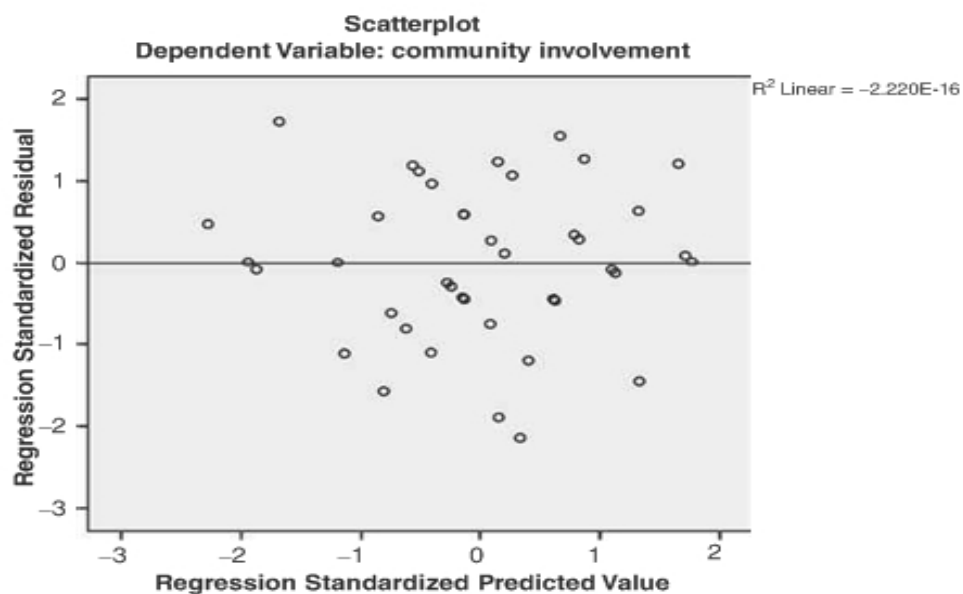


Figure 13.25 The scatterplot between standardized residuals and predicted values.

2. What are the findings (omnibus test, omnibus effect size, individual predictor significance)? (Predictors entered together as a single block.)

Omnibus test – The overall model is significant ($F = 19.610, p < 0.000$). Community involvement is significantly predicted by respondent perceptions of job and recreational opportunities, health services, and fear of crime.

Effect size – The R Square = 0.685 (Adjusted R Square = 0.650) indicating 68.5% of the variance in community involvement is accounted for by the predictor variables in this respondent sample.

Individual predictor significance – Job Opportunities ($t = 3.640, p = 0.001$) and Recreational Opportunities ($t = 3.004, p = 0.005$) are significant predictors of community involvement. Health service and fear of crime are not significant predictors.

3. What are the unique explanations of variance in community involvement by each of the predictor variables? Table 13.5 shows the squared part correlations and interpretations for each predictor variable. As you can see, job opportunities and recreational opportunities account for the greatest proportion of community involvement in this example study.

TABLE 13.5 The Unique Contribution to Outcome Variance by Predictor Variables

Predictor	Part Correlation	Squared Part Correlation	% of unique variance in Community Involvement accounted for by the Predictor
Job Opportunities	0.340	0.116	11.60%
Recreational Opportunities	0.281	0.079	7.90%
Health Services	0.139	0.019	1.90%
Fear of Crime	-0.080	0.006	0.60%

4. What conclusions can you draw about the research question from the findings of this example study? The respondents in this example study indicated that their perceptions of job opportunities and recreational opportunities were most important in their perceptions of community involvement. Perceptions of health services and fear of crime were minor considerations.

5. Briefly examine diagnostics (Standardized Residuals, Cooks, CLV). Observations?

- One ZResidual value (-2.14) is a bit extreme, but not excessive.
- One Cook's value (0.25883) is high and may be problematic; a second value (0.21808) may also bear investigation.
- One CLV value (0.35911) appears to be high, but not inordinately so.