5

Control Modeling

Elementary Control Modeling

One of the fundamental problems that researchers address when using regression analysis is determining the degree to which an effect on a dependent variable that is associated with a particular independent variable occurs as a result of the relationship between that independent variable and other independent variables. In regression analysis, we call these other independent variables "control variables."[1]

Table 5.1 is a simple example of using a control variable in ordinary least-squares regression analysis. The dependent variable is math score. The independent variables include a dummy variable for attends private school or public school and a set of dummy variables for parental education with high school or less as the excluded category. The first model shows that those students in private school score 5.06 higher than those

1. Allison (1999), pages 16–19, provides an enlightening, introductory discussion of the issue of control in regression analysis.

Table 5.1 Control Model With Math Score as Dependent Variable

	Model	
	1	2
Independent Variable	***B***	***B***
Private	5.06*	2.66*
2-Yr. Degree	—	1.32*
4-Yr. Degree	—	5.57*
Grad. Degree	—	8.72*
Intercept	50.80	47.88
R^2	.035	.140

*$p < .05$.

students in public school. The coefficient in the second model for private is 2.66, and this coefficient decreased 47% from 5.06 in the first model.

The coefficient for private in the first model is the difference in mean math scores between those who attend private school and those who attend public school. The difference in the means is 5.06, and that is the same value as the coefficient for the dummy variable for private in Table 5.1.

When the three dummy variables for parental education are added to the model, the coefficient for private school decreased to 2.66. An important question for understanding control modeling is why does the coefficient decrease? I use two ways to explain what happens to the coefficient for the independent variable of interest when a control variable is added to a regression.[2]

The first way to explain how control works is to use the analytical technique of elaboration.[3] In this instance, elaboration involves considering the difference between the mean math score for those in private and public schools within categories of parental education.

This method "controls" for parental education by considering the mean difference in math scores between respondents in private and public schools only for those respondents who have parents with the same level of education. This method produces

2. An alternative way of thinking about control variables is to use the concepts of confounding, mediating, and suppressing variables as discussed in Demaris (2004), pages 98–104; Gordon (2010), Chapter 10; and Agresti and Finlay (2009), pages 307–313.

3. Linneman (2014), Chapter 10, discusses how to use elaboration to understand how control works in regression. The textbook also explains how to use small and big models to do control modeling.

four mean differences, one for each level of parental education. Each difference is thus calculated with parental education controlled.

Elaboration for Controlling

Table 5.2 shows the difference between those respondents in private and public schools in mean math scores within categories of parental education. Although the overall difference is 5.06, the difference in each parental education category is less. A rough estimate of how much this method for "controlling" for parental education lowered the mean differences is the mean of these differences, which is 2.95. The mean of the differences is close to the coefficient for private, 2.66, in the regression that controlled for parental education.

Table 5.2 Means for Math Score for School Control Within Parental Education

Parental Education	Private/Public	Mean Math	Difference
< HS or HS Only	Public Private	47.79 51.77	3.98
2-Yr. Degree	Public Private	49.09 52.76	3.67
4-Yr. Degree	Public Private	53.55 55.79	2.24
Grad. Degree	Public Private	56.84 58.75	1.91
Mean Difference			2.95
Total	Public Private	50.80 55.86	5.06

Demographic Standardization for Controlling

A second way to explain how control works is by using the method of demographic standardization.[4] In this method, an overall mean is viewed as a weighted sum of subgroup means. In direct demographic standardization, an overall mean is adjusted by changing the weights applied to the subgroup means.

4. Treiman (2009), Chapter 2, presents an excellent discussion of using elaboration and standardization for statistical control.

A classic example of direct demographic standardization is a standardized crude death rate. The distribution of deaths in human populations shows higher rates near birth and then lower rates for childhood, adolescence, and young adulthood. Death rates start rising in middle adulthood.

The crude death rate is the total number of deaths divided by the midyear population. We can view this rate as a weighted sum of death rates for age groups weighted by the size of the age groups. Two populations can differ in overall crude death rates as a result of underlying differences in age-specific death rates and in underlying differences in age structure. Populations with age structures that feature higher proportions in younger age groups will have a lower crude death rate from that influence. Populations that feature higher proportions in older age groups will have a higher crude death rate from that influence. Directly standardized crude death rates use a standard age distribution and by doing so “control” for the effect of age structure on the overall crude death rate.

The first step in considering the issue of control is to specify an independent variable of interest that is related to the dependent variable. In Table 5.3, we use school control as the independent variable of interest and we see that those in private schools score higher in math than those in public schools. The second step is to specify a control variable that, first, has an effect on the dependent variable and, second, is related to the independent variable of interest.

We will use parental education as a control variable in our consideration of the effect of school control on math scores. The rationale is that we know that parents with more education are more likely to send their children to private schools than parents with less education. We see in Table 5.3 that that those respondents with parents with more education score higher in math than those with parents with less education.

Table 5.3 Means for School Control and Parental Education

School Control	Math Mean
Public	50.80
Private	55.86
Parental Ed.	
< HS or HS Only	48.07
2-Yr. Degree	49.48
4-Yr. Degree	54.09
Grad. Degree	57.44

The second characteristic of an effective control variable is that the variable be related to the independent variable of interest. In Table 5.4, we see that those respondents in private school are more likely to have a parent who is a 4-year college graduate or has a graduate degree than those respondents in public school.

Table 5.4 Percentages for Type of School

Parental Ed.	Percentages	
	Public	Private
< HS or HS Only	39.6	15.0
2-Yr. Degree	22.2	13.1
4-Yr. Degree	22.6	35.9
Grad. Degree	15.6	36.0
Total	100.0	100.0

Demographic standardization rests on the idea that the mean for any group is the weighted mean of the means for subgroups within the group. In this example, the means for respondents in public and private schools are the weighted means of the means in parental education subgroups.

Table 5.5 shows the public and private means expressed as weighted sums of the means for parental education subgroups. The product of the proportion in a subgroup times the mean for the subgroup is the contribution of the subgroup mean to the overall mean. Subgroups with higher proportions contribute more than subgroups with lower proportions.

The difference between the mean for respondents in private school and the mean for those in public school is 5.06. The difference is a result of two factors. One factor is that the respondents in private school have a higher math score in each category of parental education. The other factor is that the respondents in private school have higher proportions in the parental education categories where students score higher in math. The result from these two factors is a higher mean math score for those in private school compared with those in public school.

Table 5.5 shows the calculation of the mean for those respondents in private school and for those in public school as the weighted sum of the means in parental education subgroups. The calculation shows that the college graduate and graduate school categories contribute more to the mean for those in private school than for those in public school. On the other hand, the high school or less and some college categories contribute less.

Table 5.5 Group Means Expressed as Sum of Weighted Means for Subgroups

	Public			Private		
Parental Ed.	Prop.	Math Mean	Product	Prop.	Math Mean	Product
< HS or HS Only	.396	47.79	18.92	.150	51.77	7.77
2-Yr. Degree	.222	49.09	10.90	.131	52.76	6.91
4-Yr. Degree	.226	53.55	12.11	.359	55.79	20.03
Grad. Degree	.156	56.84	8.87	.360	58.75	21.15
Product Sum			50.80			55.86

In applying direct demographic standardization, the mean for those in private school is recalculated by using the parental education proportions for those in public school. Table 5.6 shows the calculation. The mean for those in private decreases from 55.86 to 53.99 when the parental education proportions for those in public school are used in place of the parental education proportions for those in private school.

Table 5.6 Calculation of Standardized Mean

	Private			Private Standardized on Public		
Parental Ed.	Prop.	Math Mean	Product	Prop.	Math Mean	Product
< HS or HS Only	.150	51.77	7.77	.396	51.77	20.50
2-Yr. Degree	.131	52.76	6.91	.222	52.76	11.71
4-Yr. Degree	.359	55.79	20.03	.226	55.79	12.61
Grad. Degree	.360	58.75	21.15	.156	58.75	9.17
Product Sum			55.86			53.99

The difference between those in private school and those in public school in mean math scores decreases from 5.06 (55.86 – 50.80) to 3.19 (53.99 – 50.80). Notice that the difference of 3.19 is fairly close in size to the coefficient for private in the regression model in Table 5.1, where parental education was controlled (2.66). Thus, demographic standardization produces an approximation to a control model in regression. The idea of controlling, then, is to constrain statistically two groups to have the same distribution on the control variable.

Small and Big Models

Professor Arthur Goldberger, a noted econometrician and teacher at the University of Wisconsin–Madison, referred to "short" and "long" models when discussing control models, and I use similar terms here.[5] A "big" model essentially adds more variables to the "small" model.

The simplest control model is shown in Table 5.7. A dummy variable for private school is the independent variable of interest. The control variables in the model include an interval variable measuring family income, a set of dummy variables measuring parental education (high school or less excluded), and a set of dummy variables measuring family structure of the respondent (two biological parents excluded). Those in private school score 1.78 higher in math scores than those in public school when family income, parental education, and family structure are controlled.

One way to think about this is to view the result as what the difference would be if those in private school and those in public school had the same distributions on family income, parental education, and family structure. We would use this one-model approach if we wanted to know the effect of the independent variable of interest when correlated, but theoretically less interesting factors are controlled. In this case, attendance at private school is related to family income, parental education, and family structure, but perhaps we are interested in the influence of attending private school itself and not interested in the characteristics of the children who attend private school. However, often we want to determine how much of the initial difference can be attributed to the control variables, and in this instance in Table 5.8, we use two regression models.

Table 5.8 shows that the effect for private school is 5.06 in the small model and 1.78 in the big model. The coefficient decreased 3.28 in actual magnitude for a decrease of 65%. Thus, more than one half of the difference between those in private school and those in public school in math scores is a result of differences in the distributions between the two groups on family income, parental education, and family structure:

5. Goldberger (1998) calls the models "short" and "long" because he has equations for regression models in mind. I call the models "small" and "big" having presentation tables in mind. A small model has a smaller number of coefficients in a column of a table, and a big model has a bigger number of coefficients.

Table 5.7 Linear Regression of Independent Variables on Math Score

Independent Variable	*B*
Private	1.78*
Family Income	.13*
2-Yr. Degree	1.14*
4-Yr. Degree	4.79*
Grad. Degree	7.34*
Bio./Step.	−1.33*
Single	−.82*
Other Fam.	−1.87*
Intercept	47.92
R^2	.157

*$p < .05$.

Table 5.8 Small and Big Models With Math Score as Dependent Variable

Independent Variable	Model	
	1	2
	B	*B*
Private	5.06*	1.78*
Family Income	—	.13*
2-Yr. Degree	—	1.14*
4-Yr. Degree	—	4.79*
Grad. Degree	—	7.34*
Bio./Step.	—	−1.33*
Single	—	−.82*
Other Fam.	—	−1.87*
Intercept	50.80	47.92
R^2	.035	.157

*$p < .05$.

$$\text{Change} = \frac{\text{big model coefficient} - \text{small model coefficient}}{\text{small model coefficient}} \times 100$$

$$= [(1.78 - 5.06)/5.06] \times 100 = 64.8\%$$

Family income, parental education, and family structure were effective control variables because the variables were strongly related to the dependent variable and strongly related to the independent variable of interest. In the regression in Table 5.8, those with higher family incomes had higher math scores than those not with high family incomes, those with higher parental education had higher math scores than those not with higher parental education, and those in two-biological-parent families had higher math scores than those not in two-biological-parent families.

Table 5.9 illustrates how family income, parental education, and family structure are strongly related to attending private school. Those in private school have higher

Table 5.9 Percentages for Control Variables by Type of School

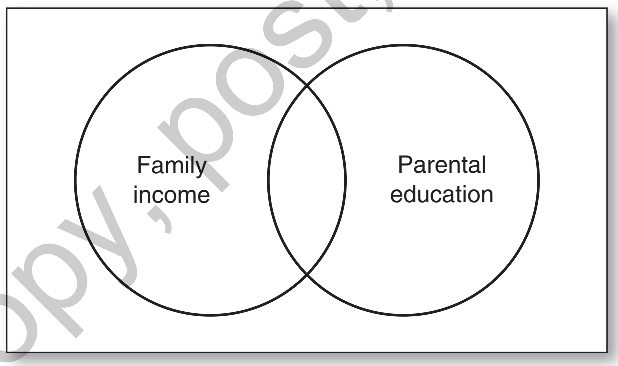
Family Income ($)	**Public**	**Private**
0–35,000	30.2	10.5
36,000–75,000	33.5	21.5
76,000–115,000	19.0	23.7
116,000+	17.3	44.3
Total	100.0	100.0
Parental Ed.	**Public**	**Private**
< HS or HS Only	39.5	15.0
2-Yr. Degree	22.2	13.1
4-Yr. Degree	22.6	35.9
Grad. Degree	15.7	36.0
Total	100.0	100.0
Family Structure	**Public**	**Private**
Two Bio. Parents	53.7	72.3
Bio./Step.	15.1	8.3
Single Parent	21.7	14.3
Other Fam.	9.5	5.1
Total	100.0	100.0

family income and are more likely to have higher parental education and to be in two-biological-parent families than those in public school.

The big model in which family income, parental education, and family structure were added explained 65% of the effect for private that we found in the small model. The big model allows us to determine how much the coefficient for private changes when the three control variables were added, but it does not allow us to determine the role of each control variable in explaining the coefficient. To determine the role of each control variable, we will need to estimate a series of control models. Complicating this decision is the fact that the control variables are usually correlated with one another.

Allocating Influence With Multiple Control Variables

The Venn diagram that follows illustrates the problem of determining the relative influences of family income and parental education in explaining the effect of private school on math scores.[6] The circle for family income captures the part of the private school effect explained by family income, whereas the parental education circle does the same for parental education. The overlap captures the correlation between family income and parental education.[7]



The problem in regard to control modeling is that whichever variable is added to the regression model first will capture the part of the private coefficient explained by the joint correlation of family income and parental education.[8] If we choose a modeling approach where we add variables in steps, the variable added earlier will have an

6. Agresti and Finlay (2009), pages 304–307, provide a brief discussion of control. They also use the idea of a Venn diagram to illustrate control, page 445.

7. The classic way to illustrate control is to use three-dimensional graphs such as in Fox (2015).

8. Goldberger (1998) points out that the first variable entered when doing a small–big model analysis always captures more of the R^2 and, therefore, allocating R^2 serves "no useful purpose."

advantage over the variable added later. If we add each control variable separately, then the joint influence is captured repeatedly in the models. Although there is no solution to this problem, we can understand the extent of the problem by modeling the control variables in different ways.

One-at-a-Time Without Controls

Table 5.10 shows the one-at-a-time model without controls approach. That is, each control variable is added without the other control variables in the model. The coefficient for attending private school decreased to 3.14 (−38%) when we added family income, to 2.66 (−47%) when we added parental education, and to 4.61 (−9%) when we added family structure. Model 5 includes all control variables, and the coefficient for private decreased to 1.78 (−65%).

Notice that the percentage change when all the control variables were added to the model is less than the sum of the percentage changes when each variable was added separately. The overlap shown in the Venn diagram is captured more than once.

Table 5.10 One-at-a-Time Without Controls With Math Score as Dependent Variable

Independent Variable	Model				
	1	2	3	4	5
	B	*B*	*B*	*B*	*B*
Private	5.06*	3.14*	2.66*	4.61*	1.78*
Family Income	—	.23*	—	—	.13*
2-Yr. Degree	—	—	1.32*	—	1.14*
4-Yr. Degree	—	—	5.57*	—	4.79*
Grad. Degree	—	—	8.72*	—	7.34*
Bio./Step.	—	—	—	−2.04*	−1.33*
Single	—	—	—	−2.30*	−.82*
Other Fam.	—	—	—	−3.25*	−1.87*
Intercept	50.80	49.01	47.88	51.91	47.92
R^2	.035	.084	.140	.049	.157

*$p < .05$.

Step Approach

The step approach in Table 5.11 involves adding one control variable, then an additional control variable, and then one more control variable.[9] Rather than comparing the models with the control variable with Model 1, in this approach, Model 2 is compared with Model 1, Model 3 with Model 2, and Model 4 with Model 3.

The coefficient for attending private school decreased to 3.14 (−38%) when we added family income, to 1.91 (−39%) when we added parental education, and to 1.78 (an additional −7%) when we added family structure. The explanatory power of parental education decreases when using this approach compared with the one-at-a-time without controls approach because the first variable added gets the overlap in explanatory power as illustrated in the Venn diagram. Thus, only family income gets the overlap in explanatory power.

In Table 5.12, we add the dummy variables for parental education first and then the interval variable for family income. The coefficient for attending private school

Table 5.11 Step Model With Math Score as Dependent Variable

Independent Variable	Model			
	1	2	3	4
	B	*B*	*B*	*B*
Private	5.06*	3.14*	1.91*	1.78*
Family Income	—	.23*	.13*	.13*
2-Yr. Degree	—	—	1.16*	1.14*
4-Yr. Degree	—	—	4.93*	4.79*
Grad. Degree	—	—	7.51*	7.34*
Bio./Step.	—	—	—	−1.33*
Single	—	—	—	−.82*
Other Fam.	—	—	—	−1.87*
Intercept	50.80	49.01	47.26	47.92
R^2	.035	.084	.153	.157

*$p < .05$.

9. Demaris (2004), page 88, provides an example of a step model.

decreased to 2.66 (–47%) when we added parental education and to 1.91 (–28%) when we added family income. Family income explained 38% of the coefficient when it was added first compared with 28% when added second. This shows that order of entry definitely matters.

In the case of parental education and family income in Table 5.12, the variable entered first not only captures the joint influence with the other variable, but also it captures the joint influence with other possible explanatory variables, whether analyzed in the analysis or not.

Please note that the step model described earlier is not created by the "stepwise" procedure found in the SPSS statistical program and in other statistical programs. Here, the step model refers to starting with a small model and adding variables by steps to create larger models that answer theoretical questions. The stepwise procedure in SPSS includes in the first model the independent variable that adds the most to R^2, then includes in the second model the variable that adds the second most to R^2, and so on. Using the step approach described in this book requires the researcher to enter variables in an order that is dictated by theoretical considerations, not by statistical considerations.

Table 5.12 Step Model With Math Score as Dependent Variable

Independent Variable	Model			
	1	2	3	4
	B	***B***	***B***	***B***
Private	5.06*	2.66*	1.91*	1.78*
2-Yr. Degree	—	1.32*	1.16*	1.14*
4-Yr. Degree	—	5.57*	4.93*	4.79*
Grad. Degree	—	8.72*	7.51*	7.34*
Family Income	—	—	.13*	.13*
Bio./Step.	—	—	—	−1.33*
Single	—	—	—	−.82*
Other Fam.	—	—	—	−1.87*
Intercept	50.80	47.88	47.26	47.92
R^2	.035	.140	.153	.157

*$p < .05$.

One-at-a-Time With Controls

The objective of control modeling is to determine how much of the effect of the independent variable of interest can be explained by a control variable. We have observed how correlation between control variables complicates the analysis. A one-at-a-time model with controls deals with the correlation issue by considering the influence of each control variable with all other control variables in the model.

Model 1 determines the baseline effect for attending private school. To determine the influence of the control variables, Model 5 is compared with Models 2, 3, and 4. Models 2, 3, and 4 each contain two control variables, and Model 5 adds the third control variable. This allows us to consider the influence of each control variable above that of the other control variables.

In Table 5.13, the coefficient for attending private school decreased to 1.78 from 2.47 when family income was added (–28%), comparing Models 2 and 5. The coefficient for private school decreased to 1.78 from 2.90 when parental education was added (–39%), comparing Models 3 and 5. Finally, the coefficient decreased to 1.78 from 1.91 when family structure was added (–7%), comparing Models 4 and 5.

Table 5.13 One-at-a-Time With Controls With Math Score as Dependent Variable

Independent Variable	Model				
	1	2	3	4	5
	B	*B*	*B*	*B*	*B*
Private	5.06*	2.47*	2.90*	1.91*	1.78*
Family Income	—	—	.22*	.13*	.13*
2-Yr. Degree	—	1.28*	—	1.16*	1.14*
4-Yr. Degree	—	5.37*	—	4.93*	4.79*
Grad. Degree	—	8.46*	—	7.51*	7.34*
Bio./Step.	—	–1.42*	–1.73*	—	–1.33*
Single	—	–1.08*	–1.57*	—	–.82*
Other Fam.	—	–2.14*	–2.55*	—	–1.87*
Intercept	50.80	48.62	49.95	47.26	47.92
R^2	.035	.145	.092	.153	.157

*$p < .05$.

Unlike the other approaches, the one-at-a-time with controls approach has a series of intermediate models that have less analytical value than the intermediate models in the other approaches. The one-at-a-time with controls allows us to affirm the results from the one-at-a-time without controls approach that family income and parental education both have strong explanatory power for explaining the effect for private school. The model also affirms that parental education has more explanatory power than family income.

The difference in the explanation between the one-at-a-time without controls approach and the one-at-a-time with controls approach lies in the handling of the part of the effect of the independent variable of interest explained by the joint correlation between variables. In the first approach, the explanation provided by the joint correlation is allocated to both variables, whereas in the second approach, the explanation is allocated to neither variable.

Hybrid Approach

All control variables do not play the same role in the analysis. Some control variables are simply correlated with the independent variable of interest. We need to control for

Table 5.14 Hybrid Approach With Math Score as Dependent Variable

Independent Variable	Model		
	1	2	3
	B	*B*	*B*
Private	5.06*	1.91*	1.78*
Family Income	—	.13*	.13*
2-Yr. Degree	—	1.16*	1.14*
4-Yr. Degree	—	4.93*	4.79*
Grad. Degree	—	7.51*	7.34*
Bio./Step.	—	—	−1.33*
Single	—	—	−.82*
Other Fam.	—	—	−1.87*
Intercept	50.80	47.26	47.92
R^2	.035	.153	.157

*$p < .05$.

these variables, but they are not high in theoretical interest. Other control variables may have direct influences on the independent variable of interest, or they may represent the intermediate mechanism by which the independent variable of interest influences the dependent variable.

The hybrid approach is similar to the step approach in that the variables are added in steps. The first step involves adding control variables that are correlated with the independent variable of interest. Later steps add control variables that are the focus of the analysis. In this way, the hybrid model resembles a one-at-a-time with controls model.

In Table 5.14, the independent variable of interest is attending private school. It is well known that those students who attend private school are more likely to come from families with higher incomes and have parents with higher levels of education than those who attend public school. Higher family income and higher parental education are both known to increase chances of attending private school. Model 1 in Table 5.14 includes only the variable for private. Model 2 includes family income and parental education. The coefficient for private decreases from Model 1 to Model 2 as the influences of family income and parental education are controlled but remains significantly different from zero.

Suppose the researcher wants to focus on family structure and believes that part of the reason attending private school influences math scores is that those who live in less advantageous family structures are less likely to attend private school. Family structure operates like family income and parental education in that these variables represent who goes to private school rather than what happens in private school.

Model 3 controls for family structure. The coefficient for private is 1.78 and still significantly different from zero. This represents a decrease from 1.91 in Model 2 and a modest 7% decrease in the private coefficient. Once family income and parental education are controlled, we find that family structure plays a relatively small role in explaining the overall effect of private school on math scores.

Nestedness and Constraints

We can use the concepts of nestedness and constraints to analyze what happens when we use control models in regression analysis. Nestedness is the idea that the variables in one model are a subset of the variables in another model. In the case of control models, variables in the small model are a subset of variables in the big model. The fact that the small model is nested in the big model allows us to use the *F* test for improvement of model fit to determine whether the additional variables significantly increased R^2.[10]

When we add one variable to the small model to obtain the big model, the *t* test for the added variable will produce the same result as the *F* test. However, if we wanted to test whether adding the three dummy variables for family structure, for example, significantly improved model fit, the *F* test would measure whether adding all three

10. Allen (1997), pages 113–117, discusses how to use the *F* test to test changes in model fit between nested models. Agresti and Finlay (2009), page 337, shows how to calculate an *F* test.

variables at once significantly improved model fit. Sometimes we might add three variables to a model, and only one has a significant *t* test. In this case, we do an *F* test to see whether the one significant variable in the set of three variables improved R^2 enough to warrant adding all three variables.

One approach to setting constraints is to set two coefficients equal. The small and big model approach to control modeling is a different way of setting constraints than setting coefficients equal. The small model is a constrained big model where certain coefficients are set to be equal to zero. We set this constraint by not adding the variables to the model.

Example Using Logistic Regression

Private schools in the United States are generally considered to be higher quality schools than public schools partly as a result of advantages in resources and teacher expertise. The analysis in Table 5.15 considers differences in private school attendance between non-Hispanic White students and Black students.

Table 5.15 Small and Big Models With Private High School as Dependent Variable

	Model	
	1	**2**
Independent Variable	***B***	***B***
Black	−.37*	−.02
Other Race/Ethnicity	−.34*	−.14*
Family Income	—	.04*
2-Yr. Degree	—	.36*
4-Yr. Degree.	—	1.10*
Grad. Degree	—	1.24*
Bio./Step.	—	−.71*
Single	—	−.32*
Other Fam.	—	−.54*
Intercept	−1.47	−2.56
−2 log likelihood	17,999.3	15,859.1

*$p < .05$.

White and Black students differ on factors correlated with private school attendance. The regression analysis in Table 5.15 considers the degree to which Black/White differences in private school attendance are a result of differences in family income, parental education, and family structure. The analysis shows that Black students are significantly less likely to attend private school than White students with other factors not controlled. The second model in Table 5.15 shows that when family income, parental education, and family structure are controlled, the Black/White difference is no longer significant.

Table 5.16 Percentages for School Control by Independent Variables

	School Control		
	Public	Private	Total
Race/Ethnicity			
White	81.4	18.6	100
Black	86.4	13.6	100
Other	86.0	14.0	100
Family Income ($)			
0–35,000	93.6	6.4	100
36,000–75,000	88.7	11.3	100
76,000–115,000	80.2	19.8	100
116,000+	66.2	33.8	100
Parental Educ.			
< HS or HS Only	93.0	7.0	100
2-Yr. Degree	89.5	10.5	100
4-Yr. Degree	76.0	24.0	100
Grad. Degree	68.6	31.4	100
Family Structure			
Two Bio. Parents	78.9	21.1	100
Bio./Step.	90.1	9.9	100
Single	88.4	11.6	100
Other Fam.	90.3	9.9	100

A preliminary step in conducting a control modeling analysis is to examine a bivariate analysis for the relationship between the independent variables and the dependent variable and then to examine a bivariate analysis for the relationship between the independent variable of interest and the control variables. Table 5.16 shows percentages attending private school by race/ethnicity, family income, parental education, and family structure. The results show that Blacks have a smaller percentage attending private school than Whites. The control variables are also strongly related to the dependent variable. Those with higher family incomes are more likely to attend than those with lower family incomes, those with more educated parents are more likely to attend than those with less educated parents, and those in two-parent families are more likely to attend than those not in two-parent families.

Table 5.17 Percentages for Control Variables by Race/Ethnicity

	Race/Ethnicity		
	White	**Black**	**Other**
Family Income ($)			
0–35,000	20.9	38.7	33.4
36,000–75,000	30.9	34.5	31.7
76,000–115,000	22.5	14.0	17.0
116,000+	25.7	12.8	17.9
Total	100	100	100
Parental Educ.			
< HS or HS Only	31.0	40.0	41.5
2-Yr. Degree	20.7	25.0	19.5
4-Yr. Degree	27.1	20.4	22.4
Grad. Degree	21.2	14.6	16.6
Total	100	100	100
Family Structure			
Two Bio. Parents	59.6	44.9	55.9
Bio./Step.	13.9	13.6	14.2
Single	18.8	28.9	20.6
Other Fam.	7.7	12.6	9.3
Total	100	100	100

Table 5.17 shows that the control variables are also related to the independent variable of interest. Blacks are less likely to have higher family incomes than are Whites, Blacks are less likely to have parents with a higher education than Whites, and Blacks are less likely to be in two-parent families than are Whites. The results of the bivariate analysis show that family income, parental education, and family structure are related to the dependent variable and that Blacks are disadvantaged on these variables compared with Whites. The results of the bivariate analysis suggest that controlling for family income, parental education, and family structure should explain part of the disadvantage that Blacks have compared with Whites in attending private school. The next question then is, to what degree do the control variables explain the disadvantage in attending private school for Blacks compared with Whites?

Table 5.18 shows the one-at-a-time model without controls. When we compare Models 2, 3, and 4 with Model 1, we see that although the coefficient for the

Table 5.18 One-at-a-Time Without Controls With Private High School as Dependent Variable

Independent Variable	Model				
	1	2	3	4	5
	B	*B*	*B*	*B*	*B*
Black	–.37*	–.13	–.19*	–.27*	–.02
Other Race/ Ethnicity	–.34*	–.21*	–.21*	–.32*	–.14*
Family Income	—	.06*	—	—	.04*
2-Yr. Degree	—	—	.42*	—	.36*
4-Yr. Degree.	—	—	1.41*	—	1.10*
Grad. Degree	—	—	1.78*	—	1.24*
Bio./Step.	—	—	—	–.89*	–.71*
Single	—	—	—	–.69*	–.32*
Other Fam.	—	—	—	–.89*	–.54*
Intercept	–1.47	–2.16	–2.48	–1.20	–2.56
–2 log-likelihood	17,999.3	16,639.8	16,660.5	17,606.9	15,859.1

*$p < .05$.

Black/White difference decreases when each control variable was added to the model, the decrease was largest for family income (−.37 to −.13). In fact, the Black/White difference was not significant when family income was controlled. However, the Black/White difference decreased by almost one half when parental education was controlled, so parental education is also an important explanatory factor (−.37 to −.19). In addition, the Black/White difference decreased by about one third when family structure was controlled so family structure also has noticeable explanatory power (−.37 to −.27). The drawback to this one-at-a-time without controls analysis is that family income, parental education, and family structure are related to one another so they share explanatory power and each one captures that shared explanatory power when added to the model separately.

Table 5.19 shows the one-at-a-time with controls approach. In this case, Models 2, 3, and 4 are compared with Model 5. Again, the biggest decrease in the Black/White coefficient occurs when family income is controlled. The Black

Table 5.19 One-at-a-Time With Controls With Private High School as Dependent Variable

Independent Variable	Model				
	1	2	3	4	5
	B	*B*	*B*	*B*	*B*
Black	−.37*	−.13	−.07	−.05	−.02
Other Race/ Ethnicity	−.34*	−.20*	−.20*	−.15*	−.14*
Family Income	—	—	.06*	.04*	.04*
2-Yr. Degree	—	.41*	—	.36*	.36*
4-Yr. Degree.	—	1.34*	—	1.15*	1.10*
Grad. Degree	—	1.68*	—	1.30*	1.24*
Bio./Step.	—	−.76*	−.79*	—	−.71*
Single	—	−.43*	−.47*	—	−.32*
Other Fam.	—	−.65*	−.67*	—	−.54*
Intercept	−1.47	−2.23	−1.91	−2.77	−2.56
−2 log-likelihood	17,999.3	16,455.8	16,414.5	16,010.4	15,859.1

$^*p < .05$.

coefficient in Model 2 is –.19 and decreases to –.02 in Model 5 for a decrease in the coefficient of .17. The Black coefficient also decreases when parental education is added, going from –.07 in Model 3 to –.02 in Model 5 for a decrease of .05. The decrease in the Black coefficient is smallest when family structure is added, going from –.05 in Model 4 to –.02 in Model 5 for a decrease of .03.

The analyses of the relative contributions of family income, parental education, and family structure that used the one-at-a-time without controls approach and the one-at-a-time with control approach were similar in saying the family income had the strongest explanatory power in explaining Black/White differences in attending private school. Both approaches showed that parental education was next in explanatory power followed by family structure. However, the relative amount of the Black coefficient explained by each variable was different in each approach as a result of the large amount of influence shared by the three factors.

Table 5.20 illustrates a step approach. The focus is on what role does family structure play in explaining the lower chances of attending private school for Black students compared with White students. First, family income and parental education are controlled since both variables are related to private school attendance and to family structure. Model 2 shows that the Black/White difference in private school attendance decreases to nonsignificance when family income and parental education

Table 5.20 Step Model With Private High School as Dependent Variable

	Model		
	1	2	3
Independent Variable	***B***	***B***	***B***
Black	–.37*	–.05	–.02
Other Race/Ethnicity	–.34*	–.15*	–.14*
Family Income	—	.04*	.04*
2-Yr. Degree	—	.36*	.36*
4-Yr. Degree.	—	1.15*	1.10*
Grad. Degree	—	1.30*	1.24*
Bio./Step.	—	—	–.71*
Single	—	—	–.32*
Other Fam.	—	—	–.54*
Intercept	–1.47	–2.77	–2.56
–2 log-likelihood	17,999.3	16,010.4	15,859.1

$^*p < .05$.

are controlled. Model 3 shows that adding family structure to the model that includes family income and parental education seems to lead to little further decrease in the Black/White coefficient.

The step model allows the variables added in the earlier steps to explain more of the coefficient for the independent variable of interest than the variables added in later steps as a result of the explanatory power of shared variation being captured in earlier steps. The contribution in the step model of adding family structure to the model in explaining the Black coefficient for Blacks of .03 [(−.02) − (−.05) = .03] is small compared with the joint contribution of family income and parental education [(−.05) − (−.37) = .32]. However, the contribution of family structure in the one-at-time without control, while the smallest of the three, was relatively larger. Thus, I suggest that a researcher explore different control modeling approaches so that the researcher has a firm idea about the relative contributions of the control variables used in the analysis under different circumstances before making conclusions.

Summary

Control modeling is the most widely used regression modeling approach that I discuss in this book. Control modeling starts with the concept of an independent variable of interest. The objective of regression modeling is to eliminate related influences that might explain the relationship, as measured by a regression coefficient, between the independent variable of interest and the dependent variable.

Standardization is a method used by demographers to hold constant the influence of a control variable on the relationship between an independent variable of interest and a dependent variable. Although demographic standardization does not produce the same result as using a control variable in dummy variable regression, the result from standardization is close enough that we can use the underlying concept in demographic standardization to understand control in regression analysis. That concept from demographic standardization is that control involves examining the difference between two groups on a dependent variable by making the underlying distribution on a third control variable equal for both groups.

A key underlying issue in control modeling is allocating the joint influence of two control variables in explaining the influence of an independent variable of interest on a dependent variable. The difficulty is there is no accepted method for allocating the joint influence of two control variables. If two variables are related to one another and to a dependent variable, then the first variable entered as a control variable will capture its unique influence and the joint influence that the variable shares with a second control variable.

The one-at-a time without controls approach to control modeling allows each control variable to capture its unique influence in explaining the effect of the independent variable of interest and any joint influences that it shares with any other control variable. On the other hand, the one-at-a time with controls approach allows each control variable only to capture its unique influence in explaining the effect of the independent variable of interest. The control variables added first in the step approach capture the joint influence of variables added in subsequent steps. Thus, the order that variables are added in the step approach has a great influence on the overall interpretation of the influence of control variables in explaining the independent variable of interest.

Researchers usually use only one control modeling approach to discuss their results in a research paper. However, I suggest that researchers carefully examine the results obtained by using alternative regression modeling approaches to understand the impact that adding control variables in some particular order had on the nature of the final result.

Key Concepts

control modeling: a modeling approach that involves first estimating the coefficient for an independent variable of interest and then adding control variables to take into account related influences.

elaboration: examining a relationship between two variables within categories of a third variable to control for the influence of the third variable.

demographic standardization: starts with the concept of a weighted mean where an overall mean is viewed as the sum of subgroup means weighted by the proportions for the subgroups; standardization involves creating an adjusted weighted mean for one group by using the subgroup proportions from a second group.

small and big models: a small model is a regression model where the variables included in the model are a subset of the variables included in a big model.

one-at-a-time without controls: a regression modeling procedure where only one control variable at a time is added to the small model, which includes the independent variable of interest.

step approach: a regression modeling procedure were one control variable is first added to the small model, which includes the independent variable of interest and then a second control variable is added to the second model to create a third model.

one-at-a-time with controls: a regression modeling procedure were only one control variable at a time is added to a smaller model, which includes the independent variable of interest and all other control variables.

hybrid approach: a regression modeling procedure that combines the step model regression modeling approach with the one-at-a-time with controls approach.

Chapter Exercises

1. Replicate the regressions and the table for the "one-at-a-time without controls" example in Table 5.10 by using X2TXMTSCOR, PRIVATE, FAMINC, TWOYR, FOURYR, GRAD, STEP, SINGLE, and FAMOTH.

2. Conduct a "one-at-a-time without controls" analysis like in Table 5.10, and create a table to present the results. Include a bivariate preliminary analysis in your answer. Use the dummy variable for two-parent family as the independent variable of interest and family income, 2-year degree, 4-year degree, graduate degree, and private as control variables. Use X2TXMTSCOR, TWOPAR, FAMINC, PAREDFOUR, PRIVATE, TWOYR, FOURYR, and GRAD in the analysis.

   What is the relationship between the independent variables and the dependent variable? What is the relationship between the independent variable of interest and the control variables? What percentage of the coefficient for the independent variable of interest was explained by controlling for all three sets of variables?

Use the following formula to calculate that percentage:

$$\frac{\text{big model coefficient} - \text{small model coefficient}}{\text{small model coefficient}} \times 100$$

What percentage of the coefficient for the independent variable of interest was explained by controlling for each set of variables separately?

3. Conduct a "hybrid" analysis like in Table 5.14, create a table to present the results, and describe your findings. Use top 25% in math as the dependent variable, and use the dummy variable for two-parent family as the independent variable of interest; add family income, 2-year degree, 4-year degree, and graduate degree in the second model, and add private in the third model as control variables. Use HIGHMATH, TWOPAR, FAMINC, PAREDFOUR, PRIVATE, TWOYR, FOURYR, and GRAD in the analysis.

   What is the relationship between the independent variables and the dependent variable? What is the relationship between the independent variable of interest and the control variables? What percentage of the coefficient for the independent variable of interest was explained by controlling for each set of variables in this stepwise manner?