where e is the vector of residuals. Express the standardized coefficients in terms of the unstandardized coefficients and the sample variances of U, V, Y.

6.4 Inferring causation by regression

The key to making causal inferences by regression is a *response schedule*. This is a new idea, and a complicated one. We'll start with a mathematical example to illustrate the idea of a "place holder." Logarithms can be defined by the equation

$$\log x = \int_1^x \frac{1}{z}\,dz \text{ for } 0 < x < \infty. \tag{11}$$

The symbol ∞ stands for "infinity." But what does the x stand for? Not much. It's a place holder. You could change both x's in (11) to u's without changing the content, namely, the equality between the two sides of the equation. Similarly, z is a place holder inside the integral. You could change both z's to v's without changing the value of the integral. (Mathematicians refer to place holders as "dummy variables," but statisticians use the language differently: section 6 below.)

Now let's take an example that's closer to regression—Hooke's law (section 2). Suppose we're going to hang some weights on a spring. We do this on n occasions, indexed by $i = 1, \ldots, n$. Fix an i. If we put weight x on the spring on occasion i, our physicist assures us that the length of the spring will be

$$Y_{i,x} = 439 + 0.05x + \epsilon_i\,. \tag{12}$$

If we put a 5-unit weight on the spring, the length will be $439 + 0.05 \times 5 + \epsilon_i = 439.25 + \epsilon_i$. If instead we put a 6-unit weight on the spring, the length will be $439.30 + \epsilon_i$. A 1-unit increase in x makes the spring longer, by 0.05 units—causation has come into the picture. The random disturbance term ϵ_i represents measurement error. These random errors are IID for $i = 1, \ldots, n$, with mean 0 and known variance σ^2. The units for x are kilograms; the units for length are centimeters, so ϵ_i and σ must be in centimeters too. (Reminder: IID is shorthand for independent and identically distributed.)

Equation (12) looks like a regression equation, but it isn't. It is a response schedule that describes a theoretical relationship between weight and length. Conceptually, x is a weight that you could hang on the spring. If you did, equation (12) tells you what the spring would do. This is all in the subjunctive.

Formally, x is a place holder. The equation gives length $Y_{i,x}$ as a function of weight x, with a bit of random error. For any particular i, we can choose *one* x, electing to observe $Y_{i,x}$ for that x and that x only. The rest of the response schedule—the $Y_{i,x}$ for the other x's—would be lost to history.

Let's make the example a notch closer to social science. We might not know (12), but only

$$Y_{i,x} = a + bx + \epsilon_i\,, \tag{13}$$

where the ϵ_i are IID with mean 0 and variance σ^2. This time, a, b, and σ^2 are unknown. These parameters have to be estimated. More troublesome: we can't do an experiment. However, observational data are available. On occasion i, weight X_i is found on the spring; we just don't quite know how it got there. The length of the spring is measured as Y_i. We're still in business, *if*

(i) Y_i was determined from the response schedule (13), so $Y_i = Y_{i,X_i} = a + bX_i + \epsilon_i$, and

(ii) the X_i's were chosen at random by Nature, independent of the ϵ_i's.

Condition (i) ties the observational data to the response schedule (13), and gives us most of the statistical conditions we need on the random errors: these errors are IID with mean 0 and variance σ^2. Condition (ii) is *exogeneity*. Exogeneity—$X \perp\!\!\!\perp \epsilon$—is the rest of what we need. With these assumptions, OLS gives unbiased estimates for a and b. Example 4.1 explains how to set up the design matrix. Conditions (4.1–5) are all satisfied.

The response schedule tells us that the parameter b we're estimating has a causal interpretation: if we intervene and change x to x', then y is expected to change by $b(x' - x)$. The response schedule tells us that the relation is linear rather than quadratic or cubic or . . . . It tells us that interventions won't affect a or b. It tells us the errors are IID. It tells us there is no confounding: X causes Y without any help from any other variable. The exogeneity condition says that Nature ran the observational study just the way we would run an experiment. We don't have to randomize. Nature did it for us. Nice.

What would happen without exogeneity? Suppose Nature puts a big weight X_i on the spring whenever ϵ_i is large and positive. Nasty. Now OLS over-estimates b. In this hypothetical, the spring doesn't stretch as much as you might think. Measurement error gets mixed up with stretch. (This is "selection bias" or "endogeneity bias," to be discussed in chapters 7 and 9.) The response schedule is a powerful assumption, and so is exogeneity. For Hooke's law, the response schedule and exogeneity are reasonably convincing. With typical social science applications, there might be some harder questions to answer.

The discussion so far is about a one-dimensional x, but the generalization to higher dimensions is easy. The response schedule would be

$$Y_{i,x} = x\beta + \epsilon_i, \tag{14}$$

where x is $1 \times p$ vector of treatments and β is a $p \times 1$ parameter vector. Again, the errors ϵ_i are IID with mean 0 and variance σ^2. In the next section, we'll see that path models put together several response schedules like (14).

> A response schedule says how one variable would respond, if you intervened and manipulated other variables. Together with the exogeneity assumption, the response schedule is a theory of how the data were generated. If the theory is right, causal effects can be estimated from observational data by regression. If the theory is wrong, regression coefficients measure association not causation, and causal inferences can be quite misleading.

Exercise set D

1. (This is a hypothetical; SAT stands for Scholastic Achievement Test, widely used for college admissions in the US.) Dr. Sally Smith is doing a study on coaching for the Math SAT. She assumes the response schedule $Y_{i,x} = 450 + 3x + \delta_i$. In this equation, $Y_{i,x}$ is the score that subject i would get on the Math SAT with x hours of coaching. The error term δ_i is normal, with mean 0 and standard deviation 100.
   (a) If subject #77 gets 10 hours of coaching, what does Dr. Smith expect for this subject's Math SAT score?
   (b) If subject #77 gets 20 hours of coaching, what does Dr. Smith expect for this subject's Math SAT score?
   (c) If subject #99 gets 10 hours of coaching, what does Dr. Smith expect for this subject's Math SAT score?
   (d) If subject #99 gets 20 hours of coaching, what does Dr. Smith expect for this subject's Math SAT score?
2. (This continues exercise 1; it is still a hypothetical.) After thinking things over, Dr. Smith still believes that the response schedule is linear: $Y_{i,x} = a + bx + \delta_i$, the δ_i being IID $N(0, \sigma^2)$. But she decides that her values for a, b, and σ^2 are unrealistic. (They probably are.) She wants to estimate these parameters from data.
   (a) Does she need to do an experiment, or can she get by with an observational study? (The latter would be much easier to do.)

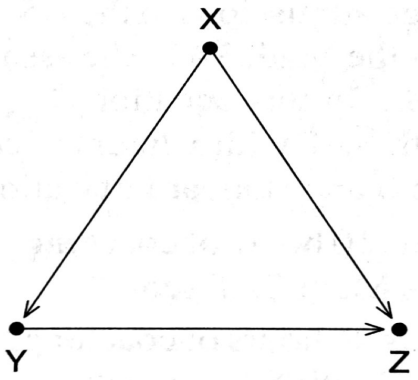
(b) If she can use observational data, what else would she have to assume, beyond the response schedule?

(c) And, how would she estimate the parameters from the observational data?

6.5 Response schedules for path diagrams

Path models are often held out as rigorous statistical engines for inferring causation from association. Statistical techniques can indeed be rigorous—given their assumptions. But the assumptions are usually imposed on the data by the analyst: this is not a rigorous process. The assumptions behind the models are of two kinds: (i) causal and (ii) statistical. This section will lay out the assumptions in more detail. A relatively simple path model is shown in figure 3, where a hypothesized causal relationship between Y and Z is confounded by X.

Figure 3. Path model. The relationship between Y and Z is confounded by X. Free arrows leading into Y and Z are not shown.



This sort of diagram is used to draw causal conclusions from observational data. The diagram is therefore more complicated than it looks: causation is a complicated business. Let's assume that Dr. Alastair Arbuthnot has collected data on X, Y, and Z in an observational study. He draws the diagram shown in figure 3, and fits the two regression equations suggested by the figure:

$$Y = \hat{a} + \hat{b}X + \text{error}, \qquad Z = \hat{c} + \hat{d}X + \hat{e}Y + \text{error}$$

Estimated coefficients are positive and significant. He is now trying to explain the findings to his colleague, Dr. Beverly Braithwaite.

Dr. A So you see, Dr. Braithwaite, if X goes up by one unit, then Y goes up by $\hat{b}$ units.

Dr. B Quite.

Dr. A Furthermore, if X goes up by one unit with Y held fixed, then Z goes up by $\hat{d}$ units. This is the direct effect of X on Z. ["Held fixed" means, kept the same; the "indirect effect" is through Y.]

Dr. B But Dr. Arbuthnot, you just told me that if X goes up by one unit, then Y will go up by $\hat{b}$ units.

Dr. A Moreover, if Y goes up by one unit with X held fixed, the change in Y makes Z go up by $\hat{e}$ units. The effect of Y on Z is $\hat{e}$.

Dr. B Dr. Arbuthnot, hello, why would Y go up unless X goes up? "Effects"? "Makes"? How did you get into causation?? And what about my first point?!?

Dr. Arbuthnot's explanation is not unusual. But Dr. Braithwaite has some good questions. Our objective in this section is to answer her, by developing a logically coherent set of assumptions which—if true—would justify Dr. Arbuthnot's data analysis and his interpretations. On the other hand, as we will see, Dr. Braithwaite has good reason for her skepticism.

At the back of his mind, Dr. Arbuthnot has two response schedules describing hypothetical experiments. In principle, these two experiments are unrelated to one another. But, to model the observational study, the experiments have to be linked in a special way. We will describe the two experiments first, and then explain how they are put together to model Dr. Arbuthnot's data.

(i) *First hypothetical experiment*. Treatment at level x is applied to a subject. A response Y is observed, corresponding to the level of treatment. There are two parameters, a and b, that describe the response. With no treatment ($x = 0$), the response level for each subject will be a, up to random error. All subjects are assumed to have the same value for a. Each additional unit of treatment adds b to the response. Again, b is the same for all subjects at all levels of x, by assumption. Thus, when treatment is applied at level x, the response Y is assumed to be

$$a + bx + \text{random error}. \tag{15}$$

For example, colleges send students with weak backgrounds to summer boot-camp with mathematics drill. In an evaluation study of such a program, x might be hours spent in math drill, and Y might be test scores.

(ii) *Second hypothetical experiment*. In the second experiment, there are two treatments and a response variable Z. There are two treatments because

there are two arrows leading into Z. The treatments are labeled X and Y in figure 3. Both treatments may be applied to a subject. In Experiment #1, Y was the response variable. But in Experiment #2, Y is one of the treatment variables: the response variable is Z.

There are three parameters, c, d, and e. With no treatment at all ($x = y = 0$), the response level for each subject will be c, up to random error. Each additional unit of treatment X adds d to the response. Likewise, each additional unit of treatment Y adds e to the response. (Here, e is a parameter not a residual vector.) The constancy of parameters across subjects and levels of treatment is an assumption. Thus, when the treatments are applied at levels x and y, the response Z is assumed to be

$$c + dx + ey + \text{random error}. \tag{16}$$

Three parameters are needed because it takes three parameters to specify the linear relationship (16), an intercept and two slopes.

Random errors in (15) and (16) are assumed to be independent from subject to subject, with a distribution that is constant across subjects: the expectation is zero and the variance is finite. The errors in (16) are assumed to be independent of the errors in (15). Equations (15) and (16) are *response schedules*: they summarize Dr. Arbuthnot's ideas about what would happen *if* he could do the experiments.

Linking the experiments. Dr. Arbuthnot collected the data on X, Y, Z in an observational study. He wants to use the observational data to figure out what would have happened *if* he could have intervened and manipulated the variables. There is a price to be paid.

To begin with, he has to assume the response schedules (15) and (16). He also has to assume that the X's are independent of the random errors in the two hypothetical experiments—"exogeneity." Thus, Dr. Arbuthnot is pretending that Nature randomized subjects to levels of X. If so, there is no need for experimental manipulation on his part, which is convenient. The exogeneity of X has a graphical representation: arrows come out of X in figure 3, but no arrows lead into X.

Dr. Arbuthnot also has to assume that Nature generates Y from X as if by substituting X into (15). Then Nature generates Z as if by substituting X and Y—the very same X that was the input to (15) and the Y that was the output from (15)—into (16). Using the output from (15) as an input to (16) is what links the two equations together.

Let's take another look at this linkage. In principle, the experiments described by the two response schedules are separable from one another. There is no a priori connection between the value of x in (15) and the value

of x in (16). There is no a priori connection between outputs from (15) and inputs to (16). However, to model his observational study, Dr. Arbuthnot links the equations "recursively." He assumes that one value of X is chosen and used as an input for both equations; that the Y generated from (15) is used as an input to (16); and there is no feedback from (16) to (15).

Given all these assumptions, the parameters a, b can be estimated by regression of Y on X. Likewise, c, d, e can be estimated by regression of Z on X and Y. Moreover, the regression estimates have legitimate causal interpretations. This is because causation is built into the response schedules (15) and (16). If causation were not assumed, causation would not be demonstrated by running the regressions.

One point of Dr. Arbuthnot's regressions is to estimate the direct effect of X on Z. The direct effect is d in (16). If X is increased by one unit with Y held fixed—i.e., kept at its old value—then Z is expected to go up by d units. This is shorthand for the mechanism in the second experiment. The response schedule (16) says what happens to Z when x and y are manipulated. In particular, y can be held at an old value while x is made to increase.

Dr. Arbuthnot imagines that he can keep the Y generated by Nature, while replacing X by $X + 1$. He just substitutes his values ($X + 1$ and Y) into the response schedule (16), getting

$$c + d(X + 1) + eY + \text{error} = \big(c + dX + eY + \text{error}\big) + d.$$

This is what Z would have been, if X had been increased by 1 unit with Y held fixed: Z would have been d units bigger.

Dr. Arbuthnot also wants to estimate the effect e of Y on Z. If Y is increased by one unit with X held fixed, then Z is expected to go up by e units. Dr. Arbuthnot thinks he can keep Nature's value for X, while replacing Y by $Y + 1$. He just substitutes X and $Y + 1$ into the response schedule (16), getting

$$c + dX + e(Y + 1) + \text{error} = \big(c + dX + eY + \text{error}\big) + e.$$

This is what Z would have been, if Y had been increased by 1 unit with X kept unchanged: Z would have been e units bigger. Of course, even Dr. Arbuthnot has to replace parameters by estimates. If $e = 0$—or could be 0 because $\hat{e}$ is statistically insignificant—then manipulating Y should not affect Z, and Y would not be a cause of Z after all. This is a qualitative inference. Again, the inference depends on the response schedule (16).

In short, Dr. Arbuthnot uses the observational data to estimate parameters. But when he interprets the results—for instance, when he talks about the

"effects" of X and Y on Z—he's thinking about the hypothetical experiments described by the response schedules (15)-(16), not about the observational data themselves. His causal interpretations depend on a rather subtle model. Among other things, the same response schedules, with the same parameter values, must apply (i) to the hypothetical experiments and (ii) to the observational data. In shorthand, the values of the parameters are stable under interventions.

To state the model more formally, we would index the subjects by a subscript i in the range from 1 to n. In this notation, X_i is the value of X for subject i. The level of treatment #1 is denoted by x, and $Y_{i,x}$ is the response for variable Y when treatment at level x is applied to subject i, as in (15). Similarly, $Z_{i,x,y}$ is the response for variable Z when treatment #1 at level x and treatment #2 at level y are applied to subject i, as in (16). The response schedules are interpreted causally.

- $Y_{i,x}$ is what Y_i would be if X_i were set to x by intervention.
- $Z_{i,x,y}$ is what Z_i would be if X_i were set to x and Y_i were set to y by intervention.

Figure 3 unpacks into two equations, which are more precise versions of (15) and (16), with subscripts for the subjects:

$$Y_{i,x} = a + bx + \delta_i, \tag{17}$$

$$Z_{i,x,y} = c + dx + ey + \epsilon_i. \tag{18}$$

The parameters a, b, c, d, e and the error terms δ_i, ϵ_i are not observed. The parameters are assumed to be the same for all subjects. There are assumptions about the error terms—the statistical component of the assumptions behind the path diagram:

(i) δ_i and ϵ_i are independent of each other within each subject i.
(ii) These error terms are independent across subjects i.
(iii) The distribution of δ_i is constant across subjects i; so is the distribution of ϵ_i. (However, δ_i and ϵ_i need not have the same distribution.)
(iv) δ_i and ϵ_i have expectation zero and finite variance.
(v) The X_i's are independent of the δ_i's and ϵ_i's, where X_i is the value of X for subject i in the observational study.

Assumption (v) says that Nature chooses X_i for us as if by randomization. In other words, the X_i's are "exogenous." By further assumption, Nature determines the response Y_i for subject i as if by substituting X_i into (17):

$$Y_i = Y_{i,X_i} = a + bX_i + \delta_i.$$

The rest of the response schedule— $Y_{i,x}$ for $x \neq X_i$—is not observed. After all, even in an experiment, subject i would be assigned to one level of treatment. The response at other levels would not be observed.

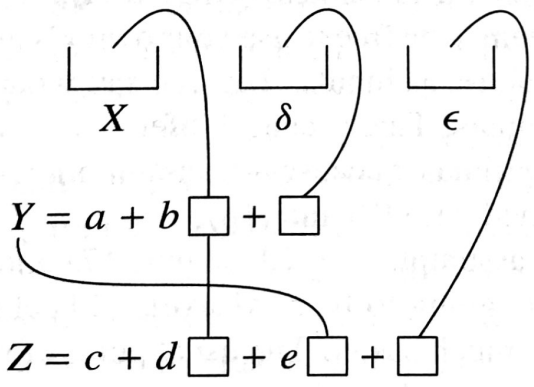
Similarly, we observe $Z_{i,x,y}$ only for $x = X_i$ and $y = Y_i$. The response for subject i is determined by Nature, as if by substituting X_i and Y_i into (18):

$$Z_i = Z_{i,X_i,Y_i} = c + dX_i + eY_i + \epsilon_i.$$

The rest of the response schedule remains unobserved, namely, the responses $Z_{i,x,y}$ for all the other possible values of x and y. Economists call the unobserved $Y_{i,x}$ and $Z_{i,x,y}$ *potential outcomes*. The model specifies unobservable response schedules, not just regression equations.

The model has another feature worth noticing: each subject's responses are determined by the levels of treatment for that subject only. Treatments applied to subject j do not affect the responses of subject i. For treating infectious diseases, this is not such a good model. (If one subject sneezes, another will catch the flu: stop the first sneeze, prevent the second flu.) There may be similar problems with social experiments, when subjects interact with each other.

Figure 4. The path diagram as a box model.



The box model in figure 4 illustrates the statistical assumptions. Independent random errors with constant distributions are represented as draws made at random with replacement from a box of potential errors (Freedman-Pisani-Purves 2007). Since the box remains the same from one draw to another, the probability distribution of one draw is the same as the distribution of any other. The distribution is constant. Furthermore, the outcome of one draw cannot affect the distribution of another. That is independence.

Figure 4 also shows how the two hypothetical causal mechanisms—response schedules (17) and (18)—are linked together to model the observational data. Let's take this apart and put it back together. We can think about each response schedule as a little machine, which accepts inputs and makes output. There are two of these machines at work.

- *First causal mechanism.* You feed an x—any x that you like—into machine #1. The output from the machine is $Y = a + bx$, plus a random draw from the δ-box.
- *Second causal mechanism.* You feed x and y—any x and y that you like—into machine #2. The output from the machine is $Z = c + dx + ey$, plus a random draw from the ϵ-box.
- *Linkage.* You don't feed anything into anything. Nature chooses X at random from the X-box, independent of the δ's and ϵ's. She puts X into machine #1, to generate a Y. She puts the same X—and the Y she just generated—into machine #2, to generate Z. You get to see (X, Y, Z) for each subject. This is Dr. Arbuthnot's model for his observational data.
- *Estimation.* You estimate a, b, c, d, e by OLS, from the observational data, namely, triples of observed values on (X, Y, Z) for many subjects.
- *Causal inference.* You can say what would happen if you could get your hands on the machines and put an x into machine #1. You can also say what would happen if you could put x and y into machine #2.

You never do touch the machines. (After all, these are purely theoretical entities.) Still, you seem to be free to use your own x's and y's, rather than the ones generated by Nature, as inputs. You can say what the machines would do *if* you chose the inputs. That is causal inference from observational data. Causal inference is legitimate because—by assumption—you know the social physics: response schedules (17) and (18).

What about the assumptions? Checking (17) and (18), which involve potential outcomes, is going to be hard work. Checking the statistical assumptions will not be much easier. The usual point of running regressions is to make causal inferences without doing real experiments. On the other hand, without the real experiments, the assumptions behind the models are going to be iffy. Inferences get made by ignoring the iffiness of the assumptions. That is the paradox of causal inference by regression, and a good reason for Dr. Braithwaite's skepticism.

Path models do not infer causation from association. Instead, path models *assume* causation through response schedules, and—using additional statistical assumptions—estimate causal effects from observational data. The statistical assumptions (independence, expectation zero, constant variance)

justify estimating coefficients by ordinary least squares. With large samples, standard errors, confidence intervals, and significance tests would follow. With small samples, the errors would have to follow a normal distribution in order to justify t-tests.

Evaluating the statistical models in chapters 1–6. Earlier in the book, we discussed several examples of causal inference by regression—Yule on poverty, Blau and Duncan on stratification, Gibson on McCarthyism. We found serious problems. These studies are among the strongest in the social sciences, in terms of clarity, interest, and data analysis. (Gibson, for example, won a prize for best paper of the year—and is still viewed as a landmark study in political behavior.) The problems are built into the assumptions behind the statistical models.

> Typically, a regression model assumes causation and uses the data to estimate the size of a causal effect. If the estimate isn't statistically significant, lack of causation is inferred. Estimation and significance testing require statistical assumptions. Therefore, you need to think about the assumptions—both causal and statistical—behind the models. If the assumptions don't hold, the conclusions don't follow from the statistics.

Selection vs intervention

The conditional expectation of Y given $X = x$ is the average of Y for subjects with $X = x$. (We ignore sampling error for now.) The response-schedule formalism connects two very different ideas of conditional expectation: (i) selecting the subjects with $X = x$, versus (ii) intervening to set $X = x$. The first is something you can actually do with observational data. The second would require manipulation. Response schedules crystallize the assumptions you need to get from selection to intervention. (*Intervention* means interrupting the natural flow of events in order to manipulate a variable, as in an experiment; the contrast is with passive observation.)

Selection is one thing, intervention is another.

Structural equations and stable parameters

In econometrics, "structural" equations describe causal relationships. Response schedules give a clearer meaning to this idea, and to the idea of "stability under intervention." The parameters in a path diagram, for instance, are defined through response schedules like (17) and (18), separately from

the data. By assumption, these parameters are constant across (i) subjects and (ii) levels of treatment. Moreover, (iii) the parameters stay the same whether you intervene or just observe the natural course of events. Response schedules bundle up these assumptions for us, along with similar assumptions on the error distributions. Assumption (iii) is sometimes called "constancy" or "invariance" or "stability under intervention."

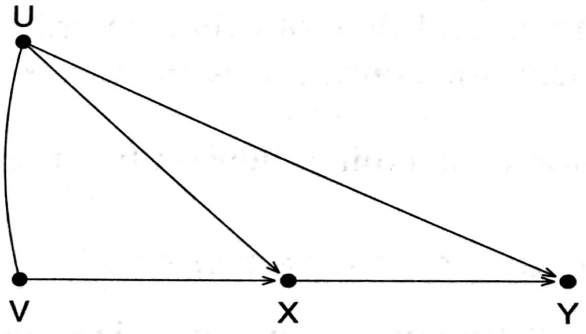
> Regression equations are structural, with parameters that are stable under intervention, when the equations derive from response schedules.

Ambiguity in notation

Look back at figure 3. In the observational study, there is an X_i for each subject i. In some contexts, X just means the X_i for a generic subject. In other contexts, X is the vector whose ith component is X_i. Often, X is the design matrix. This sort of ambiguity is commonplace. You have to pay attention to context, and figure out what is meant each time.

Exercise set E

1. In the path diagram below, free arrows are omitted. How many free arrows should there be, where do they go, and what do they mean? What does the curved line mean? The diagram represents some regression equations. What are the equations? the parameters? State the assumptions that would be needed to estimate the parameters by OLS. What data would you need? What additional assumptions would be needed to make causal inferences? Give an example of a qualitative causal inference that could be made from one of the equations. Give an example of a quantitative causal inference.



2. With the assumptions of this section, show that a regression of Y_i on X_i gives unbiased estimates, conditionally on the X_i's, of a and b in (17).

Show also that a regression of Z_i on X_i and Y_i gives unbiased estimates, conditionally on the X_i's and Y_i's, of c, d, and e in (18). Hints. What are the design matrices in the two regressions? Can you verify assumptions (4.2)–(4.5)? [Cross-references: (4.2) is equation (2) in chapter 4.]

3. Suppose you are only interested in the effects of X and Y on Z; you are not interested in the effect of X on Y. You are willing to assume the response schedule (18), with IID errors ϵ_i, independent of the X_i's and Y_i's. How would you estimate c, d, e? Do the estimates have a causal interpretation? Why?

4. True or false, and explain.
   (a) In figure 1, father's education has a direct influence on son's occupation.
   (b) In figure 1, father's education has an indirect influence on son's occupation through son's education.
   (c) In exercise 1, U has a direct influence on Y.
   (d) In exercise 1, V has a direct influence on Y.

5. Suppose Dr. Arbuthnot's models are correct; and in his data, $X_{77} = 12$, $Y_{77} = 2$, $Z_{77} = 29$.
   (a) How much bigger would Y_{77} have been, if Dr. Arbuthnot had intervened, setting X_{77} to 13?
   (b) How much bigger would Z_{77} have been, if Dr. Arbuthnot had intervened, setting X_{77} to 13 and Y_{77} to 5?

6. An investigator writes, "Statistical tests are a powerful tool for deciding whether effects are large." Do you agree or disagree? Discuss briefly.

6.6 Dummy variables

A "dummy variable" takes the value 0 or 1. Dummy variables are used to represent the effects of qualitative factors in a regression equation. Sometimes, dummies are even used to represent quantitative factors, in order to weaken linearity assumptions. (Dummy variables are also called "indicator" variables or "binary" variables; programmers call them "flags.")

Example. A company is accused of discriminating against female employees in determining salaries. The company counters that male employees have more job experience, which explains the salary differential. To explore that idea, a statistician might fit the equation

$$Y = a + b\,\text{MAN} + c\,\text{EXPERIENCE} + \text{error}.$$

Here, MAN is a dummy variable, taking the value 1 for men and 0 for women. EXPERIENCE would be years of job experience. A significant positive value for b would be taken as evidence of discrimination.

Objections could be raised to the analysis. For instance, why does EXPERIENCE have a linear effect? To meet that objection, some analysts would put in a quadratic term:

$$Y = a + b\,\text{MAN} + c\,\text{EXPERIENCE} + d\,\text{EXPERIENCE}^2 + \text{error}.$$

Others would break up EXPERIENCE into categories, e.g.,

category 1	under 5 years
category 2	5–10 years (inclusive)
category 3	over 10 years

Then dummies for the first two categories could go into the equation:

$$Y_i = a + b\,\text{MAN} + c_1\,\text{CAT}_1 + c_2\,\text{CAT}_2 + \text{error}.$$

For example, CAT_1 is 1 for all employees who have less than 5 years of experience, and 0 for the others. Don't put in all three dummies: if you do, the design matrix won't have full rank.

The coefficients are a little tricky to interpret. You have to look for the missing category, because effects are measured relative to the missing category. For MAN, it's easy. The baseline is women. The equation says that men earn b more than women, other things equal (experience). For CAT_1, it's less obvious. The baseline is the third category, over 10 years of experience. The equation says that employees in category 1 earn c_1 more than employees in category 3. Furthermore, employees in category 2 earn c_2 more than employees in category 3.

We expect c_1 and c_2 to be negative, because long-term employees get higher salaries. Similarly, we expect $c_1 < c_2$. Other things are held equal in these comparisons, namely, gender. (Saying that Harriet earns $-\$5{,}000$ more than Harry is a little perverse; ordinarily, we would talk about earning \$5,000 less: but this is statistics.)

Of course, the argument would continue. Why these categories? What about other variables? If people compete with each other for promotion, how can error terms be independent? And so forth. The point here was just to introduce the idea of dummy variables.

Types of variables

A *qualitative* or *categorical* variable is not numerical. Examples include gender and marital status, values for the latter being never-married, married,

widowed, divorced, separated. By contrast, a *quantitative* variable takes numerical values. If the possible values are few and relatively widely separated, the variable is *discrete*; otherwise, *continuous*. These are useful distinctions, but the boundaries are a little blurry. A dummy variable, for instance, can be seen as converting a categorical variable with two values into a numerical variable taking the values 0 and 1.

6.7 Discussion questions

Some of these questions cover material from previous chapters.

1. A regression of wife's educational level (years of schooling) on husband's educational level gives the equation

$$\text{WifeEdLevel} = 5.60 + 0.57 \times \text{HusbandEdLevel} + \text{residual}.$$

(Data are from the Current Population Survey in 2001.) If Mr. Wang's company sends him back to school for a year to catch up on the latest developments in his field, do you expect Mrs. Wang's educational level to go up by 0.57 years? If not, what does the 0.57 mean?

2. In equation (10), δ is a random error; there is a δ for each state. Gibson finds that $\hat{\beta}_1$ is statistically insignificant, while $\hat{\beta}_2$ is highly significant (two-tailed). Suppose that Gibson computed his P-values from the standard normal curve; the area under the curve between -2.58 and $+2.58$ is 0.99. True or false and explain—
   (a) The absolute value of $\hat{\beta}_2$ is more than 2.6 times its standard error.
   (b) The statistical model assumes that the random errors are independent across states.
   (c) However, the estimated standard errors are computed from the data.
   (d) The computation in (c) can be done whether or not the random errors are independent across states: the computation uses the tolerance scores and repression scores, but does not use the random errors themselves.
   (e) Therefore, Gibson's significance tests are fine, even if the random errors are dependent across states.

3. Timberlake and Williams (1984) offer a regression model to explain political oppression (PO) in terms of foreign investment (FI), energy development (EN), and civil liberties (CV). High values of PO correspond to authoritarian regimes that exclude most citizens from political participation. High values of CV indicate few civil liberties. Data were