Chapter 10
Introduction to Multiple Regression

In the previous chapters, we have focused on relationships between only two variables at a time. Most relationships that we are interested in social science research are more complicated, however, than can be understood with only bivariate analyses. In some cases, the relationship between two variables depends entirely on a third variable, and so a bivariate analysis can be misleading. In some cases, the relationship between two variables depends on, or operates through, additional variables. In this chapter, we will discuss multiple regression. In multiple regression analysis, a single outcome variable is modeled as a linear combination of as many additional variables as desired. Multiple regression is sometimes used simply to understand factors that are relevant to predicting an outcome, but it is also used in science to help establish that the relationship between two variables is a causal one and to understand complex relationships that simply cannot be understood with bivariate analyses. In this chapter, before introducing the details of multiple regression, we will discuss causal thinking in order to lay the foundation for seeing why multivariate analyses are necessary.

10.1 Understanding Causality

Although we have not discussed causality until now, the assumption that one of our goals in statistical analysis is often to establish causal relationships between variables has been implicit. We usually want to know whether group differences in means exist, because we may think the grouping variable *causes* the outcome variable. For example, we may be interested in examining whether there are gender differences in earnings because we believe that discrimination in the labor market is present, thus causing earnings differences. We may be interested in examining the correlation between education and earnings, because we may believe that education develops skills (or perhaps just credentials) that are valuable on the market and therefore rewarded with higher wages. We may be interested in examining differences in marital status distributions across regions of the country, because

S.M. Lynch, *Using Statistics in Social Research: A Concise Approach*,
DOI 10.1007/978-1-4614-8573-5_10,

we may believe that cultural practices vary across regions and make individuals more likely to marry, less likely to divorce, etc. Each of these examples implies a causal process in which the "independent variable," commonly denoted x, affects the "dependent variable," (or "outcome variable") commonly denoted y.

Causal arguments are ubiquitous in discussions among lay people. One only needs to look at comments on news blogs as people discuss virtually any topic. Is spanking a practice that enhances or discourages successful development in children? Does raising taxes increase or reduce growth? Does exercise reduce the risk of obesity and heart disease? Does having a strong safety net encourage dependency? While arguments on both sides of any debate often explicitly use causal language, rigorously establishing that a relationship between x and y is a causal one is incredibly difficult. First, defining causality itself is a difficult enterprise, one with which philosophers have struggled for literally millenia. Once defined, the process of determining whether the relationship between x and y meets the criteria for causality in social research is at least as difficult, because the process is riddled with the limitations of data and methods. For this reason, I have intentionally avoided discussing causality until this chapter. The previous chapters, as discussed above, have certainly alluded to the use of statistical testing as a means to peer into the window of causality, but I have intentionally avoided using the results of analyses to make causal claims, much as I have avoided claiming that the results of analyses prove any hypothesis to be true.

A very basic, initial problem with making causal claims can be seen by rediscussing the fallacy of affirming the consequent. Recall from Chap. 2 that the key reason that we cannot "prove" a hypothesis true is that alternate explanations may always exist that account for patterns we observe in data. Recalling the example in Chap. 2, I may claim that, if it rains today, my yard will be wet when I get home. When I get home, if my yard is, in fact, wet, I cannot conclude that my claim of rain is true, because my wife could have watered the lawn,[1] a water main could have broken, etc. This same rationale is true when we are seeking to make causal claims. For example, I could say: If A causes B, then I will find Z in my data (whatever Z is). Clearly, as per *modus tollens,* if I find that Z is not true in my data, I can conclude that A does not cause B. However, if I find Z, I cannot immediately conclude that A causes B. To do so requires that I meet a number of criteria, especially including most importantly that I rule out all alternate explanations.

What is Z? In other words, what should we observe if A causes B? The classic understanding of causality used in social science is that A is a cause of B if (1) there is some relationship between A and B, (2) the relationship is temporally ordered so that A precedes B in time, and (3) the relationship between A and B is not spurious.

The first requirement, correlation/association, is straightforward: If A causes B, then A and B must be associated in some fashion. If we think of A as an event, then when A happens, all else being equal, B must happen. If pushing someone

[1] In fact, my wife has informed me that she would never do so. We may have to talk about this, but doing so is beyond the scope of this chapter.

causes falling, then when I push someone, they should fall, all things considered. This requirement implies the second requirement: that the cause must precede the effect. Scientific views of causality do not allow for concepts like "destiny" in which effects produce causes: causes must come first.[2]

While establishing the first two requirements may seem straightforward, it is not necessarily easy to do so. First, two variables may be related to one another, but not linearly. That is, if the correlation coefficient is the measure you are using to capture the association between two variables, and the relationship is u-shaped, the correlation will be 0, even though a relationship clearly exists. Thus, it is important to properly model the shape of the relationship between two variables before appealing to *modus tollens* and *ruling out* a causal relationship, and multiple regression enables doing so, as we will discuss later in the chapter.

Second, it is not straightforward to establish temporal order. In cross-sectional data like the GSS in which all variables are measured simultaneously on individuals, we cannot establish which variable comes first except in rare cases. For example, parents' completed educational attainment almost certainly precedes respondent's educational attainment, and so the temporal order may be assumed. However, suppose we were interested in the relationship between health and happiness. Does having one's health make one feel happier, or does being happy make one feel healthier? Or, is the causal relationship possibly even reciprocal?

Having panel data—that is, data in which the same respondents are measured on more than one occasion—can help establish temporal order, but even with such data, temporal ordering is not always clear. For example, in examining the relationship between education measured at time 1 and health measured at time 2, we might assume that education precedes health. However, it may be the case that health prior to time 1 (time 0) is the cause of education at time 1. While it may be the case that education and health are both causes and effects of each other—so that health affects education, which then affects later health—more problematic is that health at time 0 may be the cause of both education at time 1 and health at time 2, and education is not a cause of health at all. This example brings us to the third requirement: non-spuriousness.

The word "spurious" in common language simply means false. In statistical usage, a spurious relationship between A and B is one in which a third variable, C, accounts entirely for their relationship. In other words, C is the cause of both A and B, and so the apparent relationship between A and B is a false one. A classic example of a spurious relationship is that between ice cream consumption rates (say gallons consumed per 100 people per week) and violent crime rates (say number of assaults reported per 100 people per week): the relationship is entirely explained by season (temperature). People eat more ice cream in warmer weather, and more violent crime is committed during warmer weather. Once temperature is taken into

[2]Teleological arguments—arguments in which actions in nature are purposeful, with some sort of predetermined endpoint that causes actions along the way to the endpoint—have long been rejected in science.

account, the relationship between ice cream consumption and crime vanishes. This example also suffers from a possible ecological fallacy: if the relationship were not spurious, is the implication that ice cream consumers are more likely to be criminals? As another, individual level example, consider the relationship between siblings' heights. Sibling heights are certainly correlated with one another, but the relationship is certainly not causal: the relationship is spurious because parents' height fully explains the relationship (see Davis 1985).

How do we rule out spuriousness as an explanation for the relationship between two variables of interest? There are two ways that researchers attempt to do so. One is the use of experimental methods, and the other is the use of statistical methods like multiple regression modeling.

10.1.1 The Counterfactual Model and Experimental Methods

To understand how experimental methods work, it is useful to define the counterfactual model of causality. At a most basic level, in order to demonstrate that some factor is a cause of something else, we must be able to show that the world, after receiving the cause, is different than it would have been without the cause. Consider a case in which we are trying to determine whether some new drug has an effect on reducing blood cholesterol. To demonstrate that the drug reduces cholesterol, we need to be able to show that an individual's cholesterol falls after taking the drug and that, counterfactually, his cholesterol would not have fallen had he not taken the drug, all else being equal—in other words taking vs. not taking the drug is the only thing that differs between the two conditions.

The counterfactual model is intuitively easy to grasp, and it should be employed at least as a thought experiment when thinking about causality. For example, consider the argument that gun control measures do not work, because we have some measures in effect already, and we still have a high level of gun violence. If we think counterfactually, we might ask: well, what would the level of gun violence be if all else were equal and we *didn't* have the measures in effect that we already have? Suppose we claim that toothpaste brand x prevents tooth loss, but we observe that, at age 70, a person who has used brand x for his entire adult life begins to lose his teeth. Does that imply that brand x does not prevent tooth loss? The counterfactual approach would ask: at what age would the person have begun losing his teeth if he *hadn't* used brand x?

While the counterfactual approach is an extremely useful way to think about causality, it is impossible in reality to implement. We cannot go back in time and see what our level of gun violence would have been had the gun control measures currently in effect had not been in effect. We cannot subtract 50 years of life off an individual to see the age at which tooth loss would have begun had he not used brand x. In short, it is impossible to observe a single individual, nation, or other unit of analysis in two states in which nothing differs except for some treatment of

Step 2 Group	Step 3 Pre-test	Step 4 Treatment	Step 5 Post-test	Change
R(T)	O_{it}	X	$O_{i(t+1)}$	$O_{i(t+1)} - O_{it} = \Delta_i$
R(C)	O_{jt}		$O_{j(t+1)}$	$O_{j(t+1)} - O_{jt} = \Delta_j$
Difference:	$\Delta_{(i-j)t} = 0$		$\Delta_{(i-j)(t+1)}$	$(\Delta_i - \Delta_j)$

Table 10.1. Illustration of a true experimental design applied to two individuals: i and j. $R(T)$ and $R(C)$ represent random assignment to treatment and control groups, respectively. Δ represents change or between individual differences.

interest. Suppose, for example, we measure an individual's cholesterol at time 1, administer the drug for a month, and then remeasure cholesterol at time 2. Then, we wait another month without the drug and remeasure cholesterol again at time 3. So, we have observed the person both with and without the drug. This strategy is a type of quasi-experimental design—a "time series design"—and is, in fact, commonly used in research (see Campbell and Stanley 1963). A key problem with this approach is that we have no way of knowing that nothing changed over the extended time period that may exacerbate or mitigate the estimated effectiveness of the drug. Perhaps the individual's diet changed. Perhaps the weather changed, and it somehow affects cholesterol. Perhaps the initial drug effect (assuming there was one) leads to a rebound increase in cholesterol so that the effect looks twice as large as it should. Perhaps the person is at a crucial age in which his/her cholesterol has begun a steady increase or decrease whether or not s/he took the drug. In short, the passage of time itself and factors associated with the passage of time makes the individual different under the two conditions (drug vs. no drug).

An alternative approach to obtaining a true counterfactual is to employ a "true experimental design." A true experiment involves at least two groups: a treatment group and a control group. In some cases, we include a placebo group instead of, or in addition to, the control group. A placebo group is a group that receives a fake treatment. First, we randomly select a sample from the population. Second, we randomly assign individuals to treatment and control (or placebo) groups. Third, we conduct pre-test measurement to establish baseline levels of the quantity of interest (like cholesterol level). Fourth, we provide the treatment to the treatment group and nothing to the control group (or a placebo to the placebo group). Fifth, after some passage of time, we remeasure the quantity of interest for both groups. The difference between the average change for the treatment group and the average change for the control/placebo group is the "causal" effect of the treatment.

Table 10.1 illustrates this process applied to two hypothetical individuals, i and j. Step 1—random sampling—is assumed. Step 2 is the process of random assignment (called randomization) to treatment and control groups. The second column shows the third step: the pre-test measurement for each individual at time t (O_{it} and O_{jt}). The third column shows the fourth step: the implementation of the treatment X to one, but not the other, individual. The fourth column shows the fifth step: the post-test measurement at time $t + 1$ for each individual. The final column shows the difference within individuals from time t to $t + 1$.

The bottom row of the table shows some important differences. The difference in the second column is the difference between the two individuals at baseline—the pre-test difference. The table shows that this difference is 0, and we will discuss why shortly. The difference in the bottom row of the fourth column is the post-treatment difference. Note that this is the difference *between* the two people and not a measure of change *within* the same person. Finally, the last column shows the within-individual change from pretest to posttest for both the treatment and control group, and the difference in these changes between the groups. This final difference is the causal effect of the treatment.

Under the counterfactual model, persons i and j would be the same person; however, as we discussed above, this is not possible: We cannot observe person i (aka j) both with and without the treatment. The best approximation to this ideal, instead, is to find two people who are *exactly* alike, give one the treatment and one the placebo (or no treatment), and observe the change for both as shown in the table. Obviously, it is impossible to find two identical people. However, if n is large enough and we *randomly assign* sample members to treatment and control groups, all factors that differentiate the members of the two groups will balance out, and pretest differences between treatment and control groups should be 0. This difference of 0 is for *all* characteristics, and not simply the quantity of interest. For example, suppose that our random sample consists of 50 men and 50 women. If we randomly select a person from this sample and assign him/her to the treatment group, then randomly select a second person and assign him/her to the control group, and continue this process until all sample members are assigned, we will end up with roughly 25 men and 25 women in each group. You can think of this process as being the same as the process of taking two independent random samples of some numeric characteristic from a single population: the two sample distributions should be similar to one another, in shape, mean, median, variance, etc. There will certainly be some slight difference, but if the samples are large enough, the differences will be minimal including on both characteristics you observe and characteristics you do not!

After the administration of the treatment, there are two sources of change from pre- to post-test in the treatment group: the passage of time and the implementation of the treatment. There is only one source of change for the control group: the passage of time. By subtracting the change for the control group from the change for the treatment group, we essentially eliminate the effect of time passage, and what is left is simply the effect of the treatment. Given that we have used a sample of individuals in both groups, the quantities in Table 10.1 should be means. Thus, we can place bars over the table quantities and change the subscripts to reflect groups rather than individuals. The result is that our experiment produces a quantity we can call the average treatment effect: $(\bar{\Delta}_T - \bar{\Delta}_C)$.

The process of randomization is fundamental to the validity of the true experimental design. Without randomization, we cannot completely rule out spuriousness, in part because we simply cannot measure all sources of it. A key problem is that, without randomization, we cannot know that the process of *selection* into a treatment group is ultimately responsible for differences we observe in some outcome between

treatment and control groups and not the treatment itself. By selection, I mean that differences in pre-test characteristics may be responsible for the assignment to the treatment versus control group.[3] For example, consider the relationship between education (a treatment) and earnings (an outcome). These two variables are strongly related in data (e.g., in the 2010 GSS, the correlation is .38 for working age people). Furthermore, education almost certainly precedes earnings: people tend to finish schooling before entering the labor market. Can we conclude that education causes earnings, and therefore recommend either to individuals or to policymakers that people obtain more schooling? Unfortunately, although common sense may tell us we can, in reality we cannot do so on the basis of the evidence. We have not randomly assigned education to individuals, and so there are numerous alternate explanations that may render the strong correlation spurious—in this context, such spuriousness is called selection bias. Those with greater intelligence may be more likely to select themselves into higher levels of schooling *and* into more lucrative ventures. Those with greater motivation may be more likely to continue schooling and more likely to seek promotions and job changes to increase earnings. Those whose parents have greater wealth may be better able to afford higher education, and their parents may also have better connections to help their children obtain lucrative jobs upon graduation. In terms of Table 10.1, the pre-test differences that were 0 under randomization are not guaranteed to be 0 without it.

The problem of lack of randomization applies to the relationship between *any* variables we may be interested in examining. Suppose a nutrition supplement company is seeking to determine whether a new supplement they produce helps with weight loss. So, they advertise for volunteers to try their product. They weigh volunteers before and after using the product for 1 month, and they observe that the average weight loss among the volunteers was 10 pounds. Can they conclude that their product works? Of course not. Those who volunteered are almost certainly more motivated to lose weight than those who did not volunteer, and they may therefore be more likely to engage in additional behaviors to lose weight during the treatment period. Thus, selection into the study (and therefore the treatment group)—and not the product itself—is quite likely the cause of the weight loss. This study design, even if it were to have random selection, suffers from another serious problem: there is no control (comparison) group. In this design, given that there was only one treatment group, random selection into it is equivalent to randomization, with all non-selected population members being implicit controls. The key assumption is that weight change among the general populace was 0 over the study period. But what if, over the course of the study, a natural disaster restricted food availability to everyone, so that everyone in the population lost 10 pounds over the month? If no measurement of weight change among controls was made, the

[3]This can happen either because the respondent selects him/herself into a treatment vs. control group or because the experimenter does. For example, suppose an experimenter assigns sicker patients to the control group in an experiment evaluating how well a new drug works.

effect of passage of time (or factors, like famine, associated with it) cannot be ruled out as an alternate explanation for why those in the treatment group lost weight.

Despite the importance of randomization, it is an unobtainable ideal in most social science research. Many of the treatments (causes) that we are interested in investigating simply cannot be randomly assigned to individuals. We cannot, for instance, assign individuals to be poor so that we can evaluate their long-term health. We cannot assign teen pregnancy. We cannot force some people to exercise while excluding others from doing so. Instead, social scientists generally rely on survey (or similar) data in which treatments are observed but not randomly assigned, and only outcomes (i.e., post-test measures) are measured. For example, the GSS measures education and income simultaneously. Education may be the treatment of interest, and income is only observed *after* the treatment: we usually do not know what income was before schooling was completed.

In most social science research, therefore, we turn to the second approach to ruling out spurious explanations in order to isolate the effect of treatments of interest: statistical methods. There are, in fact, a number of statistical methods that have been developed—and continue to be developed—that can help us rule out spuriousness when dealing with non-experimental data. We will focus here only on multiple regression. Many other methods for evaluating causality rely on the basics of multiple regression modeling—indeed, multiple regression is often a part of more advanced methods—and so an understanding of it is fundamental.

Before turning to multiple regression, however, it is important to note that, while true experiments are usually seen as the gold standard for making causal claims, they are not without their limitations. First, single, causal claims do not exist in a vacuum: reality is more complicated. As mentioned above, health in early life may influence education, which may then influence later health. Thus, education is an *intervening* variable (or "mediator") in a multi-step causal process. Yet, it is difficult to evaluate more than one treatment in an experiment.

Second, aside from the difficulty with handling multiple, intertwined causal relationships, it is difficult to see how one could even begin to address how early life health influences later life health in an experiment. Experiments are usually of short duration for ethical and practical reasons and cannot generally be used to evaluate causal processes that unfold over the long term.

Third, experiments are usually small in scale and involve unnatural conditions, and it is unclear whether causal effects identified in them can always scale-up. For example, suppose it were possible to learn via an experiment that obtaining more education, in fact, does increase earnings. Would this finding still hold if everyone in a society increased his/her years of schooling? As another example, in research on the effects of one's neighborhood on various outcomes, some experimental work has shown that moving to a better neighborhood produces improvement in outcomes. But, is it feasible for everyone in society to move to a better neighborhood?

These limitations of experiments, as well as others, suggest that, at least in some cases, using statistical methods may be a better—or at least more realistic—strategy

for answering research questions involving causal claims. Multiple regression is a key such method, and its strengths include that it can handle nonlinear relationships between variables, it can "control" out observed spurious threats to causality, and it can be used to attempt to disentangle direct and indirect causal relationships.

10.1.2 Statistical Control

In introducing multiple regression, let's consider one of the examples discussed in the previous section: the relationship between education and income. As mentioned above, the correlation between education and income is .38. Is this relationship causal? A respondent's current level of education almost certainly precedes his/her income, and the relationship between education and earnings is moderately strong. Yet, there may be alternate explanations for the relationship. For example, parental education may influence the respondent's educational attainment, and it may influence respondent's income for a variety of reasons, including that parents with greater education may be able to assist the respondent in choosing and obtaining a lucrative career path. Thus, the correlation between education and earnings may be spurious. In the context of Table 10.1, the implication is that pre-test differences are not 0, because respondents have not been randomized to education levels.

The correlations between parental education and respondent's education, as well as between parental education and respondent's income supports the view that respondent's educational attainment is not randomly assigned. The correlation between parent's education and respondent's education is .48, which is even larger than the correlation between respondent's education and income. Figure 10.1 illustrates the concern. In the figure, there is an arrow from parents' education to both respondent's education and respondent's income. The question is: once these two relationships are "controlled," is there any relationship remaining between respondent's education and respondent's income?

Multiple regression allows us to statistically control for parent's education to examine the remaining relationship between respondent's education and income. Before showing the process of control in multiple regression, however, we should discuss the meaning of "statistical control." Under randomization, all pre-test differences between the control and treatment groups are 0: random assignment ensures that there are no differences in any characteristic. In observational data, there is no guarantee, but statistical control is used to adjust pre-test differences so that the treatment and control groups are equal. A simple, albeit inefficient method of statistical control in the current example would be to examine the relationship between respondent's education and income among respondents with the same level of parental education. In other words, although we cannot assign individuals to the treatment and control groups (i.e., higher versus lower educational attainment) to balance the groups on pre-test characteristics, like parental education, we can hold

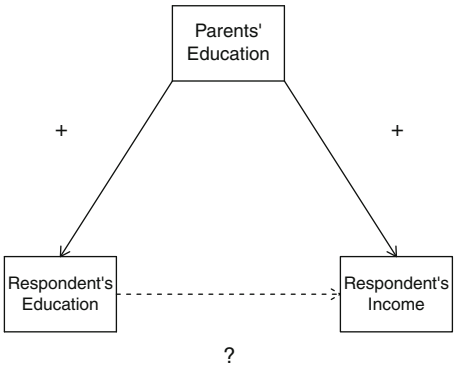


Fig. 10.1. Path model for the relationship between parents' education, respondent's education, and respondent's income.

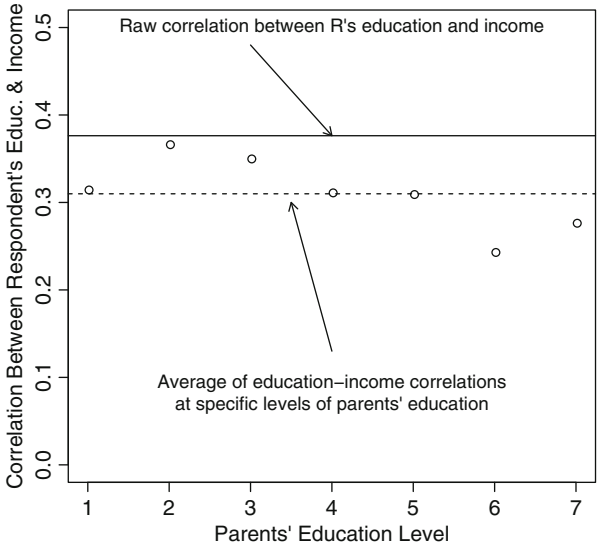


Fig. 10.2. Correlations between respondent's education and income for the sample (solid line), by level of parent's education (dots). Dashed line is the average of correlations at different levels of parent's education (levels of schooling discussed in text).

these characteristics constant by limiting our analyses to those with similar values on such characteristics. Put another way, we can manually balance respondents on parental education.

Figure 10.2 shows the results of following such a strategy. The solid horizontal line represents the correlation between respondent's education and income in the entire sample (.38). The dots are the correlations between respondent's education and income for persons with different levels of parental education. Parental education was coded into seven categories as follows: 1 for 0–7 years of schooling, 2 for

8 years of schooling, 3 for 9–11 years of schooling, 4 for 12 years of schooling, 5 for 13–15 years of schooling, 6 for 16 years of schooling, and 7 for more than 16 years of schooling. As the figure shows, while the raw (bivariate) correlation between education and income is .38, the correlation between education and income among persons with the same level of parental education tends to be much lower. The average correlation between education and income across all levels of parental education is .31. This correlation is 18 % smaller than the original correlation and illustrates that, once parents' education is controlled—that is, for respondents at comparable levels of parental education—the relationship between education and income is somewhat smaller than the raw correlation.

Why is the raw correlation between education and earnings larger than the correlations between education and earnings at each level of parental education? The answer is that respondents with lower education tend to have parents with lower education (and vice versa), and parental education influences the assignment of education level to the respondent *and* has a residual relationship with respondent's income. Ignoring parental education therefore consolidates all of the effect of parental education on respondent's income into respondent's education.

10.2 Multiple Regression Model, Estimation, and Hypothesis Testing

Multiple regression is more efficient at control than using the disaggregation method I just demonstrated. The multiple regression model extends the simple regression model to handle additional independent variables. The extended model is:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \ldots + \beta_k x_{ik} + e_i. \tag{10.1}$$

Notice that in this equation (compared to the simple linear regression model), α has been replaced by β_0, β is now more than one slope and is now subscripted, and the x variables are double-subscripted to reflect the measurement of multiple variables per individual. Notational differences aside, the multiple regression model is a straightforward extension of the simple regression model. Essentially, whereas in simple linear regression we estimated the *line* corresponding to the relationship between an x and a y, multiple regression extends the geometry to multidimensional *planes,* capturing the "tilt" of these planes in the different dimensions.

Estimation of this expanded model follows the same criteria as the simple regression model—finding the values of the parameters that minimize the sum of the squared errors—but the solution involves linear algebra and is:

$$\hat{\beta} = (X^T X)^{-1}(X^T Y). \tag{10.2}$$

Observe that I have replaced β with $\hat{\beta}$ (or b); I have done so because this is the *estimated* "vector" of slopes (or "regression coefficients"). The $(X^T X)^{-1}$ matrix (read: "x transpose x inverse") is akin to the denominator of the solution for β in the simple regression model ($\sum(x - \bar{x})^2$); the $(X^T Y)$ vector (read: "x transpose y") is akin to the numerator of the solution for β in the simple regression model ($\sum(x - \bar{x})(y - \bar{y})$). Just as in the simple regression model in which α and β are the best estimates of the population regression parameters, the estimates of β in the multiple regression are also the best estimates of the simultaneous relationships between each x variable, controlling on (or "net of") all other x variables in the model.

Given the size of most data sets in social science research, as well as the relative complexity and tediousness involved in computing the estimates of the regression coefficients, we will not dwell on the estimation of the model parameters and standard errors. In general, researchers use statistical software packages to perform the estimation. So, we will focus on hypothesis testing, understanding the model, and expanding the model's capabilities.

The ANOVA table for the multiple regression model looks exactly like the one from the simple regression model, with one exception: the degrees of freedom for the regression sums of squares is equal to the number of independent variables in the model (technically, it is the number of parameters, including the intercept, minus 1). The rest of the ANOVA table looks just as before, and so we will not repeat it here. In most tabular presentations of multiple regression model results, the only elements of the ANOVA table that are displayed include the model R-squared and possibly the model ANOVA F statistic.

The F test from the ANOVA table is now an omnibus test that tests whether *any* of the independent variables has a linear relationship with the outcome variable, and the model's R^2 reflects the proportion of the total variance in y that is explained by the linear combination of all the independent variables in the model. Ideally, if our model fits the data well, R^2 will be high, and the F test will be statistically significant. Technically, however, R^2 is a measure of both model fit and the extent of noise in the outcome: A model may fit very well, but y may simply have considerable noise built into it. So, a low R^2 should not necessarily be considered evidence that the model fits poorly.

Although the F test is important, we are usually interested in whether specific variables are related to the outcome. Thus, the main statistical tests that are of interest in the multiple regression model are t-tests on the parameters, just as with simple regression. As we discussed in the previous chapter, the logic of the test is as follows. If a particular x affects y, then we would expect the corresponding β to be non-zero. Put another way, if x and y are unrelated, then we would expect the best estimate for y to be $\bar{y}$, which does not depend on x. Thus, the slope (β) should be 0. The t-test to determine whether β is 0 can therefore be conducted for each parameter just as we did in the previous chapter for simple linear regression:

$$t = \frac{\hat{\beta} - 0}{SE(\hat{\beta})}. \tag{10.3}$$

Variable	Model 1	Model 2	Model 3
Intercept	−13.8(.75)***	8.7(.54)***	−16.1(.76)***
Education	3.1(.05)***		2.7(.06)***
Parental Educ.		1.7(.04)***	.67(.05)***
R^2	.14	.07	.15

$*p < .05, **p < .01, ***p < .001$

Table 10.2. Regression of income on respondent and parental education (GSS data 1972–2010, ages 30–64, $n = 20,409$; coefficients and (s.e.) presented)

This t-test is interpreted the same way as the t-tests we have used before: it provides us a measure of the probability we would observe the sample (regression coefficient) we did if, in reality, there is no relationship between x and y in the population. If the test value is large, the probability of observing the coefficient we did would be small under the assumption that $\beta = 0$ in the population. Thus, a significant t-test is usually considered evidence that x is related to y.

In order to illustrate the multiple regression model, I return to the example discussed previously involving parental education, respondent education, and respondent income. For this example, I use data from the 1972–2010 GSS. Only persons ages 35–64 with non-zero income were included. Education for both parents and respondents is measured in years of schooling (from 0 to 20). Parental education is measured as the maximum years of schooling for those with data on two parents. Finally, income is measured in $1,000s in real (2010) dollars.

Table 10.2 shows the results of three regression models using the data. In the first model, income was regressed on respondent's education (y is always regressed on x), and the parameter estimate (b_1) was 3.1, meaning that, on average, each year of schooling is associated with an increase of $3,100 in income. In the second model, income was regressed on parents' education, and the parameter estimate (b_1) was 1.7. That is, each year of parental education is associated, on average, with a $1,700 increase in income. Finally, Model 3 includes both independent variables. In this model, the parameter for respondent's education (b_1) is 2.7, and the parameter for parental education (b_2) is .67. Note that I have used the word "increase" although the model really simply tells us that those with more education have more income on average. The use of the word "increase" implies causality, and it implies within-individual change, neither of which we have demonstrated. This grammar usage is common, albeit not exacting.

The asterisks in the table indicate that the t-tests on all coefficients were statistically significant at the .001 level, meaning that there is a tiny probability that the estimated coefficients would be the magnitude they are if they were each 0 in the population. Substantively, education and parental education both appear to be related to respondent's income.

The results of Model 3 can be used to estimate (or predict) respondents' incomes based on their education level and the education level of their parents, following Eq. 10.1. For example, suppose we are interested in estimating income for a person with 12 years of schooling whose parental education is 16 years:

$$\hat{y}_i = b_0 + b_1 E_i + b_2 PE_i \tag{10.4}$$

$$= -16.1 + 2.7E_i + .67PE_i \tag{10.5}$$

$$= -16.1 + 2.7(12) + .67(16) \tag{10.6}$$

$$= 27.02 \tag{10.7}$$

Thus, persons with those values of education and parental education make, on average, about $27,000 in income.

In Model 3, both parameters are smaller than their counterparts in the first two models. Why? This change in magnitude illustrates the concept of control in the multiple regression model: approximately $1 - 2.7/3.1 = .13$ (13 %) of the relationship between respondent's education and income is due to parental education. Put another way, at comparable levels of parental education, there is a $2,700 difference in respondent's income on average. Compare the proportion of the relationship explained here with the reduction in the correlation between respondent's education and income shown in Fig. 10.2. The proportions are close, albeit not identical, but their similarity illustrate the notion of control: When a variable is controlled in multiple regression, it means that we are holding that variable constant to evaluate the relationship between other variables and the outcome of interest.

10.2.1 Total, Direct, and Indirect Association

We can see this idea of control in more detail if we define total, direct, and indirect association. Reconsider the path diagram shown in Fig. 10.1. In that diagram, we assumed that parents' education precedes both respondent's education and respondent's income, and the question is the extent to which respondent's education and income are related after controlling on parents' education. The regression coefficients shown in Table 10.2, with some adjustment, can help us evaluate the total association between education and income, and the indirect association of parental education with income *through* respondent's education.

The adjustment we need to make to the raw regression coefficients shown in the table is to *standardize* them. We have seen standardized variables before in previous chapters: the z score is a standardized score. It tells us the number of standard deviations a value of some variable is from its mean. In the regression context, a raw (i.e., unstandardized) coefficient tells us the expected change in y for a one-unit change in x. This change is measured in the raw metric of the variables involved, as discussed above. A standardized regression coefficient, in contrast, tells us the expected standard unit change in y per standard unit change in x. In other words, how much does a one z score unit change in x influence a z score change in y? The computation of a standardized regression coefficient is straightforward:

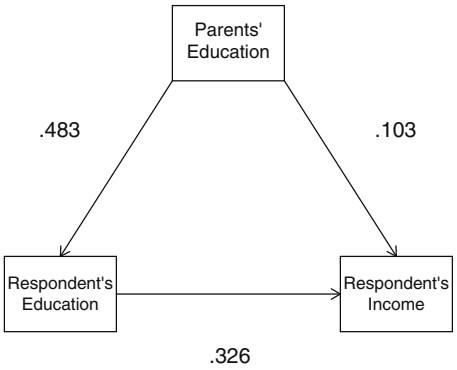


Fig. 10.3. Path model for the relationship between parents' education, respondent's education, and respondent's income with standardized coefficient estimates (GSS data from regression example)

$$\hat{\beta}_s = \left(\frac{s_x}{s_y}\right)\hat{\beta}. \tag{10.8}$$

In other words, we can convert to a standardized metric by simply multiplying the raw coefficient by the ratio of the standard deviation of x to the standard deviation of y.

The total association between x and y can be decomposed into the direct association between x and y and the indirect associations between x and y *through* other variables, using the following, simple identity:

$$Total = Direct + Indirect. \tag{10.9}$$

The indirect associations are simply the products of coefficients that comprise the pathways from x to y through other variables. For example, Fig. 10.3 revises Fig. 10.1 to include standardized regression coefficients.[4]

As the figure shows, the standardized association between parental education and income is .103, and the standardized association between respondent's education and income is .326. The standardized association between parental education and respondent's education was obtained via a regression model not shown here and was .483. Finally, two additional standardized estimates were obtained. One estimate was for Model 1 in Table 10.2: the total association between respondent's education and income was .376. The other estimate was for the total association between parental education and income: .261. That is, total associations between one variable and another are obtained from a simple linear regression model.

[4]Total, direct, and indirect effects do not need to be in a standardized metric. However, I use standardized coefficients here so that the relative magnitudes of the effects are more apparent.

Let's consider the relationship between parental education and respondent's income. The total association between parental education and income is .261, and as shown above, it can be decomposed as:

$$T = D + I \tag{10.10}$$

$$.261 = .103 + (.483)(.326). \tag{10.11}$$

Thus, of the total association between parental education and respondent's income, 61 % $(1 - .103/.261)$ of it is accounted for by respondent's education. Thus, respondent's education is called an *intervening variable* and is said to *mediate* much of the relationship between parental education and respondent's income. Put another way, respondent's education largely explains why parental education is strongly related to respondents' income: parents with high levels of schooling tend to produce children who obtain high levels of schooling and parlay it into higher incomes.

Now let's consider the total association between respondent's education and income:

$$T = D + I \tag{10.12}$$

$$.376 = .326 + (.483)(.103) \tag{10.13}$$

Here, about 13 % $(1-.326/.376)$ of the relationship between respondent's education and income is due to parental education. In other words, part of the relationship is spurious, but only a small part.

Notice that, although the path diagram has a directed arrow from parental education to respondent's education, the standardized coefficients are agnostic with respect to direction of causality. Thus, the decomposition of total association into direct and indirect is useful for either evaluating the extent to which a relationship between two variables is spurious *or* understanding mediating processes that explain why one variable has a relationship with another.

Throughout this section thus far, I have used the term "association" rather than "effect." The reason is that the methodology of decomposing total associations into direct and indirect associations is not in and of itself a method for establishing causal relationships. Nonetheless, when we construct path models with arrows delineating the direction we believe relationships "flow" between variables, we are implying causal pathways. Thus, we often use causal terminology in our discussion of such models. I will do so subsequently largely out of convenience, but keep in mind that the relationships I am discussing may not necessarily be truly causal ones.

10.2.2 Suppressor Relationships

While it is clearly important to rule out spuriousness, and thus include potential "upstream" variables in a regression model—that is, variables we think are causally prior to our variable of interest—is it important to include variables in the model

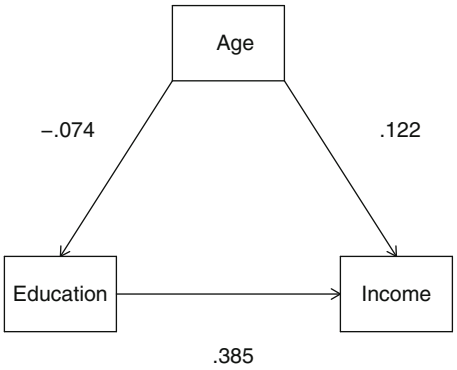


Fig. 10.4. Path model for the relationship between age, education, and income

that are "downstream," that is, intervening variables between the two variables of interest? For example, if we are really interested in the effect of parental education on income, is it necessary to control on respondent's education? In some sense, it may seem like it is unnecessary to do so, since respondent's education is an effect along the way from the ultimate cause of interest (we sometimes call such intervening variables "proximate causes," with ultimate causes called "distal causes"). However, the total effect of x on y can be misleading. In the current example, all the relationships between variables are positive: parental education is positively related to respondent's education and income, and respondent's education is positively related to income as well. Yet, this type of structure is not always present. Sometimes, a distal cause has a positive effect through one intervening variable, but a negative effect through another. Thus, the total effect may appear to be small or even zero.

We call this type of relationship a suppressor relationship, and ignoring it can produce misleading inferences. In order to illustrate suppression, consider the relationships between age, education and income as shown in Fig. 10.4. The figure contains the standardized coefficients needed to decompose the total effect of age on income into direct and indirect effects through education, obtained from the same GSS sample used in the previous example.

We might expect age to have both positive and negative indirect effects on income, in part because age represents a combination of factors. First, at the individual level, age represents the passage of time—maturation. With aging comes a number of advantages that should increase income, like increased work experience and increased savings and investments yielding higher returns, including simply interest.

At the same time, however, recall that, in cross-sectional data, we do not actually observe individuals aging: we observe differences between people at different ages at a single point in time. While such differences might include changes that occur with aging, they also include changes that differentiate *birth cohorts* from one another. One large change that has occurred across birth cohorts in the U.S. and other Western countries is increases in educational attainment. For example, in the

Variable	Model 1	Model 2	Model 3
Intercept	18.1(.81)***	−13.8(.75)***	−28.7(1.1)***
Age	.24(.02)***		.32(.02)***
Education		3.1(.05)***	3.2(.05)***
R^2	.01	.14	.16

$*p < .05, **p < .01, ***p < .001$

Table 10.3. Regression of income on respondent age and education (GSS data 1972–2010, ages 30–64, $n = 20,409$; coefficients and (s.e.) presented)

GSS data, the mean years of schooling for those between 30 and 40 in the 1970s was 13.1 years. For those between 30 and 40 in the 2000s, the mean was 14.1, a full year's difference. As we've seen, those with greater educational attainment tend to have higher incomes. Taken together, then, age is negatively related to education, producing a negative indirect effect of age on income through education:

$$T = D + I \tag{10.14}$$

$$= .122 + (-.074)(.385) \tag{10.15}$$

$$= .094. \tag{10.16}$$

Notice that the total effect of age is actually smaller than the direct effect: this result occurs because the total effect of age on income contains both negative and positive effects. Table 10.3 shows the same result in the raw metric of the variables. In Model 1, the effect of age is .24, meaning that each year of age is associated with an increase of $240 in income. Model 2 is the same as in the previous example. In Model 3, with both age and education included, the coefficient for age is .32, some 33 % (.32/.24) larger than in Model 1. In short, even though education is a mediator of the relationship between age and income, it was important to include education in the model, because the total effect was suppressed by age's negative relationship with education.

Also realize again, that, if we were primarily interested in the relationship between education and income, it would have been important to control on age because education's effect was also suppressed by age: once age is controlled, the relationship between education and income appears larger. The reason is that age increases income, perhaps due to work experience, but older cohorts have less education, making education's total effect seem smaller.

Once again, despite the fact that we have spent much of the last few pages discussing "effects" of age, parental education, and respondent's education on income, can we really claim that we have established that these relationships are causal? The use of the word "effect" implies that we are assessing causal relationships, but such is the limitation of the English language: it is much more efficient to use this terminology than a more accurate one. Yet there are several reasons why we cannot conclude that the "effects" we have found are the true causal effects.

First, in each example, we have only controlled on one variable prior to the one of interest. If our interest is in the causal effect of education on incomes, in each example we have only controlled on one prior variable. In the first example, we only controlled on parental education; in the second, we only controlled on age. Yet we have shown in these two examples that both age and parental education predict respondent's education, so both should be controlled simultaneously. Indeed, in reality, there are a number of variables that should be controlled in order to meet the goal shown in Table 10.1 of making pre-test differences between the "treatment" and "control" groups (those with more vs. less education) zero. Sex and race both certainly precede respondent's education, and both certainly influence income through mechanisms other than educational attainment. For example, there are notable differences in the *type* of education men and women choose (i.e., career paths) that influence income beyond the simple number of years of schooling each gender obtains. Men are more likely to obtain 4-year college degrees in engineering than women, and women are more likely to obtain 4-years degrees in education. These professions, while requiring the same number of years of schooling, pay substantially differently. Thus, sex and race should be controlled. Countless other variables should be controlled as well, like region, religious affiliation, and others. In other words, our simple, two-variable multiple regression models are insufficient. A casual examination of social science articles, in fact, shows that most regression models include at least half a dozen controls. Unfortunately, we can *never* control on every possible spurious threat, because many variables are unobserved: surveys rarely measure enough variables to do so. Furthermore, many spurious threats are *unobservable,* like motivation. Thus, regression modeling by itself cannot accomplish what randomization can, and we should always be hesitant to make causal claims from such models.

Second, the multiple regression models I have shown so far assume the relationships between all variables in the model are linear. In many cases, this assumption may be reasonable, but in probably many more cases, it isn't. For example, the relationship between age and income is certainly not linear across all of adulthood. Although I have restricted the data to persons of working ages (30–64) in the preceding examples, it is not clear that the relationship between age and income is linear even in that age range. At some age, there is a tradeoff between what experience adds to a worker's value to an employer vs. what knowledge of new technology adds. Thus, we might expect that income increases across age to a point, but then potentially stagnates or even decreases after. Determining the true causal effect of age, then, requires that we be able to model nonlinear relationships between the purported cause and effect.

Third, in addition to the assumption that all relationships are linear in the multiple regression models presented thus far, we have also assumed that all relationships are *additive*. That is, for example, we have assumed that the relationship between age and income is the same regardless of the respondent's level of education. It is possible that the relationship between age and income is not the same across all levels of education. The relationship between age and income may follow one pattern for those with PhDs, say, while following another pattern altogether for those

without high school diplomas. This possibility necessitates that we be able to model *interactions* between independent variables.

In the next three sections, I discuss how we can incorporate variables like sex and race, which are fundamentally non-numeric variables, into the multiple regression model; how we can incorporate nonlinear relationships into the model; and how we can incorporate interactive relationships into the model.

10.3 Expanding the Model's Capabilities

The multiple regression model would be of relatively little use in social science if it were limited to modeling linear relationships between continuous, numeric variables only. Many, if not most of the variables we use in social science research are not numeric, and the relationships between variables that our theories specify are often nonlinear. Fortunately, the multiple regression model is quite flexible and can accomodate nonlinear relationships and variables measured other than continuously. The one limitation that remains is that the outcome variable y must be continuous (and $y|x$ must be normally distributed; put another way $e \sim N(0, \sigma_e^2)$). When this assumption is violated, alternate models are needed (but discussing them is beyond the scope of this book).

10.3.1 Including Non-continuous Variables

The most useful extension of the regression model is the ability to incorporate noncontinuous covariates (predictors, x). If we are interested in race, sex, or other group differences in some y, we can incorporate indicator variables—also called "dummy variables"—into the model. For example, suppose we were interested in examining race differences in income in the GSS data. Race is measured at the nominal level (1 = white; 2 = black; 3 = other), and therefore the numeric codes assigned to racial categories in the data are meaningless. However, for regression analyses, we can construct a pair of variables indicating whether a respondent is black (=1) (or not = 0) and whether a respondent is an "other" race (=1) (or not = 0), and we can include those dummy variables in the regression model as predictors:

$$Income_i = \beta_0 + \beta_1 \times Black_i + \beta_2 \times Other_i + e_i. \tag{10.17}$$

If we estimate this model, we obtain:

$$E(Income) = 29.21 - 5.52 \times Black + 2.91 \times Other. \tag{10.18}$$

The interpretation of these coefficients is straightforward. If we were interested in the predicted value of family income for blacks, we simply insert Black = 1 and Other = 0 and compute the expected value:

Variable	Model 1	Model 2
Intercept	29.21(.19)***	−13.25(.76)**
Education		3.08(.05)***
Black	−5.52(.52)***	−3.12(.48)***
Other	2.91(.80)**	3.96(.74)**
R^2	.01	.14

* $p < .05$, ** $p < .01$, *** $p < .001$

Table 10.4. Regression of income on race and education (GSS data 1972–2010, ages 30–64, $n = 20{,}409$; coefficients and (s.e.) presented)

$$E(Income) = 29.21 - 5.52 \times (1) 2.91 \times (0) = 23.69. \tag{10.19}$$

If we are interested in the predicted value of income for persons of other races, we insert Black = 0 and Other = 1 and compute the expected value:

$$E(Income) = 29.21 - 5.52 \times (0) 2.91 \times (1) = 32.12 \tag{10.20}$$

Finally, if we are interested in the predicted value of income for whites, we insert Black = 0 and Other = 0. Doing so will leave us with only the intercept—the expected value of income for whites is 29.21. This latter finding shows that we only need to include $k - 1$ dummy variables for a variable that initially had k categories: The model intercept represents the value when all the other dummy variables are 0. The omitted group is therefore called the "reference" group, because the coefficients for the dummy variables represent how much the mean for the group represented by the dummy variable differs from the mean for the reference group. For example, in the example above, blacks have, on average, $5,520 less income *than whites*, while others have $2,910 more income *than whites.*

A couple of notes are in order regarding the use of only dummy variables in a regression model. First, the t statistic that you obtain for the regression coefficient will be identical to the t statistic you obtain when conducting a simple independent samples t test for the two groups (assuming you only have one dummy variable in the model). Second, one of the standard parts of the output of regression model software is an ANOVA table of the regression results. This ANOVA table will be identical to that which would be obtained if you performed the equivalent ANOVA. Indeed, with only dummy variables in the model, the coefficients for the dummy variables simply reproduce the mean of y for each group.

We can include both dummy variables and continuous variables in our models. Extending the example above, if we include education as a predictor of income along with race, we obtain the results shown in Table 10.4. We used the coefficients from Model 1 above in showing the interpretation of the dummy variable coefficients. Model 2 shows that education increases income (each additional year of schooling increases expected income by $3,080), and that, once education is controlled, the coefficients for the race dummy variables change. The dummy variable coefficient for blacks decreases in magnitude by $1 - 3.12/5.52 = 43\,\%$, while the dummy

variable coefficient for those of other races *increases* in magnitude by $3.96/2.91 = 36\,\%$. Substantively, we can use the concepts of direct, indirect, and total association to discuss the implications of these changing coefficients. For blacks, the total association from Model 1 was negative: blacks make less income than whites. In Model 2, once education is controlled, this negative association decreases (shrinks) toward 0. Given that education has a positive relationship with income, the relationship between the black dummy variable and education must be negative—i.e., blacks have less education than whites on average—in order for the direct association to be reduced from a larger to a smaller negative number. Thus, if blacks had comparable education to whites, their incomes would be higher, and the income gap would be smaller. Put another way, 43 % of the black-white difference in income is due to education differences between these racial groups, with blacks having less education than whites on average.

The change in the coefficient for persons of other races is slightly more difficult to consider, because the coefficient increases away from zero once education is controlled. Again, we know that the relationship between education and income is positive; thus, the relationship between the other race dummy variable and education must be negative, in order for the direct association to increase. Here, education differences between whites and those of other races, with whites having more education on average, are suppressing the "other" advantage in income. In other words, if persons of other races had average educational attainment comparable to whites, the income difference between whites and others would be 36 % larger than it already is.

10.3.2 Statistical Interactions

In the discussion regarding the inclusion of dummy variables, we implicitly assumed that the relationship between education and income was the same for each race. Specifically, consider the prediction equation from Model 2 of Table 10.4:

$$E(income) = -13.25 + 3.08 \times Education - 3.12 \times Black + 3.96 \times Other \tag{10.21}$$

For each race, the prediction equation reduces to:

$$E(income)_W = -13.25 + 3.08 \times Education \tag{10.22}$$

$$E(income)_B = (-13.25 - 3.12) + 3.08 \times Education \tag{10.23}$$

$$E(income)_O = (-13.25 + 3.96) + 3.08 \times Education. \tag{10.24}$$

As these equations show, there are race differences in the expected value of income at each level of education—the intercept—but the return for each year of education (slope) is the same—\$3,080—for each race. The result is three parallel prediction lines.

Variable	Model 1	Model 2
Intercept	−13.33(.79)**	−13.57(.84)***
Education	3.09(.06)***	3.10(.06)***
Black	−3.11(.48)***	−1.22(2.27)
Black*Education		−.14(.17)
R^2	.14	.14

* $p < .05$, ** $p < .01$, *** $p < .001$

Table 10.5. Regression of income on race and education (GSS data 1972–2010, ages 30–64, $n = 19{,}440$; coefficients and (s.e.) presented)

Often, the assumption of parallel prediction lines across subgroups in a population is an unreasonable one, or it may be an assumption our theory/hypothesis challenges. For example, with regard to race, education, and income, one may argue that blacks are discriminated against in the workforce so that each year of education does not produce comparable income returns compared to whites. In order to capture differences in the effect of one variable across subgroups of the population, we can construct "statistical interactions" and include them in our model. In this example, constructing an interaction to capture differential returns to education for blacks involves creating a new variable that is the product of the black dummy variable and education. To simplify matters, let's examine only blacks and whites and create an interaction between black and education. The prediction equation becomes:

$$E(income) = b_0 + b_1 education + b_2 black + b_3(black * education). \quad (10.25)$$

For whites, then, the prediction equation is:

$$E(income) = b_0 + b_1 education, \quad (10.26)$$

while, for blacks, the prediction equation is (after setting the "black" dummy variable to 1):

$$E(income) = (b_0 + b_2) + (b_1 + b_3) education. \quad (10.27)$$

This result shows that whites and blacks differ in both intercept and slope for the effect of education. Table 10.5 shows the results of two models: one without a statistical interaction and one with the interaction.

Given the results shown in the table, we have the following empirical prediction equations:

$$E(income)_W = -13.57 + 3.10 \times education \quad (10.28)$$

$$E(income)_B = (-13.57 - 1.22) + (3.10 - .14) \times education \quad (10.29)$$

$$= -14.79 + 2.96 \times education \quad (10.30)$$

$$(10.31)$$

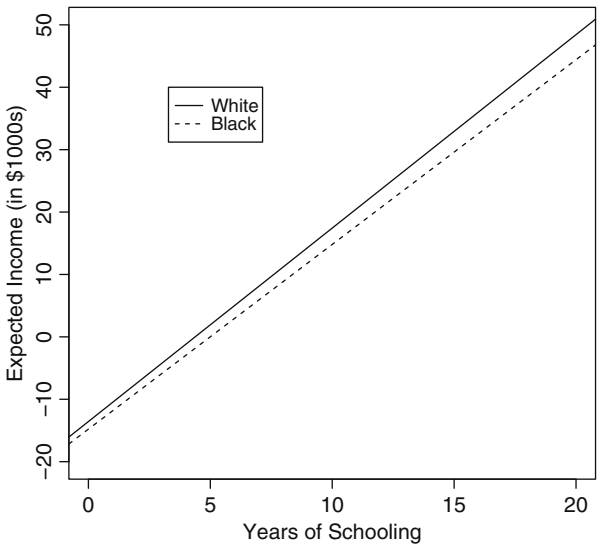


Fig. 10.5. Expected income by years of schooling and race (GSS data 1972–2010, ages 30–64)

From these results, it appears not only that blacks have lower incomes than whites at all levels of education, but also that each additional year of schooling nets less additional income for blacks relative to whites as well. Figure 10.5 displays these prediction equations graphically. As the figure shows, the prediction lines for blacks and whites are relatively close together at the lowest level of education (0 years), and the gap between the two racial groups expands across years of schooling. However, as the table indicates, in Model 2, neither the black dummy variable coefficient nor the interaction term is statistically significant. Thus, the main effects only model (i.e., the model without the interaction effect) is the better model for these data.

Interactions are not limited to two variables, nor are they restricted to a dummy and a continuous variable. However, interactions between continuous variables and three-way or higher order interactions are complex in interpretation and are beyond the scope of this book.

10.3.3 Modeling Nonlinear Relationships

Sometimes we are interested in modeling nonlinear relationships between variables. For example, in our last model predicting income as a function of race and education, we found that predicted income for persons with less than about 5 years of schooling was *negative.* Negative incomes do not occur with any regularity in reality, and so a model that predicts negative values of income may be unrealistic. Figure 10.6 shows the actual pattern for mean income by years of schooling in the

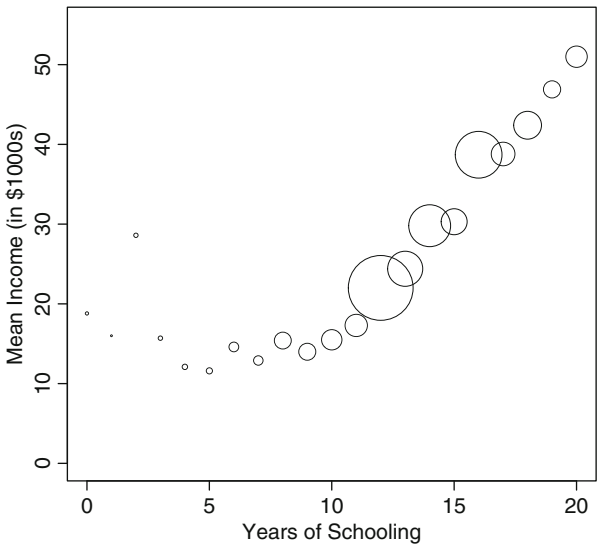


Fig. 10.6. Mean income by years of schooling with values weighted by sample size (GSS data 1972–2010, ages 30–64)

GSS data we have been using. As the figure shows, the relationship between years of schooling and income appears linear from about 8 years of schooling upward, but the relationship seems to bottom-out and perhaps even curve upward as education decreases from 8 to 0 years. Note that the plotting characters in the figure have been adjusted to reflect the number of observations at each level of schooling. As the figure shows, there are relatively few person in the sample at the lowest levels of schooling, and this explains why the model estimates a line that falls below 0 around 5 years of schooling: the estimates are driven by the much larger proportion of the sample at higher levels of schooling.

How can we capture this inverted curvilinear, or even u-shape pattern in a linear regression model? There are a variety of ways to capture nonlinearity, including applying nonlinear transformations to y variables like the logarithm. Indeed, the logarithmic transformation is a common one for income for several reasons, including that it reduces the right skew of the distribution. We will not illustrate the log or other transformations here; instead, we will focus on one particular class of models for capturing nonlinearity: using polynomial terms for x. This approach is called "polynomial regression."

Recall from algebra (and Chap. 5) that a general equation for a parabola is:

$$y = a(x - h)^2 + k, \tag{10.32}$$

where (h, k) is the vertex, and a determines the breadth/curvature and direction of the opening of the parabola. If a is positive, the parabola is u-shaped; if a is negative, the parabola is inverted. If a is small, the parabola is wide; if a is large, the parabola is narrow.

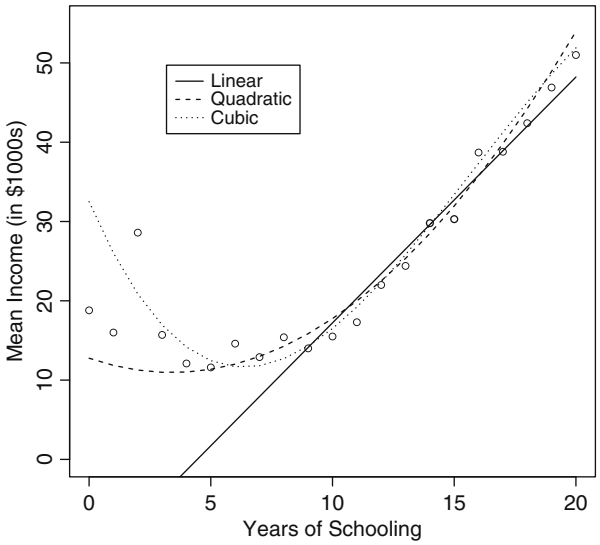


Fig. 10.7. Mean income by education with linear, quadratic, and cubic fits superimposed (GSS data 1972–2010)

If we expand the quadratic term in this formula, we obtain:

$$y = ax^2 - 2axh + ah^2 + k, \tag{10.33}$$

and if we then let $b_0 = ah^2 + k$, let $b_1 = -2ah$, and let $b_2 = a$, we obtain:

$$y = b_0 + b_1 x + b_2 x^2. \tag{10.34}$$

This result suggests that, if we include a second variable in our regression model—namely, x^2—we can capture a parabolic shape with the linear regression model. x^2 is constructed simply by taking the original x variable, squaring it, and saving this quantity as a new variable that we then include in the multiple regression model.

This result can be extended to higher order polynomials in order to capture more complex nonlinearities. Figure 10.7 shows the income-by-education data again with the linear fit, a quadratic (parabolic; second degree polynomial) fit, and a cubic (third degree polynomial—x^3) fit. The quadratic fit is better than the linear fit, but it fails to capture the higher incomes at the lowest levels of schooling, and it overestimates incomes at the very highest levels of education. The cubic fit appears to be the best. At the same time, however, the interpretation of the cubic model, which involves three terms for education—x, x^2, and x^3—is difficult. Consequently, the quadratic model is probably preferable, especially given that the sample size at the lowest levels of schooling is quite small, thus making the sample means for income highly variable at these levels.

10.4 Conclusions

In this chapter, we began by discussing the requirements that must be met in order to establish that a relationship between two variables is causal. As we saw, experimental methods are the gold standard for establishing causality, because randomization of sample members to treatment and control groups allows us to rule out spuriousness, that is, alternative explanations for the relationship between the variables of interest. However, social science research generally involves treatments that cannot be randomly assigned to respondents. Thus, social scientists tend to turn to multiple regression methods. We showed the basic extension of the simple regression model to handle multiple independent variables. We then extended the multiple regression model to handle noncontinuous independent variables, statistical interactions between independent variables, and nonlinear relationships between x and y variables. Together, these extensions of the model make the multiple regression model highly flexible and therefore extremely useful for statistical analysis of social science data. Indeed, because of its flexibility, the linear regression model, and further extensions of it, are widely used in research.

10.5 Items for Review

- Causality rules
- Spuriousness
- Counterfactual model
- True experiment
- Treatment
- Control
- Placebo
- Randomization
- Statistical control
- Interpreting coefficients in multiple regression
- F test in multiple regression
- t tests on parameters
- Total, direct, and indirect effects
- Suppressor effects
- Dummy variables
- Statistical interaction
- Capturing nonlinearity in regression

10.6 Homework

1. The following table presents three models for the relationship between education and income using GSS data only up to 2006. For all three, plot the prediction curve.

Variable	Coefficient		
	Model 1	Model 2	Model 3
Intercept	−2.99***	8.50***	24.95***
Education	4.10***	2.11***	−3.53***
Education^2		.08***	.63***
Education^3			−.016***

Regression of income on polynomials for education, 1972–2006 GSS data, n = 26,228. (*** $p < .001$)

2. Below is a table with results of a single regression model predicting income. The model contains a statistical interaction between sex and education. Plot the implied regression lines across education for men and women.

Variable	Coefficient
Intercept	−28.34***
Male	23.43**
Education	5.62***
Male*Educ.	−1.16*

Regression of income on sex and education, 2000 GSS data, n = 1,667 (* $p < .05$, ** $p < .01$, *** $p < .001$)

3. Below is a regression model predicting income. Birth cohort is constructed as year of survey minus age of respondent. Male, married, and lives in south are dummy variables with references as female, not married, and lives in other regions. Education is in years of schooling, and health is measured so that a higher score indicates better health. Interpret the results.

Variable	Coefficient
Intercept	−27.96***
Age	.09***
Birth cohort	.04*
Male	5.3***
Education	3.6***
Married	21.7***
Lives in South	−.74
Health	5.27***
R^2	.29

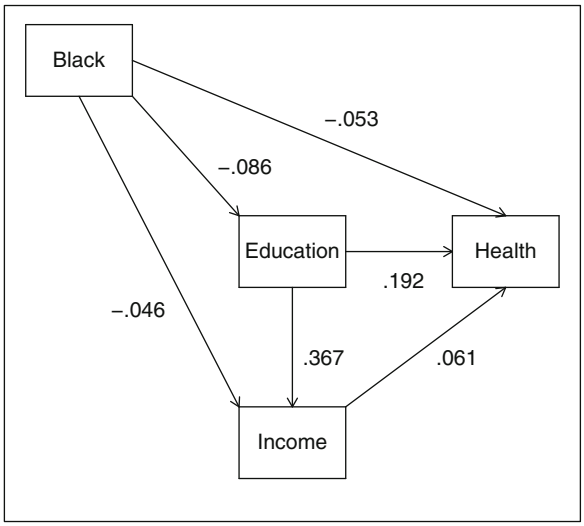
Regression of income on selected covariates, 1972–2006 GSS data, n = 26,228 (* $p < .05$, ** $p < .01$, *** $p < .001$)

4. The GSS asks a question each year about political party affiliation, with 0 being a "strong Democrat" and 6 being a "strong Republican." Values in between reflect less extremity, with 3 being moderate, independent, or unsure (as per my coding of response categories). I estimated a series of regression models as shown in the table below, for persons ages 30–64, from 1972 to 2010 to examine the pattern of party preference over time for males and females. Plot the prediction curves for each gender across year for each model and interpret. How different is the interpretation for Model 1 vs. Model 4?

Variable	Model 1	Model 2	Model 3	Model4
Intercept	1.21***	−6.04***	27.09*	29.92**
	(.13)	(1.15)	(11.01)	(11.01)
Male	.34***	.34***	.34***	−.967***
	(.03)	(.03)	(.03)	(.25)
Year	.015***	.174***	−.923*	−.997**
	(.001)	(.023)	(.36)	(.36)
Year^2		−.00086***	.0111**	.0118**
		(.0001)	(.004)	(.004)
Year^3			−.000044**	−.000046**
			(.00001)	(.00001)
Male*Year				.0141***
				(.003)
R^2	.012	.014	.015	.016

Regression of party affiliation on selected covariates, 1972–2010 GSS data, ages 30–64, n = 20,409 (* $p < .05$, ** $p < .01$, *** $p < .001$)

5. Black and white differences in health are widely studied in social science research. It is well known that whites report better health on average than blacks, and whites have longer life expectancies, as well. A key question is why. Below is a path diagram with standardized regression coefficients for direct and indirect paths from race to health. Based on the diagram, what is the total effect of race on health? What proportion of the total effect is explained by education and income differences between blacks and whites? (hint: there are THREE indirect paths from black to health).



Path diagram for direct and indirect effects of race on health (1972–2010 GSS data, ages 30–64, n = 14,149, persons of "other" races excluded).

6. Suppose I am interested in determining why some people are Republicans, while others are Democrats. So, I regress the party affiliation variable described above on a number of demographic and social variables and obtain the results shown in the table below. I have a friend with the following characteristics: He was born in 1971, he is white and lived in the south when he was 16. His parents' education level is 20 years, while his own education level is 19. He is married, currently lives in the south, he claims excellent health, and he makes $150,000 per year in income. Compute his predicted score on the party affiliation variable and speculate about his affiliation.
   Note: cohort is computed as year (as a three digit number, with 2000 being 100) minus age (so, if cohort = 71, current age in 2013 would be 42). Male, Black, Other, South at 16, Married, and South are all dummy variables. Health is rated on a 4 point scale coded as: 0 = Poor, 1 = Fair, 2 = Good, 3 = Excellent. Income is in $1,000 units. The outcome variable, as described above, ranges from 0 (strong Democrat) to 6 (strong Republican). The central value on the scale—3—is not affiliated with either party.

Variable	Coefficient
Intercept	.72(.20)***
Cohort	.014(.002)***
Age	.014(.002)***
Male	.200(.03)***
Black	−1.58(.05)***
Other	−.53(.07)***
South at 16	−.09(.05)#
Parents' Education	.04(.005)***
South	.17(.05)***
Married	.28(.03)***
Education	−.006(.006)
Health	.07(.02)**
Income	.002(.001)*
R^2	.11

Regression of party affiliation on selected covariates, 1972–2010 GSS data, ages 30–64, n = 14,835 (* $p < .05$, ** $p < .01$, *** $p < .001$)