Workers Party, Libertarian Party, etc.). None of the three dummies has a statistically significant coefficient. Is it reasonable to conclude that political affiliation doesn't affect contributions? Is there a better way to set up the dummy variables?

7. For a sample of 474 employees, current salary was regressed on

   Minority Classification: 1 = minority, 0 = nonminority

   Educational Level: years of schooling

   EDMIN: Minority Classification × Years of Schooling

Box 8.1 shows the results of the regression, as reported by SPSS. How would you interpret these results?

BOX 8.1. Coefficients[a]

Model	Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1 (Constant)	-23028.8	3077.537		-7.483	.000
Minority Classification	31477.287	6953.429	.764	4.527	.000
Educational Level (years)	4312.077	219.726	.729	19.625	.000
EDMIN	-2725.021	526.904	-.865	-5.172	.000

a. Dependent variable: Current salary

9 How Is Multiple Regression Related to Other Statistical Techniques?

Multiple regression is closely related to many other statistical methods that are in common use today. In fact, several methods are exactly equivalent to multiple regression but go by different names. Other methods involve extensions of multiple regression to more complicated situations. Finally, there are several classes of nonlinear models that have the same general aim as multiple regression but make allowance for special features of the dependent variable. The goal of this chapter is to give you a quick overview of these methods and to explain how they are related to multiple regression. Obviously, I can't provide enough information here to enable you to actually use any of these methods.

9.1. How Is Regression Related to Correlation?

I already answered this question in Chapter 5, but a brief review may be helpful here. Both regression and correlation assume that the relationship between the dependent and independent variables can be described by a straight line. Regression focuses on the slope of the line, whereas correlation focuses on the degree to which the data points are scattered around the line. In the bivariate case, testing whether the regression slope is equal to zero is equivalent to testing whether the correlation is equal to zero. When there are multiple independent variables, testing whether a regression slope is zero is equivalent to testing whether the corresponding *partial* correlation is zero. Finally, the R^2 for a regression is equal to the squared correlation between the dependent variable and the predicted value of the dependent variable, based on the estimated regression model.

9.2. How Is Regression Related to the *t* Test for a Difference Between Means?

Suppose you have a sample with both men and women, and you want to test whether the average income for men is greater than the average income for women. The conventional way of doing this is to compute a *t* statistic for the difference between the means (using a pooled variance estimate). Alternatively, you could do a regression of income on a dummy variable for sex. The *t* statistic for the coefficient of this dummy variable will be *identical* to that for the conventional *t* test. The advantage of doing it with regression is that you can control for other variables by putting them in the equation.

9.3. What Is Analysis of Variance?

Analysis of variance (ANOVA) is a method that is widely used to analyze data from experiments with complex treatment designs (Iversen & Norpoth, 1987). In its simplest form, one-way ANOVA is used to test whether the means of the outcome variable are different across two or more groups. When there are only two groups, ANOVA is equivalent to a conventional *t* test for a difference in means and, therefore, it's also equivalent to a multiple regression with a single, independent dummy variable. When there are more than two groups, one-way ANOVA is exactly equivalent to a multiple regression with dummy variables for all but one of the groups. Even for multifactor designs, there is virtually always a multiple regression model that is exactly equivalent to the ANOVA model. Many statistical packages don't even have separate ANOVA and multiple regression programs anymore. Instead, they have a single procedure that estimates the so-called *general linear model*.

9.4. What Is Analysis of Covariance?

Analysis of covariance (ANCOVA) is an extension of ANOVA to allow for situations where some of the independent variables are categorical (as with experimental treatments) and some are measured on quantitative scales, often called covariates (Wildt & Ahtola,

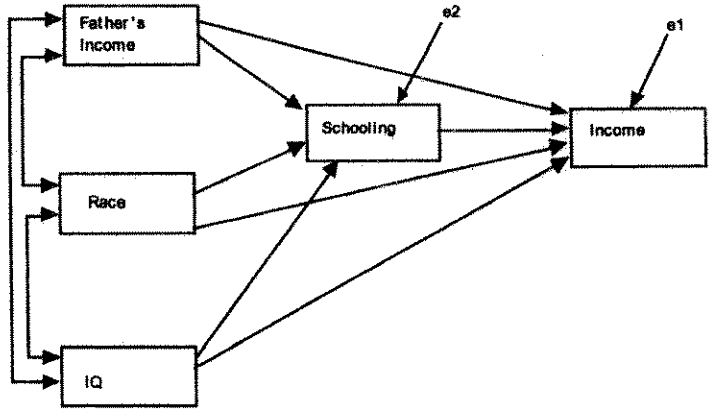

Figure 9.1. Path Diagram for Regressions of Income and Schooling

1978). Of course, we already know how to do that with multiple regression: Just put the quantitative variables in the regression and include a set of dummy variables to represent each of the categorical variables. As you might expect from the last two questions, the multiple regression model is exactly equivalent to the corresponding ANCOVA model.

9.5. What Is Path Analysis?

Path analysis is a technique that became extremely popular in sociology in the late 1960s and then spread to other social sciences (Duncan, 1975). Essentially, it is a way of representing a set of regression equations with "causal" diagrams. Its popularity stemmed from the fact that the diagrams made it easy to understand a complex set of relationships that might otherwise be obscure if you looked only at the equations.

Figure 9.1 shows a path diagram for the dependence of income and schooling on father's income, race, and IQ. This diagram corresponds to two regression equations:

$$\text{Schooling} = a_0 + a_1\text{Race} + a_2\text{IQ} + a_3\text{Father's Income} + e_2$$

$$\text{Income} = b_0 + b_1\text{Schooling} + b_2\text{Race} + b_3\text{IQ} + b_4\text{Father's Income} + e_1.$$

In a path analysis, the *a* and *b* coefficients can be estimated separately for each equation by ordinary least squares.

In the diagram, each of the single-headed arrows represents a causal effect of one variable on another. The double-headed arrows represent correlations that don't involve any presumption of causality. It is common (but not essential) to put numbers on the single-headed arrows. Usually these are standardized regression coefficients, but they may also be unstandardized coefficients.

Although path analysis doesn't really involve anything new with regard to statistical estimation, it can be very helpful in the analysis of direct and indirect effects, an issue that we discussed in Chapter 3. In particular, path analysis has some easy to use rules for calculating indirect effects by tracing along pathways in the diagrams.

9.6. What Are Simultaneous Equation Models?

The path model that we just examined is an example of a more general class of linear models called simultaneous equation models. The essence of such models is that there is more than one equation, and all equations are simultaneously true.

Recursive simultaneous equation models, like the one in Figure 9.1, can be estimated by ordinary least squares applied to each equation separately. A model is recursive if the causal relationships flow in only one direction. In terms of path diagrams, that means that if you follow the single-headed arrows, you can't get back to where you started from.

Figure 9.2 shows a simple *non*recursive system. In this model, occupational aspirations affect educational aspirations but, at the same time, educational aspirations affect occupational aspirations. The two equations in this model *must* be estimated simultaneously, and ordinary least squares will not do the job. The best known methods for estimating nonrecursive models are two-stage least squares, three-stage least squares, and maximum likelihood, but there are many others. Unfortunately, the validity of any of these methods rests on assumptions that may be difficult to justify. For the model in Figure 9.2, for example, it is necessary to assume that father's income does *not* directly affect educational aspirations, and that father's education does *not* directly affect occupational aspirations.

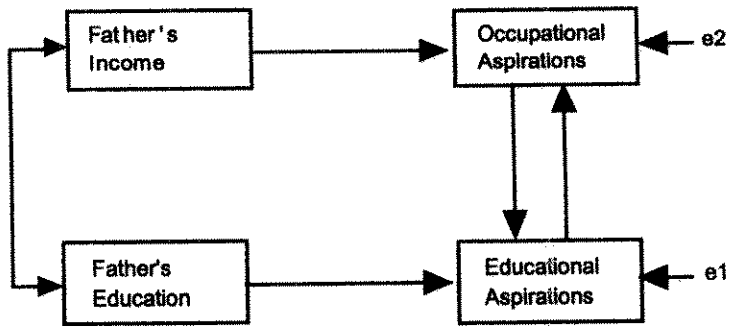

Figure 9.2. Path Diagram of a Nonrecursive Model

9.7. What Is Factor Analysis?

Factor analysis is designed to take a bunch of variables and reduce them to a much smaller number of dimensions (Kim & Mueller, 1978a, 1978b). Suppose, for example, that you have an interview schedule that contains 40 questions about attitudes toward illegal drugs. If you did a factor analysis of these questions, you might find that the responses to these questions could be accounted for by three underlying dimensions: parental concern, liberal versus conservative, and fear of crime. For each observed variable, you would get an estimate of how strongly it depends on one or more of these underlying dimensions.

The model that forms the basis of factor analysis is a linear, simultaneous equations model. Each observed variable is the dependent variable in a single equation. The independent variables are all latent, unobserved variables called *factors*. For four observed variables (y_1 through y_4) and a single factor F, the factor model can be written as

$$y_1 = a_1F + e_1$$

$$y_2 = a_2F + e_2$$

$$y_3 = a_3F + e_3$$

$$y_4 = a_4F + e_4$$

The *a*s are like ordinary regression coefficients, but they are commonly referred to as *factor loadings*. Like all factor models, this one

says that whatever correlations we find among the *y*s can be completely explained by their mutual dependence on *F*.

Because the independent variables in the linear equations are not directly measured, ordinary least squares cannot be used to estimate the factor loadings. Many different methods are available, but they generally fall into two classes, *exploratory factor analysis* and *confirmatory factor analysis*. In confirmatory factor analysis, the analyst must specify exactly how many factors there are in the model, and which observed variables depend on which factors. In exploratory factor analysis, these decisions are left to the computer algorithm. Although that may seem like a big advantage, the problem with exploratory factor analysis is that different algorithms may come up with very different results. Nowadays, confirmatory factor analysis is generally preferred, at least in those situations where you can make some reasonable guesses about what the model should look like.

9.8. What Are Structural Equation Models?

Structural equation models combine confirmatory factor analysis with simultaneous equations models (Bollen, 1989; Hayduk, 1988). The basic idea is that the observed variables depend on latent, unobserved variables, and those latent variables may have causal relationships among them. Because the models are often rather complicated, they are frequently presented in the form of path diagrams. Figure 9.3 is a path diagram showing how a son's socioeconomic status (SES) depends on his father's SES and his mother's SES. The three SES variables are all latent variables. Each has two observed indicators: occupational prestige and years of schooling. The diagram follows the usual convention that all the observed variables are rectangles and all the unobserved variables are ovals; it is incomplete because I have omitted the error terms for all the dependent variables.

When this model is written in the form of equations (five of them), each equation looks like an ordinary regression equation. But again, because some of the variables are not directly observed, specialized software is needed to estimate the model. The best-known programs are LISREL (Jöreskog, 1997), EQS (Byrne, 1994), and Amos (Arbuckle, 1998).

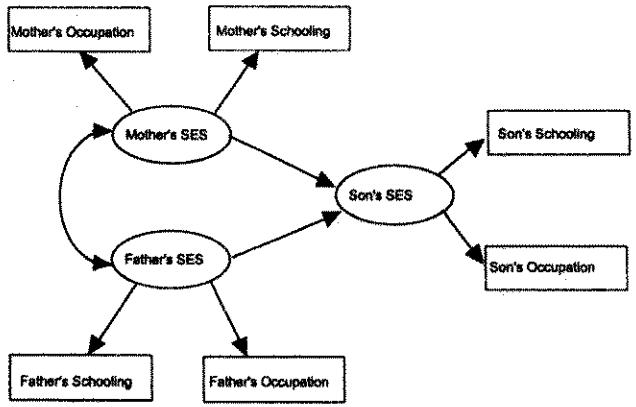

Figure 9.3. Path Diagram of a Structural Equation Model

Why would anyone want to estimate a model like this? There are two reasons. First, these methods can correct for the biasing effect of measurement error that I discussed in Chapter 3. Second, the methods can be helpful in dealing with multicollinearity (Chapter 7) that often occurs when you have several independent variables that may be measuring the same thing.

9.9. What Are Multilevel Models?

Multilevel models, also known as *hierarchical linear models*, are designed for data in which the observations are naturally clustered into groups (Bryk & Raudenbush, 1992; Kreft & De Leeuw, 1998). Imagine a survey of fourth-grade students in which the first step is to take a sample of *classrooms* from around the country. Then within each classroom, 10 students are chosen to be in the study, so students are clustered within classrooms. Suppose that the dependent variable y is a score on an academic achievement test, and the independent variable x is a measure of the child's IQ. (I'll stick with a single independent variable to keep things simple.)

Ignoring the classrooms, we could use ordinary least squares to estimate a bivariate regression model for the entire sample of the form

$$y = A + Bx + U,$$

where U is a random disturbance term. Although this would not be completely wrong (the coefficients should be unbiased), it's likely that this regression would violate the "uncorrelated disturbances" assumption discussed in Chapter 6. Specifically, we would expect the U variables for students in the same classroom to be correlated. That's because, for a variety of reasons, students in the same classroom tend to be more alike in academic performance than students in different classrooms. The consequences of violating this assumption are standard errors that are too low and test statistics that are too high.

One way to solve this problem is to estimate a multilevel model that takes the classroom structure into account. To do this, let's put some subscripts on the variables, with i denoting the student and j denoting the classroom:

$$y_{ij} = A_j + B_j x_{ij} + U_{ij}.$$

This is the student-level equation. Notice that besides putting subscripts on the variables, I've also put a j subscript on the intercept and the slope. This allows for the possibility that the slope and intercept may vary from one classroom to another. But how?

Let's suppose we've also measured the variable z_j, which is the number of years of experience of the teacher in each classroom. We can now write a classroom-level equation in which the slope and intercept at the student level depend on the teacher's experience:

$$A_j = C + Dz_j + E_j$$

$$B_j = F + Gz_j + H_j.$$

If we had several classrooms within each school, we could have additional school-level equations in which the intercepts and coefficients at the classroom level depend on variables measured at the school level.

Besides producing correct standard errors, the appeal of the multilevel model lies in its capacity to reveal how processes at the individual level are affected by things going on at the group level. As with many of the more advanced methods we've looked at in this chapter, the estimation of multilevel models requires specialized software. There are well-known standalone packages like HLM and MLn, but some of the more comprehensive packages like SAS and Stata can also estimate these models.

9.10. What Is Nonlinear Regression?

In Chapter 8, we saw how different kinds of nonlinear models could be estimated with ordinary least squares by performing transformations on the variables before doing the analysis. Some nonlinear models cannot be estimated in this way, however. Suppose, for example, that you think the relationship between x and y can best be described by the equation

$$y = 1 + Ax^B$$

where A and B are constants to be estimated. In this case, there's no way to transform the variables so that the equation can be estimated by ordinary least squares.

You may be asking yourself, "Why would I ever want to estimate an equation like this?" In sociology or political science, that's a hard question to answer, but economists do this sort of thing all the time. They construct mathematical theories of economic relationships, from which they derive testable equations that are often too nonlinear for ordinary least squares. So how do they do it? There are a variety of different techniques for estimating nonlinear regression models, but the best known is nonlinear least squares (Seber & Wild, 1989).

The basic principle of nonlinear least squares is exactly the same as ordinary least squares: Choose values of the parameters that minimize the sum of the squared prediction errors. Unfortunately, in the nonlinear case there's usually no explicit formula that produces parameter estimates that satisfy the least squares criterion. Instead, it's necessary to use an *iterative algorithm*, which can be described as a process of successive approximations. You start with reasonable guesses of the unknown parameters, called *starting values*. You plug the starting values into a formula that produces another set of numbers that are, you hope, closer to the correct solution. Then you take those numbers and plug them into the same formula to produce a third set of estimates. The process continues in this way until there's virtually no change from one *iteration* to the

next. Then we say that the algorithm has converged. Of course, the computer handles all the details of this process. Many statistical packages have procedures that perform nonlinear least squares, or some similar estimation technique.

9.11. What Is Logit Analysis?

Logit analysis—also known as *logistic regression* analysis—is a very popular regression method for *dichotomous dependent variables* (Long, 1997; Menard, 1995). Suppose, for example, that you want to estimate a regression model for explaining why some students attend college while others do not. For a sample of 19-year-olds, you could create a dummy variable with a value of 1 if they were enrolled in college, and a value of 0 if they were not enrolled. Then you could estimate a linear regression of this variable on explanatory variables like high school GPA, family income, and parents' education.

Although this isn't such a bad method—it was widely used 20 years ago—most researchers now consider it unacceptable. The problem is that dummy dependent variables necessarily violate two assumptions of the linear regression model: homoscedasticity and normality of the disturbance. The details aren't worth our attention here. Alternative methods that don't suffer from these problems include logit analysis, probit analysis, and complementary log-log analysis. The most popular by far is logit analysis. Here's a brief explanation.

Let p be the probability that a student is enrolled in college. If we have two independent variables, x_1 and x_2, the logit model says that

$$\log\left(\frac{p}{1-p}\right) = A + B_1x_1 + B_2x_2.$$

The expression on the left-hand side of this equation is called the *logit* of p. The purpose of this transformation is to convert a variable that is bounded by 0 and 1 into a variable that has no upper or lower bounds.

The most widely used method for estimating the logit model is something called *maximum likelihood*. Like nonlinear regression, maximum likelihood usually requires an iterative algorithm to produce the estimates. Although logit analysis has many things going for it,

the coefficients can be more difficult to interpret than ordinary regression coefficients.

9.12. What Is Event History Analysis?

Event history analysis is a regression method that is used for explaining or predicting the *timing of events* (Allison, 1984, 1995). The method was first developed by biostatisticians who called it *survival analysis* because they wanted to model the survival times of cancer patients. It turns out that the same methodology can be used for all sorts of events in the social sciences: births, marriages, divorces, job changes, residence changes, promotions, school dropouts, arrests, and so on. In fact, it's remarkable just how many of the things that social scientists study can be treated as events.

Why do we need a special methodology for events? For two reasons: *censoring* and *time dependent explanatory variables*. Suppose you want to develop a regression model to explain why some marriages last longer than others. You get a sample of couples who were married in, say, 1990, and you follow them for 10 years to see how long the marriages last. Let y be length of each marriage, in months. You want to estimate a regression of y on such variables as wife's age at marriage, husband's age at marriage, wife's education, and husband's education. There's a problem: Two-thirds of the marriages were still intact at the end of 10 years. The marriages that are still in progress when the study ends are called *censored*. You could delete the censored cases from the analysis, but that's losing an awful lot of data. Or you could set the marriage length at 120 months, but that's clearly an underestimate.

Before we discuss solutions to the censoring problem, there's another problem to consider: time dependent explanatory variables. Suppose you want to include number of children as one of the explanatory variables. This seems like a natural thing to do, but the number of children changes over time. You wouldn't want to count the number of children only at the end of the marriage because that would produce an artifactual positive relationship. On average, if a marriage lasts longer, there's more time available to have children.

The most popular method for dealing with both these problems is something called Cox regression, named after its inventor, Sir David Cox (1972). Cox regression has two components: a model

known as the *proportional hazards model* and an estimation method called *partial likelihood*. The proportional hazards model (for two explanatory variables) can be written as

$$\log h(t) = A + B_1x_1 + B_2x_2.$$

In this equation, $h(t)$ is the *instantaneous likelihood* of an event at time t. Partial likelihood is an iterative estimation method that is very similar to the maximum likelihood method used in logit analysis.

There are many other methods of event history analysis, including Kaplan-Meier estimation of survival curves, log-rank tests for differences in survival curves, accelerated failure time models, and discrete-time logit models. All these methods solve the problem of censoring, but not all deal with time-dependent explanatory variables.

Chapter Highlights

1. Correlation is intimately related to regression. The correlation coefficient merely describes the degree of scatter around a regression line.
2. The usual t test for a difference between two means is equivalent to regressing the variable of interest y on a dummy variable that distinguishes the two groups.
3. Analysis of variance is equivalent to multiple regression in which the independent variables are dummy variables representing different categories.
4. Analysis of covariance is equivalent to multiple regression in which some of the independent variables are dummy variables and some are interval level variables.
5. Path analysis is a method for representing a set of regression equations by way of diagrams. It is helpful in understanding and analyzing direct and indirect effects.
6. Simultaneous equations models are used to represent *reciprocal* (two-way) relationships among variables. The estimation of such models requires special techniques and somewhat demanding assumptions.
7. Factor analysis is a method for taking a set of variables and reducing them to a smaller number of dimensions. It assumes that the observed variables are linear functions of a set of unobserved or latent variables.
8. Structural equations models combine confirmatory factor analysis with simultaneous equation models.
9. Multilevel models are linear models designed for situations in which units at one level (e.g., persons) are clustered into larger units (e.g., schools). They can help us understand how individual behavior is affected by group characteristics.
10. Nonlinear regression analysis is a set of methods for estimating nonlinear equations that are too complicated to be handled by ordinary least squares.
11. Logit analysis is the most popular method for regression analysis when the dependent variable is a dichotomy. It avoids the violations of assumptions that occur when you do OLS regression with a dummy dependent variable.
12. Event history analysis is a regression method for predicting the timing of events. It is designed to solve two problems: censoring and time dependent explanatory variables.