Tutorial 4

A Brief Primer of Principles of Formulating and Comparing Models

This book approaches questions of designing experiments and analyzing data from the perspective of building and comparing models. As such, much of the emphasis pertains to using sample data to make inferences about the population. In this respect, the goal of data analysis is typically to decide whether the inclusion of certain effects in a statistical model significantly improves the model. However, this begs the question of what effects are considered as possible candidates for inclusion in the first place. The answer to this question inevitably hinges much more on knowledge of the subject area than on statistics. Nevertheless, there are certain statistical principles of model formulation that can help guide a researcher's thought processes about choosing potential effects to include in a model. The purpose of this tutorial is to provide a brief introduction to these general principles. We approach these principles through consideration of the consequences of mistakenly excluding relevant variables or including irrelevant variables. In a very general sense, as a researcher contemplates possible candidate effects (i.e., predictor variables) to include in a model, is it better statistically to think small or large? Is leaving out relevant effects more or less problematic than including irrelevant effects? From a narrower statistical perspective, should our full model contain only the effects of specific theoretical interest, or might there be advantages to including additional effects even if we are not directly interested in them? Even though we should emphasize that data analysis should be driven primarily by research questions (and not the other way around), nevertheless we believe that an understanding of the principles we present in this tutorial may lead to more informed choices of how best to formulate and build models, which in turn will lead to better answers to theoretical and practical questions in behavioral science research.

Our intent in writing this tutorial is to present the material in such a way that it can be read at any point after Chapter 3. Reading this tutorial immediately after Chapter 3 can ideally provide a framework for understanding how and why more complex models are developed throughout the remainder of the book. Alternatively, reading this tutorial after having read all of Chapters 1 through 16 can ideally provide a synthesis and closure for the "big picture" inherent throughout the book and thus allow readers to be certain they see the "forest amidst the trees." We leave this choice to the discretion of instructors, students, and other readers.

FORMULATING MODELS

We saw near the beginning of Chapter 3 that the general form of a statistical model can be expressed as the following.

observed value		sum of effects		sum of effects
on dependent	=	of "allowed-for"	+	of other
variable		factors		factors

This is sometimes written more succinctly as

$$\text{data} = \text{fit} + \text{error},$$

or as

$$\text{observed} = \text{predicted} + \text{error}.$$

Regardless of which way we express the general idea of a statistical model, we must decide what effects to include in the model and what other effects to exclude. By implication, any effects not explicitly included in the model become components of the error term.[1] At first glance, it might seem that the proper approach would be to include "everything" in our model, because it seems natural that to believe that "error" must be bad, so we should do as much as possible to avoid error in our model. While this perspective is not entirely wrong, it is not necessarily entirely right either. The first problem with this mind-set is that we typically do not know what "everything" even means. If we already knew with certainty all of the effects that truly influence our dependent variable, we would probably not need to conduct a study. In reality, there is an immediate risk that we may end up including some effects that unbeknownst to us are unnecessary, and at the same time, no matter how diligent we are, we will probably omit some other effects that are truly relevant. A second problem is yet more practical. Namely, even if we could identify all relevant effects, it would usually be impractical to measure all of them and include them in a single study.

A natural alternative to including "everything" in the model might be to include only the X variables of specific research interest. For example, we might be able to express our hypothesis in terms of a single X variable (say X_1) and Y. It would seem reasonable in this situation to think of a very simple model—namely, a model that included only X_1 as a predictor of Y. In fact, in this situation, it may seem rather unnecessary even to think in terms of a model. Instead, why not see whether X_1 and Y are in fact related. However, we will see later in the tutorial that even in this seemingly simple situation, there may be reasons to think in terms of models and that doing so will sometimes reveal reasons for including other X variables in our model whether or not they are a central part of our research question.

The main purpose of this tutorial is to consider several related questions: (1) How can we decide whether a variable is relevant for including in a statistical model? (2) What are the consequences of omitting a relevant variable? (3) What are the consequences of including an irrelevant variable?

RELEVANCE OF PREDICTOR VARIABLES

To understand whether a particular effect is relevant, it is helpful to return to our basic model formulation from Chapter 3. There we saw that we can write the general case of the univariate general linear model with p predictor variables (not counting the intercept) as

$$Y_i = \beta_0 X_{0i} + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \ldots + \beta_p X_{pi} + \varepsilon_i. \qquad \text{(3.1, repeated)}$$

Before proceeding, we will mention a word about notation. This tutorial discusses model formulation in terms of X variables and corresponding β coefficients. However, the principles we present here also apply if we write our model in terms of effect parameters such as

$$Y_{ij} = \mu + \alpha_j + \varepsilon_{ij}. \qquad \text{(3.59, repeated)}$$

Tutorial 3 describes the relationship between these two formulations in more detail, but for our purposes here, it suffices to say that we can think of an effect in terms of one or more X variables.

At least in principle, it is straightforward to state whether any particular variable is relevant to the model shown in Equation 3.1. Specifically, any variable with a nonzero population β coefficient is part of the "fit" of the model. By implication, any variable with a zero β coefficient is not really part of the "fit" and thus belongs in the error term of the model. Thus, if our goal is to build a complete model for Y, we should in principle include all X variables with nonzero β coefficients and exclude all X variables with zero β coefficients.[2] We could debate how well this describes the typical goal of a research program, but for our purposes here, it is important to be sure we understand at least in principle what determines whether a population β coefficient is zero or nonzero.

For example, how would we determine whether β_1 in the model shown in Equation 3.1 is zero or nonzero? Although in practice we might use sample data in an attempt to answer this question, for the moment, we will concentrate on understanding what it means to say that the population value itself is zero. To arrive at this understanding, we need to consider an alternative model for Y, namely a model that omits X_1. We could write such a model as

$$Y_i = \beta_0 X_{0i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \ldots + \beta_p X_{pi} + \varepsilon_i. \qquad (1)$$

This model could be used to obtain a predicted Y score for every individual i. We will denote these predicted scores from this model as $\hat{Y}_i(-1)$. The −1 term in parentheses serves as a reminder that we have omitted X_1 from our model. Notice that we could write the error in prediction for individual i from this model as $Y_i - \hat{Y}_i(-1)$. We can then state that the population value of β_1 is zero if and only if X_1 and the error $Y_i - \hat{Y}_i(-1)$ are uncorrelated in the population. In other words, if X_1 does not improve the prediction of Y after we use the remaining $p - 1$ X variables as predictors, then the β coefficient for X_1 is zero, and X_1 is irrelevant to the model. On the other hand, if X_1 does improve the prediction of Y after we use the remaining $p - 1$ X variables as predictors, then the β coefficient for X_1 is nonzero, and X_1 is relevant to the model.

In many situations, it is natural to think of the correlation between X_1 and the error $Y_i - \hat{Y}_i(-1)$ in terms of the correlation between X_1 and Y itself. From a practical perspective, it often seems plausible that variables that correlate with the error $Y_i - \hat{Y}_i(-1)$ will also correlate with Y, and vice versa. While this may frequently be true in behavioral science data, it is not a mathematical necessity. For example, X_1 and Y could be highly correlated with one another, and yet after we predict Y from the other X variables, the relationship between Y and X_1 has vanished in the sense that $Y_i - \hat{Y}_i(-1)$ and X_1 are uncorrelated. To understand how this might happen, suppose we are interested in understanding differences among reading ability in elementary school children. It is a virtual certainty that among all elementary school children, taller children read better than younger children. Thus if we let Y represent reading ability and X_1 represent height, Y and X_1 are undoubtedly correlated. But the fact that this correlation is nonzero does not necessarily mean that we should include height in our model of reading ability, because the correlation does not necessarily imply that the β coefficient for height predicting reading ability is nonzero. To see why, suppose we have also measured age, which we will denote as X_2. Once we predict Y from X_2, the subsequent errors will in all likelihood be uncorrelated with height. Thus even though height and reading ability are correlated, height does not improve the prediction of reading ability after we remove the predictability of age. Thus, in this scenario, β_1 is zero even though the correlation between X_1 and Y is nonzero, and we would not need to include X_1 in the model despite its correlation with Y. It turns out that the opposite pattern is also possible. In other words, it is possible that X_1 and Y are uncorrelated, and yet a relationship emerges between $Y_i - \hat{Y}_i(-1)$ and X_1. In this situation, X_1 is said to be a suppressor variable. Thus, from a mathematical perspective, we cannot judge the relevance of a predictor in a model simply on the basis of its correlation with the dependent variable. Even so, in some situations contemplating likely correlates of Y may be a useful way to assemble candidate X variables for inclusion in a model.

Of course, if we have measured Y as well as all p of the X variables, we can use sample data to test whether we need to include X_1 in our model (for example, we could test whether the regression coefficient is statistically significant). But the purpose of our discussion here is more subtle. How should we decide whether to measure X_1 in the first place? What variables are candidates for inclusion in our statistical model? It is unlikely that we can measure “everything” and then pick and choose once we have obtained our data, and even if we could measure “everything,” picking and choosing is typically not straightforward because of such complications as performing multiple tests (discussed at length in Chapter 5) and concerns about inappropriate levels of statistical power.

So how do we decide what variables to measure in our study? Should we measure X_1 or should we forgo it? If there were a simple answer to this question, science could presumably follow a simple recipe. In reality, there is no simple answer. However, there are a variety of reasons we might decide to measure a specific variable. Most obvious of these is that this variable is the center of our research attention. In fact, maybe all we care about is whether X_1 and Y are related to one another. Do we really need to worry about measuring any other variables besides X_1 and Y?

CONSEQUENCES OF OMITTING A RELEVANT PREDICTOR

Suppose for simplicity that our research question involves only Y and X_1. When, if ever, should we include other predictor variables in our model? If all other potential predictor variables would have β coefficients of zero, there would be no additional relevant predictors. We will consider the role of such predictor variables in the next section. Here we will instead consider the possibility that one or more other X variables in addition to X_1 might have nonzero β coefficients in the population model for Y. Does the existence of such additional X variables imply that they should not only be measured but also included in our statistical model and ensuing data analysis? We will see that the answer to this question depends in no small part on how one interprets the word “should” in the question of whether these additional variables should be included.

To begin to answer this question, we will consider the simplest case of only one additional X variable, which we will denote as X_2. The principles we illustrate in this simple case generalize in a straightforward fashion to the more complicated case of multiple X variables. Suppose we are planning a study investigating the relationship between X_1 and Y. Further suppose we suspect that both X_1 and X_2 may be relevant predictors (as defined in the previous section) in a model of the form:

$$Y_i = \beta_0 X_{0i} + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i. \tag{2}$$

What are the consequences of omitting X_2 from this model? In other words, what are the consequences of assessing the relationship between X_1 and Y in a model that does not include X_2, such as

$$Y_i = \beta_0 X_{0i} + \beta_1 X_{1i} + \varepsilon_i. \tag{3}$$

How important is it to include X_2 in our model? Might there be reasons to exclude X_2 even though we suspect that its regression coefficient is nonzero? All of these questions fall under a general study of specification error, which refers to the problem (typically unknown to the researcher) whereby the model is not correctly specified. What are the consequences of specification error?

We will consider the answers to these questions according to two criteria. First, does omitting X_2 affect the value we will expect to observe for $\hat{\beta}_1$? In other words, does omitting X_2 tend to produce an estimate of β_1 that is either systematically too small or too large, or on average is our estimate still likely to be correct? Second, does omitting X_2 affect how precisely we can estimate β_1? For example, if we were to form a confidence interval for β_1, will the width of the interval tend to be different if we omit X_2 instead of including it in our model? If so, statistical power will also be affected, so we will say more about this as well momentarily.

In order to explore the effects of excluding X_2 on our two criteria, we need to develop expressions for a regression coefficient and its confidence interval. Although it is possible to write an expression for an estimated regression coefficient in the general case of p predictors, we will keep things simple by considering only 1 or 2 predictor variables. First, if X_1 is the only predictor as in Equation 3, the estimated regression coefficient for X_1 can be written as

$$\hat{\beta}_1 = r_{Y1} \frac{s_Y}{s_{X_1}}, \tag{4}$$

where r_{Y1} is the correlation between Y and X_1, and s_Y and s_X are the standard deviations of Y and X, respectively. When both X_1 and X_2 are included in the model, the estimated regression coefficient for X_1 becomes

$$\hat{\beta}_{1\cdot2} = \left(\frac{r_{Y1} - r_{Y2} r_{12}}{1 - r_{12}^2} \right) \left(\frac{s_Y}{s_{X_1}} \right), \tag{5}$$

where the 1.2 subscript for $\hat{\beta}_1$ emphasizes that this is the coefficient for X_1 with X_2 also included in the model. The other new terms in Equation 5 that did not appear in Equation 4 are r_{Y2}, which is the correlation between Y and X_2, and r_{12}, which is the correlation between X_1 and X_2.

Of special interest to us in the remainder of this tutorial will be the extent to which the regression coefficient for X_1 excluding X_2 is different from the regression coefficient for X_1 including X_2. However, this is simply the difference between $\hat{\beta}_1$ in Equation 4 and $\hat{\beta}_{1\cdot2}$ in Equation 5. After some algebraic manipulation, we can write this difference as

$$\hat{\beta}_{1\cdot 2} - \hat{\beta}_1 = \left(r_{12}\frac{s_Y}{s_{X_1}}\right)\left(\frac{r_{Y1}r_{12} - r_{Y2}}{1-r_{12}^2}\right). \tag{6}$$

We can simplify this expression yet further because the term in the second set of parentheses turns out to be closely related to $\hat{\beta}_{2\cdot 1}$. In particular, the formula for the regression coefficient for X_2 when X_1 is in the model is just the same as the formula for the regression coefficient for X_1 when X_2 is in the model, except we need to reverse all subscripts for X_1 and X_2. Incorporating this modification into Equation 5, we can write $\hat{\beta}_{2\cdot 1}$ as

$$\hat{\beta}_{2\cdot 1} = \left(\frac{r_{Y2} - r_{Y1}r_{12}}{1-r_{12}^2}\right)\left(\frac{s_Y}{s_{X_2}}\right). \tag{7}$$

Substituting from Equation 7 into Equation 6 and performing a bit of algebraic manipulation allows us to rewrite the expression for the difference between regression coefficients as

$$\hat{\beta}_{1\cdot 2} - \hat{\beta}_1 = \left(-r_{12}\hat{\beta}_{2\cdot 1}\right)\left(\frac{s_{X_2}}{s_{X_1}}\right). \tag{8}$$

We will examine Equation 8 and its implications in a moment. In particular, we will often be interested in finding conditions under which we might expect the difference between the two coefficients for X_1 to equal zero. Before pursuing this question, however, we will first develop some necessary information on confidence intervals.

A general expression for a confidence interval can be written as

estimate ± (critical value)(standard error).

In the specific case of a regression coefficient, we can rewrite this expression as

$$\hat{\beta}_1 \pm (t_{.05;N-p-1})\sqrt{\frac{SS_Y(1-R^2_{Y\cdot X_1,X_2,X_3,\cdots,X_p})}{SS_{X_1}(N-p-1)(1-R^2_{X_1\cdot X_2,X_3,\cdots,X_p})}}. \tag{9}$$

In the general case of p predictor variables, the critical value comes from a t distribution with $N - p - 1$ degrees of freedom. The standard error of $\hat{\beta}_1$ depends on five values: (1) the sum of squares of Y, (2) the sum of squares of X_1, (3) the proportion of variance in Y explained by the p predictor variables, (4) the proportion of variance in X_1 explained by the remaining $p - 1$ predictor variables, and (5) the degrees of freedom.

Now we are ready to return to our questions of how omitting X_2 will affect our conclusions about X_1. Implicit in Equation 8 is the fact that the effect of excluding X_2 depends in part on whether X_1 and X_2 are correlated with one another. We will begin with the case where the two predictor variables are correlated and then consider what is different if they are not correlated.

Correlated Predictors

When the population correlation between X_1 and X_2 is nonzero, excluding a relevant X_2 variable from our model leads to a biased estimate of the regression coefficient for X_1. To see why, return to Equation 8. If both r_{12} and $\hat{\beta}_{2\cdot 1}$ are nonzero, the coefficient for X_1 when X_2 is included in the model will be different from the coefficient for X_1 when X_2 is omitted. The same principle extends from a sample to the population. Thus when the true model for Y includes X_2, and X_2

correlates with X_1, we need to include X_2 in the model in order to obtain an unbiased estimate of the true population regression coefficient for X_1. Otherwise, even in a very large sample, our estimated value of β_1 is likely to be inaccurate. More generally, this principle applies to the case of p predictor variables. If we hope to obtain an unbiased estimate of the true population regression coefficient for X_1, we must include in the model all variables that correlate with X_1 and also have nonzero regression coefficients in the model for Y. This illustrates the difficulties facing researchers who are unable to assign individuals to groups at random, because the number of variables possibly impacting Y and correlated with X_1 sometimes seems virtually boundless in the behavioral and social sciences.

In this case, excluding X_2 from the model also influences the width of the confidence interval we might form for β_1. However, the width of the interval usually becomes a secondary consideration when the center of the interval is likely to be inaccurate because of bias in the estimation of β_1. For this reason, we will not pursue considerations of width further in this case.

Examples

Chapter 9 contains an extensive discussion of the role of a covariate in adjusting group differences in the absence of random assignment to groups. For example, Figure 9.4 and the accompanying verbal description explain why adjusted differences that arise by including the covariate in the model may be different from unadjusted differences obtained when the covariate is not included in the model. Specifically, in accordance with Equation 8, adjusted differences will differ from unadjusted differences to the extent that both: (a) the covariate is related to the dependent variable, and (b) groups have different means on the covariate (because this implies that the covariate is correlated with group membership). Another example of this general principle occurs in Chapter 7, where we discuss unequal n designs with more than one factor. In so-called nonorthogonal designs, the variables representing the main effects and interaction are correlated with one another. To the extent that main effects and an interaction are non-null, the estimates we obtain for an effect will depend on what other effects we include in our model. In particular, Chapter 7 presents three types of sums of squares (literally Type I, Type II, and Type III) of possible relevance in factorial designs. Each of these types of sums of squares corresponds to a specific question, and we may get different answers to these questions even with the same data in an unequal n factorial design.

Uncorrelated Predictors

When the population correlation between X_1 and X_2 is zero, excluding a relevant X_2 variable from our model does not affect the expected value of $\hat{\beta}_1$. Notice that this is a very different outcome from the previous situation we discussed, where X_1 and X_2 were correlated. To see why the outcome is so different now, return to Equation 8. If r_{12} is zero, the coefficient for X_1 when X_2 is included in the model will be equal to the coefficient for X_1 when X_2 is omitted, even if $\hat{\beta}_{2\cdot 1}$ is nonzero. Of course, in a sample, r_{12} will typically not be exactly zero even if ρ_{12} does equal zero. However, when ρ_{12} equals zero, the long-run average value of $\hat{\beta}_1$ will equal the long-run average value of $\hat{\beta}_{1\cdot 2}$. Said another way, in a very large sample, $\hat{\beta}_1$ and $\hat{\beta}_{1\cdot 2}$ will essentially be equal to one another. On average, we will tend to obtain the correct value for the regression coefficient for X_1 even if we leave X_2 out of our model. Thus, although the true model for Y includes X_2, if X_2 does not correlate with X_1, we do not need to include X_2 in the model in order to obtain an unbiased estimate of the true population regression coefficient for X_1.

When can we be reasonably certain that some potential predictor X_2 will not correlate with our predictor variable of interest X_1? The one situation where we can typically develop a strong theoretical argument that X_2 will not correlate with X_1 is when we have formed experimental groups through random assignment, and X_1 represents membership in the groups we have formed. From this perspective, the primary benefit of random assignment is that it ensures (in a probabilistic sense) that no other X variables can correlate with X_1. Thus the beauty of random assignment is that, in this respect, it relieves us of the responsibility to identify additional predictor variables to include in our model in order to obtain an unbiased estimate of the regression coefficient for our predictor of main theoretical interest. At least in the long run, we can expect to obtain the same estimated effect of X_1 on Y regardless of whether we include other predictors in the model whenever we have randomly assigned individuals to levels of X_1.

The previous discussion might seem to imply that it makes no difference whatsoever whether we include X_2 or exclude X_2 in our model when X_2 is uncorrelated with Y. Although this is true in the long run, it is not necessarily true in general. To begin to understand why, you need to realize that so far our focus has been on the average value we would expect to obtain for $\hat{\beta}_1$. Suppose, however, that one approach might yield values for $\hat{\beta}_1$ that range from 2 to 12 from sample to sample, while a second approach yields values that range from 6 to 8. Clearly, we would prefer to use the second approach to estimate the population value of β_1. Although both approaches appear to produce an average value of 7, the second approach is preferable because estimates vary less from sample to sample, and thus any given sample is likely to provide a more precise estimate of the population parameter. Another way of thinking about this is that the second approach will provide narrower confidence intervals than the first approach.

What does this have to do with including or excluding X_2 from our model? We can compare the likely width of confidence interval estimates for the regression coefficient for X_1 in the model with X_2 versus the model without X_2. If there are reasons to expect one of these intervals to be narrower than the other, that will provide grounds for preferring one model over the other. To make this comparison, we need to return to Equation 9. When X_1 is the only predictor variable in the model, we can write this equation more simply as

$$\hat{\beta}_1 \pm (t_{\cdot 05;N-2})\sqrt{\frac{SS_Y(1-r_{Y1}^2)}{SS_{X_1}(N-2)}} \quad . \tag{10}$$

Similarly, when X_2 is included in the model, Equation 9 becomes

$$\hat{\beta}_{1\cdot 2} \pm (t_{\cdot 05;\,N-3})\sqrt{\frac{SS_Y(1-R_{Y.X_1X_2}^2)}{SS_{X_1}(N-3)(1-r_{12}^2)}} \; . \tag{11}$$

We can use Equations 10 and 11 to see that the ratio of the width of the confidence interval excluding X_2 to the width including X_2 will be given by

$$\frac{t_{\cdot 05;N-2}}{t_{\cdot 05;N-3}}\sqrt{\frac{(N-3)(1-r_{Y1}^2)}{(N-2)(1-R_{Y\cdot X_1,X_2}^2)}}\sqrt{1-r_{12}^2} \; . \tag{12}$$

What does Equation 12 tell us about the widths of the two confidence intervals? The first term—i.e., the ratio of critical t values—will always be smaller than 1.0, but will be very close to 1.0 unless N is very small. For example, even if the total sample size N is only 20, the ratio of critical t values is between 0.99 and 1.00. For larger values of N, the ratio is even closer to 1.0. Thus there is a slight advantage for the interval excluding X_2, but the advantage will be of no practical value unless N is very small. Now let's jump to the third term (we will consider the second term in a moment). The third term will tend to be very close to 1.0 with random assignment. In fact, when X_1 and X_2 represent two manipulated factors, it will usually be the case that their correlation is exactly zero. Now let's consider the second term. How will the numerator compare to the denominator? There is a potential trade-off here. The $N - 3$ portion of the numerator will be smaller than the $N - 2$ portion of the denominator, so this part of the ratio is less than 1.0. However, to the extent that X_2 is a relevant predictor of Y, the multiple correlation predicting Y from both X_1 and X_2 will be larger than the correlation predicting Y from X_1 alone. Thus this portion of the denominator will tend to be smaller than the comparable portion of the numerator, which means that the ratio will exceed 1.0. Thus one portion of the ratio will be less than 1.0, while another portion will tend to be greater than 1.0. What can we say about the ratio as a whole? After some algebraic manipulation, it is possible to show that the middle term in Equation 12 will be less than 1.0 if and only if

$$\frac{r_{Y2}^2}{(1 - R_{Y\cdot12}^2)/(N - 3)} < 1 . \tag{13}$$

We can simplify this by realizing that the left side of the inequality is identical to the observed F value comparing a model with both X_1 and X_2 as predictors to a model with only X_1 as a predictor, when X_1 and X_2 are uncorrelated. Thus, letting F_{β_2} denote this F statistic, we can say that the middle term of Equation 12 will be less than 1 if and only if

$$F_{\beta_2} < 1 . \tag{14}$$

In other words, when X_1 and X_2 are uncorrelated and N is large enough (say 20 or more) so that the ratio of critical t values is essentially 1, excluding X_2 will yield a more precise interval than including X_2 if and only if the F statistic associated with X_2 is less than 1. To see why this matters, we need to realize that if the population coefficient for X_2 is zero, we would expect the F value for X_2 to exceed 1 roughly half the time and be less than 1 the other half of the time. However, to the extent that the coefficient for X_2 is nonzero, its F statistic will tend to be larger than 1, in which case including it in our model will lead to a more precise interval for β_1.

So where does all of this leave us? Consider an example where the correlation between Y and X_1 is .40, the correlation between Y and X_2 is .50, and X_1 and X_2 are uncorrelated. It follows that the squared multiple correlation predicting Y from both X_1 and X_2 then equals .41, in which case the second term in Equation 12 will equal 1.19 if the total sample size is large enough that the square root of the ratio of $N - 3$ to $N - 2$ can be assumed to essentially equal 1.0. This means that in this scenario, a confidence interval formed for β_1 without including X_2 in the model will tend to be approximately 20% wider than an interval formed while including X_2 in the model. Thus even though X_2 is unrelated to X_1, including it in the model improves our ability to estimate the coefficient for X_1 precisely. Including X_2 helps us because X_2 improves the predictability of Y, and thus reduces error variance in the model.

There are two practical implications of this discussion:

1. When X_2 is uncorrelated with X_1, we do not need to include it in our model in order to obtain an unbiased estimate of the regression coefficient for X_1.
2. When X_2 is imcorrelated with X_1, including X_2 in the model can increase the precision with which we estimate the coefficient for X_1 to the extent that X_2 correlates with Y. So although we do not have to include X_2 in our model, there nevertheless may be definite advantages to doing so. From the perspective of hypothesis testing, a narrower confidence interval leads to increased statistical power, so we could say that including X_2 in our model in this scenario increases the power of our test for X_1.

Examples

A classic example of this situation occurs in Chapter 9, where including a covariate in the model for Y can sometimes greatly increase the power to detect a treatment effect (represented by X_1) when we have randomly assigned individuals to groups. Notice that as long as we have random assignment, we do not need to include the covariate in our model in order to obtain an unbiased estimate of the treatment effect. For this reason, many researchers would overlook the option to add the covariate to the model. However, to the extent that the covariate is related to the dependent variable, failing to include it in the model misses an opportunity to explain additional variance and thus increasing power for detecting a treatment effect while also increasing precision to estimate the magnitude of the effect. Another example occurs in the within-subjects designs we present in Chapters 11 through 14, where we (explicitly or implicitly) include a parameter for each subject in the model. For example, the full model we present in Chapter 11 for a single-factor within-subjects design is

$$Y_{ij} = \mu + \alpha_j + \pi_i + \varepsilon_{ij}, \qquad \text{(11.22, repeated)}$$

where π_i, is an effect associated with person (i.e., subject) i. Notice how this model is different from the corresponding full model we developed for a single-factor between-subjects design:

$$Y_{ij} = \mu + \alpha_j + \varepsilon_{ij}. \qquad \text{(3.59, repeated)}$$

The only difference between these models is that the model for within-subjects designs includes extra predictors to reflect potential differences between individual subjects. As long as there are no missing data, variables representing group differences will be uncorrected with variables representing subject effects. Thus we could omit the subject variables and still obtain an unbiased estimate of group differences. However, to the extent that the subject variables are predictive of Y (i.e., there are systematic individual differences between subjects), including these extra predictors will (as shown in Equations 12 and 14) increase the precision of our estimated group differences. Indeed, this is precisely why for a fixed number of data points, within-subjects designs are usually more powerful than between-subjects designs.

CONSEQUENCES OF INCLUDING AN IRRELEVANT PREDICTOR

The implication of the previous section might seem to be to include everything in the "kitchen sink" in our model. After all, it seems there can be advantages to including an additional predictor whether or not it is correlated with X_1, the predictor we have presumed to be of particular theoretical interest. However, you need to realize that the previous section assumed throughout that the additional predictor X_2 was a relevant predictor. In other words, we assumed in all of the previous section that the regression coefficient for predicting Y from X_2 was nonzero. In actual research practice, however, we will typically not know before we collect data whether our X_2 variable is in fact a relevant predictor. Suppose we thought that X_2 would be relevant, but unbeknownst to us, in reality, X_2 is irrelevant. In other words, suppose we thought that X_2 had a nonzero regression coefficient, but in actuality the population value of the regression coefficient for X_2 is zero. Remember that an "irrelevant" predictor is one whose population regression coefficient equals zero. What are the consequences of including an irrelevant predictor variable in our model?

As we did in our consideration of a relevant X_2 variable, we will consider the effects of including an irrelevant predictor for two separate scenarios. First, we will consider a situation where the irrelevant predictor correlates with X_1. Second, we will consider a scenario where X_2 does not correlate with X_1. Fortunately, in both cases, we will be able to rely on many of the formulas we developed in the previous section.

Correlated Predictors

Suppose that X_2 is correlated with X_1, but that the regression coefficient for X_2 predicting Y is zero with both X_1 and X_2 in the model. Will our estimated coefficient for X_1 still be unbiased? The answer to this question follows immediately from reconsideration of Equation 8:

$$\hat{\beta}_{1.2} - \hat{\beta}_1 = (-r_{12}\hat{\beta}_{2.1})\left(\frac{S_{X_2}}{S_{X_1}}\right). \qquad \text{(8, repeated)}$$

When X_2 is an irrelevant predictor, we would expect its regression coefficient $\hat{\beta}_{2\cdot1}$ to equal zero, which means that we would expect the difference between $\hat{\beta}_{1\cdot2}$ and $\hat{\beta}_1$ to equal zero. Thus the estimated regression coefficient for X_1 will be unbiased regardless of whether we include X_2 in our model.

In the long run, it does not matter whether we include an irrelevant correlated predictor in our model. However, what about the possible effects in any particular sample? To address this question, we need to consider the confidence interval for the regression coefficient of X_1. Recall from Equation 12 that the ratio of the width of the confidence interval excluding X_2 to the width including X_2 will be given by

$$\frac{t_{.05;\,N-2}}{t_{.05;\,N-3}}\sqrt{\frac{(N-3)(1-r_{Y1}^2)}{(N-2)(1-R_{Y.X_1,X_2}^2)}}\sqrt{1-r_{12}^2}. \qquad \text{(12, repeated)}$$

Once again, the ratio of critical t values can be thought of as equal to 1.0 except in very small samples. In this case, it turns out that we would expect the second term to be close to 1.0 as well. The reason is that now X_2 does not add to the predictability of Y. However, the third term will not equal 1.0 to the extent that X_1 and X_2 are correlated. For example, suppose X_1 and X_2 correlate .50 with one another. In this case, the third term equals .87, which means that the confidence interval for β_1 excluding X_2 will be 87% as wide as the interval we would obtain if we included X_2. Or, equivalently, by including X_2 in the model, we have increased the likely width of our confidence interval by 15% (i.e., 100 times the reciprocal of .87). Thus if we could realize that X_2 is irrelevant, we should exclude it from our model, because its inclusion decreases the precision with which we can estimate the regression coefficient for X_1.

Example

It might seem difficult to come up with an example of a situation where we would include an irrelevant predictor. Of course, if we could know with certainty that a predictor is in fact irrelevant, we would exclude it. The problem is that in reality we typically cannot know whether a predictor is truly irrelevant. And keep in mind that omitting a relevant predictor that is correlated with X_1 yields a biased estimate of β_1. For this reason, we may prefer to include some predictor X_2 in our model if it correlates with X_1, even if there is some reasonable chance that X_2 may be irrelevant. Indeed, this is exactly the logic in unequal n factorial designs in Chapter 7. Specifically, the Type II sum of squares for a main effect is calculated by omitting the interaction from the model. This is unquestionably the best approach if the interaction (which we can think of here as X_2) is truly irrelevant in the population. In an unequal *n* design, the predictor representing the interaction will typically correlate with the predictor representing a main effect, so we can increase precision by excluding the interaction from our model when it is irrelevant. However, the rub here is that in so doing, we are running a risk that we have in fact excluded a relevant correlated predictor from our model, in which case we saw in the previous section that our estimate of the main effect will be biased. Most statisticians have come to believe that it is generally preferable to sacrifice some precision in order to assure an unbiased estimate, in which case the interaction term would be included in the model. However, if there is strong evidence (perhaps based on a combination of data and theory) that the interaction is truly zero, greater power and precision can be obtained for the main effect by excluding the interaction term from the model.

Uncorrelated Predictors

Suppose that X_2 is uncorrelated with X_1 and that the population regression coefficient for X_2 predicting Y is zero with both X_1 and X_2 in the model. Notice in this case that X_2 is unrelated to any other variables in the system. As such, X_2 is literally random noise. Fortunately, this is precisely what the error term of the model is designed to accommodate. As such, the likely intuition that it should not make much difference whether we include or exclude X_2 in this case is basically correct. Even here, however, there is a theoretical difference, and that theoretical difference can become a practical difference when we realize that in actual research the number of such irrelevant variables we may be contemplating could be much larger than 1.

We can immediately see from Equation 8 that omitting an irrelevant uncorrelated predictor does not bias our estimate of the regression coefficient of X_1. Once again, in the long run or in very large samples, it will not matter whether we include or exclude this type of X_2 in our model.

Including or excluding X_2 in the model does have some effect on the precision with which we can estimate X_1, but the effect is typically quite small. Notice that we would expect both the second and third terms in Equation 12 to be essentially 1.0 when X_2 correlates with neither Y nor X_1. In particular, based on Equation 14, we can conclude that roughly half the time the second term will be larger and half the time the second term will be smaller if we include X_2. Thus the only systematic effect of including or excluding X_2 manifests itself in the critical value. As we have already seen, this effect can be ignored unless sample size is very small or unless the number of additional variables we might include becomes large. Not surprisingly, it is preferable to exclude variables that have nothing to do with either Y or X_1, but their inclusion causes little harm when the number of such variables is small relative to sample size.

SUMMARY

It is hardly a surprise to learn that relevant predictors should be included in a model and that irrelevant predictors should be excluded. However, understanding the principles behind this conclusion may be less intuitive. In particular, understanding the consequences of mistakenly excluding a relevant predictor as compared to including an irrelevant predictor may be useful in guiding decisions about what variables to measure and include in a statistical model.

Table T4.1 provides a summary of the principles we have developed in this tutorial. Implicit in the table is the suggestion that in general, it is most important to include variables that are likely to have nonzero coefficients and to correlate with other variables of interest. On the other hand, the variables most important to exclude are those that are correlated with variables of interest but whose coefficients are zero. Of course, in practice, the difficulty with this distinction is that it is typically difficult to know whether the coefficient for a predictor is truly zero or nonzero until we have included it in the model, at least provisionally. Even then, the decision may not be obvious, especially if statistical power may be lacking.

Another implication of the table is the enormous benefit of random assignment to treatments. When our goal is to assess the effect of a variable whose levels are determined by random assignment, we know (probabilistically) that we are in the second row of Table T4.1. However, the consequences of mistakenly excluding or including a variable are usually much less in the second row of the table than in the first row. Thus random assignment offers a degree of protection rarely available in nonrandomized studies.

Finally, we will close with six additional points. First, it is important to realize that X_2^2 is not the same variable as X_2. In other words, at least in theory, it may not be enough to include X_2 in our model because it may relate nonlinearly to Y. For example, in order to obtain an unbiased estimate of β_1, it may be necessary to include both X_2 and X_2^2 in the model. Obviously, this can create sizable complications, because even if we are so insightful as to know all of the relevant variables, our work is still not finished because we also have to know how each of them relates to Y. Second, a related type of effect is an interaction effect, initially described in Chapter 7. An interaction effect can be thought of as the product of two or more variables with one another, such as X_1X_2. Notice from this perspective that X_2^2 could be conceptualized as X_2 interacting with itself. In fact, consistent with the idea of an interaction we develop in Chapter 7, an effect due to X_2^2 suggests that the effect of X_2 on Y changes depending on the value of X_2 itself. Readers interested in further similarities between X_1X_2 and X_2^2, and subsequent implications for interpreting interactions and higher-order trends, are advised to read Lubinski and Humphreys's (1990) excellent article as well as MacCallum and Mar's (1995) follow-up. Third, it is also important to realize that an observed score on X_2 will frequently be different from the corresponding true score on X_2. To the extent that Y depends on the true score of X_2, but we include an imperfectly measured version of X_2 in our model, we have not completely succeeded in including the truly relevant predictor (Figure 9.7 and the accompanying discussion in Chapter 9 provides an illustration of this problem in the context of analysis of covariance). Structural equation modeling (also referred to as latent variable modeling) provides a viable solution to this dilemma. Many books and articles have been written on this topic. Good introductions are available in such sources as Bollen (1989), Kaplan (2009), and Raykov and Marcoulides (2006). Fourth, we need to point out an additional complication when X_2, the variable we may mistakenly exclude, represents a random effects factor. As we discuss at some length in Chapter 10, the presence of a random effects factor often implies the need for an error term that takes this factor into account, even if our statistical test or interval pertains to a different factor (i.e., predictor) in the model. Thus omitting relevant random effects factors can have especially unfortunate consequences. Fifth, we have described the goals of model building and the consequences of misspecifying models. These principles are exemplified at various points throughout the body of the text. However, our focus is primarily on the inclusion or exclusion of design factors, especially in randomized studies. As we have seen, issues of including or excluding predictors are often less complex in randomized designs. A broader discussion emphasizing observational studies lacking random assignment is a highly complex topic requiring a book unto itself. Fortunately, we can highly recommend Harrell (2015) for readers who are interested in such a book-length treatment. Sixth, this entire tutorial could be thought of as exemplifying the role of parsimony in building statistical models for data. In this respect, it seems fitting to close with a quote from Einstein: "Everything should be made as simple as possible, but not simpler."

TABLE T4.1
EFFECTS OF EXCLUDING A RELEVANT VARIABLE OR INCLUDING AN IRRELEVANT VARIABLE

	Excluding a ***Relevant X***	*Including an* ***Irrelevant X***
Correlated with X_1, predictor of theoretical interest	Biased estimate of β_1	Unbiased estimate of β_1, but estimate is less precise than if irrelevant X were excluded
Uncorrelated with X_1, predictor of theoretical interest	Unbiased estimate of β_1, but estimate is less precise than if relevant X were included	Unbiased estimate of β_1, but estimate is slightly less precise than if irrelevant X were excluded

NOTES

1. More technically, it could be said that the part of any excluded effect that does not correlate with effects already included in the model becomes a component of the error term.
2. It is tempting to assume that an effect that correlates with Y should have a nonzero regression coefficient. However, we will see momentarily that this is not necessarily true. An effect can correlate with Y and yet still have a regression coefficient of zero. Similarly, it might seem that an effect with a nonzero regression coefficient must correlate with Y. However, as we describe later in this tutorial, it is possible for a predictor variable to have a nonzero regression coefficient, even though the variable fails to correlate with Y. Thus, in principle, identifying effects with nonzero regression coefficients involves more than identifying variables that can be expected to correlate with Y.

REFERENCES TUTORIAL 4

Bollen, K. A. (1989). *Structural equations with latent variables*. New York: Wiley.

Harrell, F. E. (2015). *Regression modeling strategies: With applications to linear models, logistic and ordinal regression, and survival analysis (2nd ed.)*. New York: Springer-Verlag.

Kaplan, D. (2009). *Structural equation modeling: Foundations and extensions (2nd ed.)*. Thousand Oaks, CA: Sage.

Lubinski, D., & Humphreys, L. (1990). Assessing spurious "moderator effects": Illustrated substantively with the hypothesized ("synergistic") relation between spatial and mathematical ability. *Psychological Bulletin*, *107*, 385–393.

MacCallum, R. C., & Mar, C. (1995). Distinguishing between moderator and quadratic effects in multiple regression. *Psychological Bulletin*, *118*, 405–421.

Raykov, T., & Marcoulides, G. A. (2006). *A first course in structural equation modeling (2nd ed.)*. Mahwah, NJ: Lawrence Erlbaum Associates.