Ordinary Least Squares

Hormoz Ramian

□

Introduction

This chapter attempts to accomplish two purposes. First, it is a self-contained introduction to linear regression with measurement error in the explanatory variables, suitable as a supplement to an ordinary regression course. Second, it is an introduction to the study of structural equation models in general. Without confronting the general formulation at first, the student will learn why structural equation models are important and see what can be done with them. Some of the ideas and definitions are repeated later in the book, so that the theoretical treatment of structural equation modeling does not depend much on this chapter. On the other hand, the material in this chapter will be used throughout the rest of the book as a source of examples. It should not be skipped by most readers.

1 Conditional Expectation

Consider the usual version of univariate multiple regression. For $i = 1, \ldots, n$,

$$Y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \cdots + \beta_{p-1} x_{i,p-1} + \epsilon_i,$$

where $\epsilon_1, \ldots \epsilon_n$ are independent random variables with expected value zero and common variance σ^2, and $x_{i,1}, \ldots x_{i,p-1}$ are fixed constants. For testing and constructing confidence intervals, $\epsilon_1, \ldots \epsilon_n$ are typically assumed normal.

Alternatively, the regression model may be written in matrix notation, as follows. Let

$$\mathbf{y} \ = \ \mathbf{X}\boldsymbol{\beta} \ + \ \boldsymbol{\epsilon}, \tag{1}$$

where $\mathbf{X}$ is an $n \times p$ matrix of known constants, $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown constants, and $\boldsymbol{\epsilon}$ is multivariate normal with mean zero and covariance matrix $\sigma^2 \mathbf{I}_n$; the variance $\sigma^2 > 0$ is a constant.

Now please take a step back and think about this model, rather than just accepting it without question. In particular, think about why the x variables should be constants. It's true that if they are constants then all the calculations are easier, but in the typical application of regression to observational[1] data, it makes more sense to view the explanatory variables as random variables rather than constants. Why? Because if you took repeated samples from the same population, the values of the explanatory variables would be different each time. Even for an experimental study with random assignment of cases (say dogs) to experimental conditions, suppose that the data are recorded in the order they were collected. Again, with high probability the values of the explanatory variables would be different each time.

So, why are the x variables a set of constants in the formal model? One response is that the regression model is a conditional one, and all the conclusions hold conditionally upon the values of the explanatory variables. This is technically correct, but consider the reaction of a zoologist using multiple regression, assuming he or she really appreciated the point. She would be horrified

[1] *Observational* data are just observed, rather than being controlled by the investigator. For example, the average number of minutes per day spent outside could be recorded for a sample of dogs. In contrast to observational data are *experimental* data, in which the values of the variable in question are controlled by the investigator. For example, dogs could be randomly assigned to several different values of the variable "time outside." Based on this, some dogs would always be taken for longer walks than others.

at the idea that the conclusions of the study would be limited to this particular configuration of explanatory variable values. No! The sample was taken from a population, and the conclusions should apply to that population, not to the subset of the population with these particular values of the explanatory variables.

At this point you might be a bit puzzled and perhaps uneasy, realizing that you have accepted something uncritically from authorities you trusted, even though it seems to be full of holes. In fact, everything is okay this time. It is perfectly all right to apply a conditional regression model even though the predictors are clearly random. But it's not so very obvious why it's all right, or in what sense it's all right. This section will give the missing details. These are skipped in every regression textbook I have seen; I'm not sure why.

Unbiased Estimation Under the standard conditional regression model (1), it is straightforward to show that the vector of least-squares regression coefficients $\widehat{\boldsymbol{\beta}}$ is unbiased for $\boldsymbol{\beta}$ (both of these are $p \times 1$ vectors). This means that it's unbiased *conditionally* upon $\mathbf{X} = \mathbf{x}$. In symbols,

$$E\{\widehat{\boldsymbol{\beta}}|\mathbf{X} = \mathbf{x}\} = \boldsymbol{\beta}.$$

This applies to every fixed $\mathbf{x}$ matrix with linearly independent columns, a condition that is necessary and sufficient for $\widehat{\boldsymbol{\beta}}$ to exist. Assume that the joint probability distribution of the random matrix $\mathbf{X}$ assigns zero probability to matrices with linearly dependent columns (which is the case for continuous distributions). Using the double expectation formula $E\{Y\} = E\{E\{Y|X\}\}$,

$$E\{\widehat{\boldsymbol{\beta}}\} = E\{E\{\widehat{\boldsymbol{\beta}}|\mathbf{X}\}\} = E\{\boldsymbol{\beta}\} = \boldsymbol{\beta},$$

since the expected value of a constant is just the constant. This means that *estimates of the regression coefficients from the conditional model are still unbiased, even when the explanatory variables are random.*

The following calculation might make the double expectation a bit clearer. The outer expected value is with respect to the joint probability distribution of the explanatory variable values – all n vectors of them; think of the $n \times p$ matrix $\mathbf{X}$. To avoid unfamiliar notation, suppose they are all continuous, with joint density $f(\mathbf{x})$. Then

$$\begin{aligned}
E\{\widehat{\boldsymbol{\beta}}\} &= E\{E\{\widehat{\boldsymbol{\beta}}|\mathbf{X}\}\} \\
&= \int \cdots \int E\{\widehat{\boldsymbol{\beta}}|\mathbf{X} = \mathbf{x}\}\, f(\mathbf{x})\, d\mathbf{x} \\
&= \int \cdots \int \boldsymbol{\beta}\, f(\mathbf{x})\, d\mathbf{x} \\
&= \boldsymbol{\beta} \int \cdots \int f(\mathbf{x})\, d\mathbf{x} \\
&= \boldsymbol{\beta} \cdot 1 = \boldsymbol{\beta}.
\end{aligned}$$

Size α Tests Suppose Model (1) is conditionally correct, and we plan to use an F test. Conditionally upon the x values, the F statistic has an F distribution when the null hypothesis is true, but unconditionally it does not. Rather, its probability distribution is a *mixture* of F distributions, with

$$Pr\{F \in A\} = \int \cdots \int Pr\{F \in A|\mathbf{X} = \mathbf{x}\} f(\mathbf{x})\, d\mathbf{x}.$$

If the null hypothesis is true and the set A is the critical region for an exact size α F-test, then $Pr\{F \in A | \mathbf{X} = \mathbf{x}\} = \alpha$ for every fixed set of explanatory variable values $\mathbf{x}$. In that case,

$$\begin{aligned} Pr\{F \in A\} &= \int \cdots \int \alpha f(\mathbf{x}) \, d\mathbf{x} \\ &= \alpha \int \cdots \int f(\mathbf{x}) \, d\mathbf{x} \\ &= \alpha. \end{aligned} \tag{2}$$

Thus, the so-called F-test has the correct Type I error rate when the explanatory variables are random (assuming the model is conditionally correct), even though the test statistic does not have an F distribution.

It might be objected that if the explanatory variables are random and we assume they are fixed, the resulting estimators and tests might be of generally low quality, even though the estimators are unbiased and the tests have the right Type I error rate. Now we will see that given a fairly reasonable set of assumptions, this objection has no merit.

Denoting the explanatory variable values by $\mathbf{X}$ and the response variable values by $\mathbf{Y}$, suppose the joint distribution of $\mathbf{X}$ and $\mathbf{Y}$ has the following structure. The distribution of $\mathbf{X}$ depends on a parameter vector $\boldsymbol{\theta}_1$. Conditionally on $\mathbf{X} = \mathbf{x}$, the distribution of $\mathbf{Y}$ depends on a parameter vector $\boldsymbol{\theta}_2$, and $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ are *not functionally related*. For a standard regression model this means that the distribution of the explanatory variables does not depend upon the values of $\boldsymbol{\beta}$ or σ^2 in any way. This is surely not too hard to believe.

Please notice that the model just described is not at all limited to linear regression. It is very general, covering almost any conceivable regression-like method including logistic regression and other forms of non-linear regression, generalized linear models and the like.

Because likelihoods are just joint densities or probability mass functions viewed as functions of the parameter, the notation of Appendix **??** may be stretched just a little bit to write the likelihood function for the unconditional model (with $\mathbf{X}$ random) in terms of conditional densities as

$$\begin{aligned} L(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \mathbf{x}, \mathbf{y}) &= f_{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2}(\mathbf{x}, \mathbf{y}) \\ &= f_{\boldsymbol{\theta}_2}(\mathbf{y}|\mathbf{x}) \, f_{\boldsymbol{\theta}_1}(\mathbf{x}) \\ &= L_2(\boldsymbol{\theta}_2, \mathbf{x}, \mathbf{y}) \, L_1(\boldsymbol{\theta}_1, \mathbf{x}) \end{aligned} \tag{3}$$

Now, take the log and partially differentiate with respect to the elements of $\boldsymbol{\theta}_2$. The marginal likelihood $L_1(\boldsymbol{\theta}_1, \mathbf{x})$ disappears, and $\widehat{\boldsymbol{\theta}}_2$ is exactly what it would have been for a conditional model.

In this setting, likelihood ratio tests are also identical under conditional and unconditional models. Suppose the null hypothesis concerns $\boldsymbol{\theta}_2$, which is most natural. Note that the structure of (3) guarantees that the MLE of $\boldsymbol{\theta}_1$ is the same under the null and alternative hypotheses. Letting $\widehat{\boldsymbol{\theta}}_{0,2}$ denote the restricted MLE of $\boldsymbol{\theta}_2$ under H_0, the likelihood ratio for the unconditional model is

$$\begin{aligned} \lambda &= \frac{L_2(\widehat{\boldsymbol{\theta}}_{0,2}, \mathbf{x}, \mathbf{y}) \, L_1(\widehat{\boldsymbol{\theta}}_1, \mathbf{x})}{L_2(\widehat{\boldsymbol{\theta}}_2, \mathbf{x}, \mathbf{y}) \, L_1(\widehat{\boldsymbol{\theta}}_1, \mathbf{x})} \\ &= \frac{L_2(\widehat{\boldsymbol{\theta}}_{0,2}, \mathbf{x}, \mathbf{y})}{L_2(\widehat{\boldsymbol{\theta}}_2, \mathbf{x}, \mathbf{y})}, \end{aligned}$$

which again is exactly what it would have been under a conditional model. While this holds only because the likelihood has the nice structure in (3), it's a fairly reasonable set of assumptions.

Thus in terms of both estimation and hypothesis testing, the fact that explanatory variables are usually random variables presents no difficulty, regardless of what the distribution of those explanatory variables may be. On the contrary, the conditional nature of the usual regression model is a virtue. Notice that in all the calculations above, the joint distribution of the explanatory variables is written in a very general way. It really doesn't matter what it is, because it disappears.

So one might say that with respect to the explanatory variables, the usual linear regression model is distribution free.

2 Covariance and Relationship

In spite of the virtues of the conditional regression model, in this book we will focus on *unconditional* regression models, in which the explanatory variables are random. The reason is that ultimately, the explanatory variables themselves may be influenced by other variables. The easiest way to represent this is to admit from the outset that they are random variables.

Most of the models we will consider are linear in the explanatory variables as well as the regression parameters, and so relationships between explanatory variables and response variables are represented by covariances. To clarify this fundamental point, first note that saying two random variables are "related" really just means that they are not independent. A non-zero covariance implies lack of independence, and therefore a relationship of some kind between the variables. Furthermore, if the random variables in question are normally distributed (a common and very useful model), zero covariance is exactly the same thing as independence.

More generally, consider two random variables X and Y whose joint distribution might not be bivariate normal. Suppose there is a tendency for higher values of X to go with higher values of Y, and for lower values of X to go with lower values of Y. This idea of a "positive" relationship is pictured in the left panel of Figure **??**. Since the probability of an (x, y) pair is roughly proportional to the height of the surface, a large sample of points will be most dense where the surface is highest[2]. On a scatterplot, the best-fitting line will have a positive slope. The right panel of Figure **??** shows a negative relationship. There, the best-fitting line will have a negative slope.

The word "covariance" suggests that it is a measure of how X and Y vary together. To see that positive relationships yield positive covariances and negative relationships yield negative covariances, look at Figure **??**.
Figure **??** shows contour plots of the densities in Figure **??**. Imagine you are looking down at a density from directly above, and that the density has been cut into slices that are parallel with the x, y plane. The ellipses are the cut marks. The outer ellipse is lowest, the next one in is a bit higher, and so on. All the points on an ellipse (contour) are at the same height. It's like a topographic map of a mountainous region, except that the contours on maps are not so regular.

The definition of covariance is

$$Cov(X,Y) = E\left\{(X-\mu_x)(Y-\mu_y)\right\} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x-\mu_x)(y-\mu_y)f(x,y)\,dx\,dy$$

In the left panel of Figure **??**, more of the probability is in the upper right and lower left, and that is where $(x-\mu_x)(y-\mu_y)$ is positive. The positive volume in these regions is greater than the negative volume in the upper left and lower right, so that the integral is positive. In the right-hand panel the opposite situation occurs, and the covariance is negative. The pictures are just of one example, but the rule is general. Positive covariances reflect positive relationships and negative covariances reflect negative relationships.

3 The Centering Rule

Since relationships between variables are represented by covariances, there will be a lot of variance and covariance calculations in this book. Anything that makes them easier will be very welcome. To make the presentation self-contained, this section re-states the *Centering Rule* given on page **??**

[2]Presumably this is why it's called a probability *density* function.

of Appendix **??**. The idea is that because adding or subtracting constants has no effect on variances and covariances, it is okay to replace random variables by "centered" versions in which the expected value has been subtracted off, and then calculate variances and covariances. Suppose $E(\mathbf{X}) = \boldsymbol{\mu}_x$. Using a non-standard but useful notation, the centered version of a random vector will be written $\overset{c}{\mathbf{X}} = \mathbf{X} - \boldsymbol{\mu}_x$, so that $E(\overset{c}{\mathbf{X}}) = \mathbf{0}$, $cov(\mathbf{X}) = E(\overset{c}{\mathbf{X}}\overset{c}{\mathbf{X}}^\top)$, and $cov(\mathbf{X}, \mathbf{Y}) = E(\overset{c}{\mathbf{X}}\overset{c}{\mathbf{Y}}^\top)$.

Consider the linear combination $\mathbf{L} = \mathbf{A}_1\mathbf{X}_1 + \cdots + \mathbf{A}_m\mathbf{X}_m + \mathbf{b}$. The centered version of $\mathbf{L}$ is obtained by dropping the constant vector $\mathbf{b}$ and centering all the variables. That is, $\overset{c}{\mathbf{L}} = \mathbf{A}_1 \overset{c}{\mathbf{X}}_1 + \cdots + \mathbf{A}_m \overset{c}{\mathbf{X}}_m$. Here is a full statement of the Centering Rule.

Let $\mathbf{X}_1, \ldots, \mathbf{X}_m$ and $\mathbf{Y}_1, \ldots, \mathbf{Y}_k$ be random vectors, and

$$\begin{aligned}
\mathbf{L}_1 &= \mathbf{A}_1\mathbf{X}_1 + \cdots + \mathbf{A}_m\mathbf{X}_m + \mathbf{b}. \text{ We have} \\
\overset{c}{\mathbf{L}}_1 &= \mathbf{A}_1 \overset{c}{\mathbf{X}}_1 + \cdots + \mathbf{A}_m \overset{c}{\mathbf{X}}_m, \text{ where} \\
\overset{c}{\mathbf{X}}_j &= \mathbf{X}_j - E(\mathbf{X}_j) \text{ for } j = 1, \ldots, m.
\end{aligned}$$

Similarly,

$$\begin{aligned}
\mathbf{L}_2 &= \mathbf{C}_1\mathbf{Y}_1 + \cdots + \mathbf{C}_k\mathbf{Y}_k + \mathbf{d} \text{ and} \\
\overset{c}{\mathbf{L}}_2 &= \mathbf{C}_1 \overset{c}{\mathbf{Y}}_1 + \cdots + \mathbf{C}_k \overset{c}{\mathbf{Y}}_k, \text{ where} \\
\overset{c}{\mathbf{Y}}_j &= \mathbf{Y}_j - E(\mathbf{Y}_j) \text{ for } j = 1, \ldots, k.
\end{aligned}$$

Then $cov(\mathbf{L}_1) = E(\overset{c}{\mathbf{L}}_1\overset{c}{\mathbf{L}}_1^\top)$, $cov(\mathbf{L}_2) = E(\overset{c}{\mathbf{L}}_2\overset{c}{\mathbf{L}}_2^\top)$, and $cov(\mathbf{L}_1, \mathbf{L}_2) = E(\overset{c}{\mathbf{L}}_1\overset{c}{\mathbf{L}}_2^\top)$.

As an example, consider the calculation of $cov(\mathbf{X} + \mathbf{Y})$.

$$\begin{aligned}
cov(\mathbf{X} + \mathbf{Y}) &= cov(\overset{c}{\mathbf{X}} + \overset{c}{\mathbf{Y}}) \\
&= E(\overset{c}{\mathbf{X}} + \overset{c}{\mathbf{Y}})(\overset{c}{\mathbf{X}} + \overset{c}{\mathbf{Y}})^\top \\
&= E(\overset{c}{\mathbf{X}} + \overset{c}{\mathbf{Y}})(\overset{c}{\mathbf{X}}^\top + \overset{c}{\mathbf{Y}}^\top) \\
&= E(\overset{c}{\mathbf{X}}\overset{c}{\mathbf{X}}^\top) + E(\overset{c}{\mathbf{Y}}\overset{c}{\mathbf{Y}}^\top) + E(\overset{c}{\mathbf{X}}\overset{c}{\mathbf{Y}}^\top) + E(\overset{c}{\mathbf{Y}}\overset{c}{\mathbf{X}}^\top) \\
&= cov(\mathbf{X}) + cov(\mathbf{Y}) + cov(\mathbf{X}, \mathbf{Y}) + cov(\mathbf{Y}, \mathbf{X})
\end{aligned}$$

This is the matrix version of the formula $Var(X + Y) = Var(X) + Var(Y) + 2Cov(X, Y)$. Note that if $\mathbf{X}$ and $\mathbf{Y}$ are not 1×1, $cov(\mathbf{X}, \mathbf{Y})$ is not in general equal to $cov(\mathbf{Y}, \mathbf{X})$, though $cov(\mathbf{Y}, \mathbf{X}) = cov(\mathbf{X}, \mathbf{Y})^\top$.

The centering rule is useful in scalar variance-covariance calculations too. For example, let $X_1, \ldots, X_n$ be a random sample from a distribution with mean μ and variance σ^2, and consider the task of showing that $Cov(\overline{X}, X_j - \overline{X}) = 0$, which is the key to proving the independence of $\overline{X}$ and S^2 for the normal distribution, and the gateway to the t distribution. Since $\overline{X}$ and $X_j - \overline{X}$ are both linear combinations,

$$
\begin{aligned}
Cov(\overline{X}, X_j - \overline{X}) &= E\left(\overline{\overset{c}{X}}\,(\overset{c}{X}_j - \overline{\overset{c}{X}})\right) \\
&= E\left(\overset{c}{X}_j \overline{\overset{c}{X}}\right) - E\left(\overline{\overset{c}{X}}^2\right) \\
&= E\left(\overset{c}{X}_j \frac{1}{n}\sum_{i=1}^{n} \overset{c}{X}_i\right) - Var\left(\overline{X}\right) \\
&= E\left(\frac{1}{n}\sum_{i=1}^{n} \overset{c}{X}_i \overset{c}{X}_j\right) - Var\left(\overline{X}\right) \\
&= \frac{1}{n}\sum_{i=1}^{n} E\left(\overset{c}{X}_i \overset{c}{X}_j\right) - \frac{\sigma^2}{n} \\
&= \frac{1}{n} E\left(\overset{c}{X}_j^2\right) + \frac{1}{n}\sum_{i\neq j} E\left(\overset{c}{X}_i\right) E\left(\overset{c}{X}_j\right) - \frac{\sigma^2}{n} \\
&= \frac{1}{n} Var(X_j) + 0 - \frac{\sigma^2}{n} \\
&= \frac{\sigma^2}{n} - \frac{\sigma^2}{n} = 0
\end{aligned}
$$

This valuable calculation is long because all the details are shown. It is significantly messier without centering.

4 Unconditional regression with observed variables

Example 4.1

Suppose that the covariance between two random variables arises from a regression. Independently for $i = 1, \ldots, n$, let

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \tag{4}$$

where

- X_i is has expected value μ_x and variance $\phi > 0$
- ϵ_i has expected value zero and variance $\sigma^2 > 0$
- X_i and ϵ_i are independent.

The pairs (X_i, Y_i) have a joint distribution that is unspecified, except for the expected value

$$E\begin{pmatrix} X_i \\ Y_i \end{pmatrix} = \boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} = \begin{pmatrix} \mu_x \\ \beta_0 + \beta_1 \mu_x \end{pmatrix},$$

and variance-covariance matrix

$$cov\begin{pmatrix} X_i \\ Y_i \end{pmatrix} = \boldsymbol{\Sigma} = [\sigma_{i,j}] = \begin{pmatrix} \phi & \beta_1 \phi \\ \beta_1 \phi & \beta_1^2 \phi + \sigma^2 \end{pmatrix}.$$

The Centering Rule of Section 3 (see Page 4) is useful for calculating the covariance between the

explanatory and response variables.

$$\begin{aligned} Cov(X_i, Y_i) &= Cov(\overset{c}{X}_i, \overset{c}{Y}_i) \\ &= E(\overset{c}{X}_i \overset{c}{Y}_i) \\ &= E\left(\overset{c}{X}_i \, (\beta_1 \, \overset{c}{X}_i + \epsilon_i)\right) \\ &= \beta_1 E(\overset{c}{X}_i^2) + E(\overset{c}{X}_i) E(\epsilon_i) \\ &= \beta_1 \phi \end{aligned}$$

Since ϕ is a variance it is greater than zero, the sign of the covariance is the sign of the regression coefficient. Positive regression coefficients produce positive relationships, negative regression coefficients produce negative relationships, and zero corresponds to no relationship as measured by the covariance.

While the sign of the covariance (and hence the direction of the relationship) is determined by β_1, the magnitude of the covariance is jointly determined by the magnitude of β_1 and the magnitude of ϕ, the variance of X_i. Consequently the covariance of X_i and Y_i depends on the scale of measurement of X_i. If X_i is measured in centimeters instead of meters, its variance is $100^2 = 10,000$ times as great, and $Cov(X_i, Y_i)$ is ten thousand times as great, as well. This makes raw covariances difficult to interpret, except for the sign.

A solution is to put the variables on a standard common scale by looking at correlations instead of covariances. Denoting the correlation of any two random variables X and Y by Greek letter "rho," which is a common notation,

$$\begin{aligned} \rho_{xy} &= \frac{Cov(X,Y)}{SD(X)SD(Y)} \\ &= \frac{E\left\{(X-\mu_x)(Y-\mu_y)\right\}}{\sqrt{Var(X)}\sqrt{Var(Y)}} \\ &= E\left\{\left(\frac{X-\mu_x}{\sigma_x}\right)\left(\frac{Y-\mu_y}{\sigma_y}\right)\right\}. \end{aligned} \tag{5}$$

That is, the correlation between two random variables is the covariance between versions of the variables that have been standardized to have mean zero and variance one. Using (5), the correlation for the regression example is

$$\begin{aligned} \rho &= \frac{\beta_1 \phi}{\sqrt{\phi}\sqrt{\beta_1^2\phi + \sigma^2}} \\ &= \frac{\beta_1 \sqrt{\phi}}{\sqrt{\beta_1^2\phi + \sigma^2}}. \end{aligned} \tag{6}$$

This may not look like much, but consider the following. In any regression, the response variable is likely to represent the phenomenon of primary interest, and explaining why it varies from unit to unit is an important scientific goal. For example, if Y_i is academic performance, we want to know why some students do better than others. If Y_i is the crime rate in neighbourhood i, we want to know why there is more crime in some neighbourhood than in others. If there were no variation in some phenomenon (it's hard to think of examples) there might still be something to explain, but it would not be a statistical question. Because X_i and ϵ_i are independent,

$$\begin{aligned} Var(Y_i) &= Var(\beta_1 X_i + \epsilon_i) \\ &= \beta_1^2 Var(X_i) + Var(\epsilon_i) \\ &= \beta_1^2 \phi + \sigma^2. \end{aligned}$$

Thus the variance of Y_i is separated into two parts[3], the part that comes from X_i and the part that comes from ϵ_i. The part that comes from X_i is $\beta_1^2\phi$, and the part that comes from ϵ_i (that

[3]The word "analysis" means splitting into parts, so this is literally analysis of variance.

is, everything else) is σ^2. From (6) the *squared* correlation between X_i and Y_i is

$$\rho^2 = \frac{\beta_1^2\phi}{\beta_1^2\phi + \sigma^2}, \tag{7}$$

the proportion of the variance in Y_i that comes from X_i. This quantity does not depend on the scale of X_i or the scale of Y_i, because both variables are standardized.

Now consider multiple regression. In ordinary multiple regression (the conditional model), one speaks of the relationship between and explanatory variable and the response variable "controlling" for other variables in the model[4]. This really refers to the conditional expectation of Y as a function of x_j for fixed values of the other x variables, say in the sense of a partial derivative. In unconditional regression with random explanatory variables one talks about it in the same way, but the technical version is a bit different and perhaps easier to understand.

Example 4.2

Independently for $i = 1, \ldots, n$, let $Y_i = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \epsilon_i$, where $E(X_{i,1}) = \mu_1$, $E(X_{i,2}) = \mu_2$, $E(\epsilon_i) = 0$, $Var(\epsilon_i) = \sigma^2$, ϵ_i is independent of both $X_{i,1}$ and $X_{i,2}$, and

$$cov\begin{pmatrix} X_{i,1} \\ X_{i,2} \end{pmatrix} = \begin{pmatrix} \phi_{11} & \phi_{12} \\ \phi_{12} & \phi_{22} \end{pmatrix}.$$

Figure **??** shows a path diagram for this model. The explanatory and response variables are all observed, so they are enclosed in boxes. The double-headed curved arrow between the explanatory variables represents a possibly non-zero covariance. This covariance might arises from interesting and important processes including common influences on the X variables, but those processes are not part of the model. Curved double-headed arrows represent *unanalyzed* covariances between explanatory variables.

The straight arrows from the explanatory to response variables represent direct influence, or at least that we are interested in predicting y from x rather than the other way around. There is a regression coefficient β on each straight arrow, and a covariance ϕ_{12} on the curved double-headed arrow.

For this model, the covariance of $X_{i,1}$ and Y_i is

$$\begin{aligned} Cov(X_{i,1}, Y_i) &= E(\overset{c}{X}_{i,1}\overset{c}{Y}_i) \\ &= E\left(\overset{c}{X}_{i,1}\,(\beta_1\,\overset{c}{X}_{i,1} + \beta_2\,\overset{c}{X}_{i,2} + \epsilon_i)\right) \\ &= \beta_1 E(\overset{c}{X}{}^2_{i,1}) + \beta_2 E(\overset{c}{X}_{i,1}\overset{c}{X}_{i,2}) + E(\overset{c}{X}_{i,1})E(\epsilon_i) \\ &= \beta_1\phi_{11} + \beta_2\phi_{12}. \end{aligned}$$

This means that the relationship between X_1 and Y has two sources. One is the direct link from X_1 to Y through the straight arrow represented by β_1, and the other is through the curved arrow between X_1 and X_2 and then through the straight arrow linking X_2 to Y. Even if $\beta_1 = 0$, there still will be a relationship provided that X_1 is related to X_2 and X_2 is related to Y[5]. Furthermore, $\beta_2\phi_{12}$ may overwhelm $\beta_1\phi_{11}$, so that the covariance between X_1 and Y may be positive even though β_1 is negative.

All this is true of the unconditional relationship between X_1 and Y, but what if you "control" for X_2 by holding it constant at some fixed value? In the classical conditional regression model the meaning of holding a variable constant is a bit subtle, because the explanatory variable values are already constants. For unconditional regression the interpretation is more straightforward. When the explanatory variables are all random, the relationship between X_1 and Y controlling

[4]One can also speak of "correcting" for the other variables, or "holding them constant," or "allowing" for them, or "taking them into account." These are all ways of saying exactly the same thing.

[5]Yes, body weight may be positively related to income because men are bigger on average and they tend to make more money for the same work.

for X_2 simply refers to a conditional distribution — the joint distribution of X_1 and Y given $X_2 = x_2$. In this case the regression equation is

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_{i,1} + \beta_2 x_{i,2} + \epsilon_i \\ &= (\beta_0 + \beta_2 x_{i,2}) + \beta_1 X_{i,1} + \epsilon_i \\ &= \beta_0' + \beta_1 X_{i,1} + \epsilon_i \end{aligned}$$

The constant is simply absorbed into the intercept. It's a little strange in that the intercept is potentially different for $i = 1, \ldots, n$, but that doesn't affect the covariance. Following the calculations in Example 4.1, the conditional covariance between $X_{i,1}$ and Y_i is $\beta_1\phi_{11}$. Thus to test whether X_1 is connected to Y controlling for X_2 (or correcting for it, or allowing for it or some such term), it is appropriate to test $H_0 : \beta_1 = 0$. If the null hypothesis is rejected, the sign of the estimated regression coefficient guides your conclusion as to whether the conditional relationship is positive or negative. These considerations extend immediately to multiple regression.

In terms of interpreting the regression coefficients, it is helpful to decompose (analyze) the variance of Y_i.

$$\begin{aligned} Var(Y_i) &= Var(\beta_1 X_{i,1} + \beta_2 X_{i,2} + \epsilon_i) \\ &= \beta_1^2\phi_{11} + \beta_2^2\phi_{22} + 2\beta_1\beta_2\phi_{12} + \sigma^2 \end{aligned}$$

The explanatory variables contribute to the variance of the response individually through their variances and squared regression coefficients, and also jointly through their regression coefficients and their covariance. This joint effect is not an interaction in the ordinary sense of the term; the model of Example 4.2 has no product term. The null hypothesis $H_0 : \beta_1 = 0$ means that X_1 does not contribute at all to the variance of Y, either directly or through its covariance with X_2.

Estimation

Here is some useful terminology, repeated from Appendix **??**.

Definition 4.1 Moments *of a distribution are quantities such* $E(X)$, $E(Y^2)$, $Var(X)$, $E(X^2Y^2)$, $Cov(X, Y)$, *and so on.*

Definition 4.2 Moment structure equations *are a set of equations expressing moments of the distribution of the data in terms of the model parameters. If the moments involved are limited to variances and covariances, the moment structure equations are called* covariance structure equations.

For the simple (one explanatory variable) regression model of Example 4.1, the moments are the elements of the mean vector $\boldsymbol{\mu} = E\begin{pmatrix} X_i \\ Y_i \end{pmatrix}$, and the unique elements of the covariance matrix $\boldsymbol{\Sigma} = cov\begin{pmatrix} X_i \\ Y_i \end{pmatrix}$. The moments structure equations are

$$\begin{aligned} \mu_1 &= \mu_x \\ \mu_2 &= \beta_0 + \beta_1\mu_x \\ \sigma_{1,1} &= \phi \\ \sigma_{1,2} &= \beta_1\phi \\ \sigma_{2,2} &= \beta_1^2\phi + \psi. \end{aligned} \tag{8}$$

In this model, the parameters are μ_x, ϕ, β_0, β_1, ψ, and also the unknown distribution functions of X_i and ϵ_i. Our interest is in the Greek-letter parameters, especially β_0 and β_1. Method of Moments estimates (See Section **??** in Appendix **??**) can be obtained by solving the moment structure equations (8) for the unknown parameters and putting hats on the result. The moment

structure equations form a system of 5 equations in five unknowns, and may be readily be solved to yield

$$\begin{aligned}
\beta_0 &= \mu_2 - \frac{\sigma_{1,2}}{\sigma_{1,1}}\mu_1 \\
\mu_x &= \mu_1 \\
\phi &= \sigma_{1,1} \\
\beta_1 &= \frac{\sigma_{1,2}}{\sigma_{1,1}} \\
\psi &= \sigma_{2,2} - \frac{\sigma_{1,2}^2}{\sigma_{1,1}}.
\end{aligned} \tag{9}$$

Thus, even though the distributions of X_i and ϵ_i are unknown, we have nice consistent estimators of the interesting part of the unknown parameter. Putting hats on the parameters in Expression 9,

$$\begin{aligned}
\widehat{\beta}_0 &= \overline{y} - \frac{\widehat{\sigma}_{1,2}}{\widehat{\sigma}_{1,1}}\overline{x} \\
\widehat{\mu}_x &= \widehat{\mu}_1 = \overline{x} \\
\widehat{\phi} &= \widehat{\sigma}_{1,1} \\
\widehat{\beta}_1 &= \frac{\widehat{\sigma}_{1,2}}{\widehat{\sigma}_{1,1}} \\
\widehat{\psi} &= \widehat{\sigma}_{2,2} - \frac{\widehat{\sigma}_{1,2}^2}{\widehat{\sigma}_{1,1}}.
\end{aligned}$$

It is very standard to assume that X_i and ϵ_i are normally distributed. In this case, the existence of the solution (9) tells us that the parameters of the normal version of this regression model stand in a one-to-one-relationship with the mean and covariance matrix of the bivariate normal distribution posessed by the observable data. In fact, the two sets of parameter values are 100% equivalent; they are just different ways of expressing the same thing. For some purposes, the parameterization represented by the regression model may be more informative.

Furthermore, the Invariance Principle of maximum likelihood estimation (see Section **??** in Appendix **??**) says that the MLE of a one-to-one function is just that function of the MLE. So, the Method of Moments estimates are also the Maximum Likelihood estimates in this case.

The calculations just shown are important, because they are an easy, clear example of what will be necessary again and again throughout the course. Here is the process:

- Calculate the moments of the distribution (usually means, variances and covariances) in terms of the model parameters, obtaining a system of moment structure equations.
- Solve the moment structure equations for the parameters, expressing the parameters in terms of the moments.

When the second step is successful, putting hats on all the parameters in the solution yields Method of Moments estimators, even when these do not correspond to the MLEs[6].

It turns out that for any reasonable models, a unique solution for the parameters is mathematically impossible. In such cases, successful parameter estimation by any method is impossible as well. It is vitally important to verify the *possibility* of successful parameter estimation before trying it for a given data set, and verification consists of a process like the one you have just seen. Of course it is no surprise that estimating the parameters of a regression model is technically possible.

[6]When there are the same number of moment structure equations and a unique sulution for the parametrers exists, the Mothod of Moments estimators and MLEs coincide. When there are more equations than parameters they no longer coincide in general, but still the process of "putting hats on everything" yields Method of Moments estimators.

Because the process is so important, let us take a look at the extension to multivariate multiple regression — that is, to linear regression with multiple explanatory variables and multiple response variables. This will illustrate the matrix versions of the calculations.

Example 4.3

Independently for $i = 1, \ldots, n$, let

$$\mathbf{Y}_i = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 \mathbf{X}_i + \boldsymbol{\epsilon}_i \tag{10}$$

where

$\mathbf{Y}_i$ is an $q \times 1$ random vector of observable response variables, so the regression can be multivariate; there are q response variables.

$\boldsymbol{\beta}_0$ is a $q \times 1$ vector of unknown constants, the intercepts for the q regression equations. There is one for each response variable.

$\mathbf{X}_i$ is a $p \times 1$ observable random vector; there are p explanatory variables. $\mathbf{X}_i$ has expected value $\boldsymbol{\mu}_x$ and variance-covariance matrix $\boldsymbol{\Phi}$, a $p \times p$ symmetric and positive definite matrix of unknown constants.

$\boldsymbol{\beta}_1$ is a $q \times p$ matrix of unknown constants. These are the regression coefficients, with one row for each response variable and one column for each explanatory variable.

$\boldsymbol{\epsilon}_i$ is the error term of the latent regression. It is an $q \times 1$ multivariate normal random vector with expected value zero and variance-covariance matrix $\boldsymbol{\Psi}$, a $q \times q$ symmetric and positive definite matrix of unknown constants. $\boldsymbol{\epsilon}_i$ is independent of $\mathbf{X}_i$.

The parameter vector for this model could be written $\boldsymbol{\theta} = (\boldsymbol{\beta}_0, \boldsymbol{\mu}_x, \boldsymbol{\Phi}, \boldsymbol{\beta}_1, \boldsymbol{\Psi}, F_{\mathbf{x}}, F_{\boldsymbol{\epsilon}})$, where it is understood that the symbols for the matrices really refer to their unique elements.

Figure **??** depicts a model with three explanatory variables and two response variables. The explanatory and response variables are all observed, so they are enclosed in boxes. Double-headed curved arrows between the explanatory variable represent possible non-zero covariances. The straight arrows from the explanatory to response variables represent direct influence, or at least that we are interested in predicting y from x rather than the other way around. There is a regression coefficient β on each arrow. The error terms ϵ_1 and ϵ_2 represent all other influences on Y_1 and Y_2. Since there could be common influences (omitted variables that affect both Y_1 and Y_2), the error terms are assumed to be correlated. This is the reason for the curved double-headed arrow joining ϵ_1 and ϵ_2.
There is one regression equation for each response variable. In scalar form, the model equations are

$$\begin{aligned} Y_{i,1} &= \beta_{1,0} + \beta_{1,1}X_{i,1} + \beta_{1,2}X_{i,2} + \beta_{1,3}X_{i,3} + \epsilon_{i,1} \\ Y_{i,2} &= \beta_{2,0} + \beta_{2,1}X_{i,1} + \beta_{2,2}X_{i,2} + \beta_{2,3}X_{i,3} + \epsilon_{i,2}. \end{aligned}$$

In matrix form,

$$\mathbf{Y}_i = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 \mathbf{X}_i + \boldsymbol{\epsilon}_i$$

$$\begin{pmatrix} Y_{i,1} \\ Y_{i,2} \end{pmatrix} = \begin{pmatrix} \beta_{1,0} \\ \beta_{2,0} \end{pmatrix} + \begin{pmatrix} \beta_{1,1} & \beta_{1,2} & \beta_{1,3} \\ \beta_{2,1} & \beta_{2,2} & \beta_{2,3} \end{pmatrix} \begin{pmatrix} X_{i,1} \\ X_{i,2} \\ X_{i,3} \end{pmatrix} + \begin{pmatrix} \epsilon_{i,1} \\ \epsilon_{i,2} \end{pmatrix}$$

Note that in traditional linear regression, the regression coefficients $\boldsymbol{\beta}$ form a column vector, while the explanatory variables (a row of the $\mathbf{X}$ matrix) are a row vector. Here it is the other way around, in order to allow the random vector $\mathbf{X}_i$ to be a column vector.

Returning to the general case of Example 4.3, the observable data are the random vectors $\mathbf{D}_i = \begin{pmatrix} \mathbf{X}_i \\ \hline \mathbf{Y}_i \end{pmatrix}$, for $i = 1, \ldots, n$. The notation indicates that $\mathbf{D}_i$ is a partitioned random vector, with $\mathbf{X}_i$ stacked directly on top of $\mathbf{Y}_i$. Using the notation $E(\mathbf{D}_i) = \boldsymbol{\mu}$ and $cov(\mathbf{D}_i) = \boldsymbol{\Sigma}$, one may write $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ as partitioned matrices (matrices of matrices).

$$\boldsymbol{\mu} = \begin{pmatrix} E(\mathbf{X}_i) \\ \hline E(\mathbf{Y}_i) \end{pmatrix} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \hline \boldsymbol{\mu}_2 \end{pmatrix}$$

and

$$\boldsymbol{\Sigma} = V \begin{pmatrix} \mathbf{X}_i \\ \hline \mathbf{Y}_i \end{pmatrix} = \left(\begin{array}{c|c} cov(\mathbf{X}_i) & cov(\mathbf{X}_i, \mathbf{Y}_i) \\ \hline cov(\mathbf{X}_i, \mathbf{Y}_i)^\top & cov(\mathbf{Y}_i) \end{array}\right) = \left(\begin{array}{c|c} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \hline \boldsymbol{\Sigma}_{12}^\top & \boldsymbol{\Sigma}_{22} \end{array}\right)$$

As in the univariate case, the maximum likelihood estimators may be obtained by solving the moment structure equations for the unknown parameters. The moment structure equations are obtained by calculating expected values and covariances in terms of the model parameters. All the calculations are immediate except possibly

$$\begin{aligned} \boldsymbol{\Sigma}_{12} &= cov(\mathbf{X}_i, \mathbf{Y}_i) \\ &= cov(\overset{c}{\mathbf{X}}_i, \overset{c}{\mathbf{Y}}_i) \\ &= E\left(\overset{c}{\mathbf{X}}_i \, (\boldsymbol{\beta}_1 \, \overset{c}{\mathbf{X}}_i + \boldsymbol{\epsilon}_i)^\top\right) \\ &= \boldsymbol{\Phi}\boldsymbol{\beta}_1^\top \end{aligned}$$

Thus, the moment structure equations are

$$\begin{aligned} \boldsymbol{\mu}_1 &= \boldsymbol{\mu}_x \\ \boldsymbol{\mu}_2 &= \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1\boldsymbol{\mu}_x \\ \boldsymbol{\Sigma}_{11} &= \boldsymbol{\Phi} \\ \boldsymbol{\Sigma}_{12} &= \boldsymbol{\Phi}\boldsymbol{\beta}_1^\top \\ \boldsymbol{\Sigma}_{22} &= \boldsymbol{\beta}_1\boldsymbol{\Phi}\boldsymbol{\beta}_1^\top + \boldsymbol{\Psi}. \end{aligned} \tag{11}$$

Solving for the parameter matrices is routine.

$$\begin{aligned} \boldsymbol{\beta}_0 &= \boldsymbol{\mu}_2 - \boldsymbol{\Sigma}_{12}^\top\boldsymbol{\Sigma}_{11}^{-1}\,\boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_x &= \boldsymbol{\mu}_1 \\ \boldsymbol{\Phi} &= \boldsymbol{\Sigma}_{11} \\ \boldsymbol{\beta}_1 &= \boldsymbol{\Sigma}_{12}^\top\boldsymbol{\Sigma}_{11}^{-1} \\ \boldsymbol{\Psi} &= \boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{12}^\top\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12} \end{aligned} \tag{12}$$

As in the univariate case, the Method of Moments estimates are obtained by putting hats on all the parameters in Expression (12). If the distributions of $\mathbf{X}_i$ and $\boldsymbol{\epsilon}_i$ are multivariate normal, the Invariance Principle reveals that Method of Moments estimates are also the maximum likelihood estimates.

5 Omitted Variables

Some very serious problems arise when standard regression methods are applied to non-experimental data. Note that regression methods are applied to non-experimental data *all the time*, and we teach students how to do it in almost every Statistics class where regression is mentioned. But without an understanding of the technical issues involved, the usual applications can be misleading.

The problems do not arise because the explanatory variables are random. As we saw in Section 1, that's fine. The problems arise because the random explanatory variables have non-zero correlations with other explanatory variables that are missing from the regression equation and are related to the response variable. In this section, we will see how omitting important

explanatory variables from a regression equation can cause the error term to be correlated with the explanatory variables that remain, and how that can produce incorrect results.

To appreciate the issue, it is necessary to understand what the error term in a regression equation really represents. When we write something like

$$Y_i = \beta_0 + \beta_1 X_{i,1} + \epsilon_i, \tag{13}$$

we are saying that $X_{i,1}$ contributes to Y_i, but there are also other, unspecified influences. Those other influences are all rolled together into ϵ_i.

The words "contributes" and "influences" are used deliberately. They should be setting off alarm bells, because they imply a causal connection between X_i and Y_i. Regression models with random explanatory variables are applied mostly to observational data, in which explanatory variables are merely recorded rather than being manipulated by the investigator. The correlation-causation issue applies. That is, if X and Y are related, there is in general no way to tell whether X is influencing Y, or Y is influencing X, or if other variables are influencing both X and Y.

It could be argued that a *conditional* regression model (the usual model in which the explanatory variable values are fixed constants) is just a convenient way to represent dependence between X and Y by specifying a generic, more or less reasonable conditional distribution for Y given $X = x$. In this case, the correlation-causation issue can be set aside, and taken up when it is time to interpret the results. But if the explanatory variables are explicitly random, it is harder to avoid the obvious. In the simple regression model (13), the random variable Y_i is a function of the random variables X_i and ϵ_i. It is being directly produced by them. If this is taken seriously as a *scientific* model as well as a statistical model[7], it is inescapably causal; it is a model of what affects what. That's why the straight arrows in path diagrams are directional. The issue of whether X is influencing Y, or Y is influencing X or both is a modelling issue that will mostly be decided based on subject-matter theory.

It is natural to ask whether the data can be used to decide which way the arrows should be pointing. The answer is that sometimes it can, and sometimes it can't. We will return to this issue later in the book. In the meantime, regression models with random explanatory variables, like the general structural equation models that are their extensions, will be recognized as causal models.

Again, Equation (13) says that X_i is influencing Y_i. All other influences are represented ϵ_i. It is common practice to assume that $X_{i,1}$ and ϵ_i are independent, or at least uncorrelated. But that does not mean the assumption can be justified in practice. Prepare yourself for a dose of reality.

Example 5.1

Suppose that the variables X_2 and X_3 have an impact on Y and are correlated with X_1, but they are not part of the data set. The values of the response variable are generated as follows:

$$Y_i = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \beta_3 X_{i,3} + \epsilon_i, \tag{14}$$

independently for $i = 1, \ldots, n$, where $\epsilon_i \sim N(0, \sigma^2)$. The explanatory variables are random, with expected value and variance-covariance matrix

$$E\begin{pmatrix} X_{i,1} \\ X_{i,2} \\ X_{i,3} \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{pmatrix} \quad \text{and} \quad cov\begin{pmatrix} X_{i,1} \\ X_{i,2} \\ X_{i,3} \end{pmatrix} = \begin{pmatrix} \phi_{11} & \phi_{12} & \phi_{13} \\ & \phi_{22} & \phi_{23} \\ & & \phi_{33} \end{pmatrix},$$

where ϵ_i is independent of $X_{i,1}$, $X_{i,2}$ and $X_{i,3}$. Values of the variables $X_{i,2}$ and $X_{i,3}$ are latent, and are not included in the data set.

Figure 1 shows a path diagram of this situation. Because the explanatory variables $X_{i,2}$ and $X_{i,3}$ are not observable, they are *latent* variables, and so they are enclsed by ovals in the path

[7] In structural equation modelling, the models are both statistical models and primitive scientific models of the data. Once the general linear structural model is introduced, you will see that regression is a special case.

diagram. Their covariances with $X_{i,1}$ and each other are represented by two-headed curved arrows.

Figure 1: Omitted explanatory variables

includegraphics[width=4in]Pictures/OmittedPath1

Since X_2 and X_3 are not observed, they are absorbed by the intercept and error term.

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \beta_3 X_{i,3} + \epsilon_i \\ &= (\beta_0 + \beta_2\mu_2 + \beta_3\mu_3) + \beta_1 X_{i,1} + (\beta_2 X_{i,2} + \beta_3 X_{i,3} - \beta_2\mu_2 - \beta_3\mu_3 + \epsilon_i) \\ &= \beta_0' + \beta_1 X_{i,1} + \epsilon_i'. \end{aligned}$$

The primes just denote a new β_0 and a new ϵ; the addition and subtraction of $\beta_2\mu_2 + \beta_3\mu_3$ serve to make $E(\epsilon_i') = 0$. And of course there could be any number of omitted variables. They would all get swallowed by the intercept and error term, the garbage bins of regression analysis.

Notice that although the original error term ϵ_i is independent of $X_{i,1}$, the new error term ϵ_i' is not.

$$\begin{aligned} Cov(X_{i,1}, \epsilon_i') &= Cov(X_{i,1}, \beta_2 X_{i,2} + \beta_3 X_{i,3} - \beta_2\mu_2 - \beta_3\mu_3 + \epsilon_i) \\ &= E\left(\overset{c}{X}_{i,1}\,(\beta_2\,\overset{c}{X}_{i,2} + \beta_3\,\overset{c}{X}_{i,3} + \epsilon_i)\right) \\ &= \beta_2\phi_{12} + \beta_3\phi_{13} \end{aligned} \tag{15}$$

So, when explanatory variables are omitted from the regression equation and those explanatory variables have non-zero covariance with variables that *are* in the equation, the result is non-zero covariance between the error term and the explanatory variables in the equation[8].

Response variables are almost always affected by more than one explanatory variable, and in observational data, explanatory variables usually have non-zero covariances with one another. So, the most realistic model for a regression with just one explanatory variable should include a covariance between the error term and the explanatory variable. The covariance comes from the regression coefficients and covariances of some unknown number of omitted variables; it will be represented by a single quantity because there is no hope of estimating all those parameters individually. We don't even know how many there are.

We have arrived at the following model, which will be called the *true model* in the discussion that follows. It may not be the ultimate truth of course, but for observational data it is almost always closer to the truth than the usual model. Independently for $i = 1, \ldots, n$,

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \tag{16}$$

where $E(X_i) = \mu_x$, $Var(X_i) = \sigma_x^2$, $E(\epsilon_i) = 0$, $Var(\epsilon_i) = \sigma_\epsilon^2$, and $Cov(X_i, \epsilon_i) = c$. A path diagram of the true model is given in Figure 2. The covariance c is indicated on the curved arrow connecting the explanatory variable and the error term. Consider a data set consisting

Figure 2: Omitted explanatory variables have been swallowed by ϵ

includegraphics[width=4in]Pictures/OmittedPath2

of pairs $(X_1, Y_1), \ldots, (X_n, Y_n)$ *coming from the true model*, and the interest is in the regression coefficent β_1. Who will try to estimate the parameters of the true model? Almost no one.

[8]The effects of the omitted variables could offset each other. In this example, it is possible that $\beta_2\phi_{12} + \beta_3\phi_{13} = 0$, but that is really too much to hope.

Practically everyone will use ordinary least squares, as described in countless Statistics textbooks and implemented in countless computer programs and statistical calculators.

The model underlying ordinary least squares is $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, where $x_1, \ldots, x_n$ are fixed constants, and conditionally on $x_1, \ldots, x_n$, the error terms $\epsilon_1, \ldots, \epsilon_n$ are independent normal random variables with mean zero and variance σ^2. It may not be immediately obvious, but this model implies independence of the explanatory variable and the error term. It is a conditional model, and the distribution of the error terms is *the same* for every fixed set of values $x_1, \ldots, x_n$. Using a loose but understandable notation for densities and conditional densities,

$$\begin{aligned} & f(\epsilon_i|x_i) = f(\epsilon_i) \\ \Leftrightarrow \quad & \frac{f(\epsilon_i, x_i)}{f(x_i)} = f(\epsilon_i) \\ \Leftrightarrow \quad & f(\epsilon_i, x_i) = f(\epsilon_i) f(x_i), \end{aligned}$$

which is the definition of independence. So, the usual regression model makes a hidden assumption. It assumes that *any explanatory variable that is omitted from the equation has zero covariance with the variables that are in the equation.*

Surprisingly, this does not depend on the assumption of any particular distribution for the error terms. All you need is the stipulation $E(\epsilon_i) = 0$ in a fixed-x regression model. It's worth doing this in generality, so consider the multivariate multiple regression model of Example 4.3 on page 11:

$$\mathbf{Y}_i = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 \mathbf{X}_i + \boldsymbol{\epsilon}_i.$$

If the $\mathbf{X}_i$ values are considered fixed constants, the statement $E(\boldsymbol{\epsilon}_i) = \mathbf{0}$ actually means $E(\boldsymbol{\epsilon}_i|\mathbf{X}_i = \mathbf{x}_i) = \mathbf{0}$ for all $p \times 1$ constant vectors $\mathbf{x}_i$ in the support of $\mathbf{X}_i$. Then,

$$E(\boldsymbol{\epsilon}_i) = E\{E(\boldsymbol{\epsilon}_i|\mathbf{X}_i)\} = E\{\mathbf{0}\} = \mathbf{0},$$

and

$$\begin{aligned} cov(\mathbf{X}_i, \boldsymbol{\epsilon}_i) &= E(\mathbf{X}_i \boldsymbol{\epsilon}_i^\top) - E(\mathbf{X}_i) E(\boldsymbol{\epsilon}_i)^\top \\ &= E(\mathbf{X}_i \boldsymbol{\epsilon}_i^\top) - \mathbf{0} \\ &= E\{E(\mathbf{X}_i \boldsymbol{\epsilon}_i^\top | \mathbf{X}_i)\}. \end{aligned}$$

The inner expected value is a multiple integral or sum with respect to the conditional distribution of $\boldsymbol{\epsilon}_i$ given $\mathbf{X}_i$, so $\mathbf{X}_i$ may be moved through the inner expected value sign. To see this, it may help to write the double expectation in terms of integrals of a general kind[9]. Continuing the calculation,

$$\begin{aligned} E\{E(\mathbf{X}_i \boldsymbol{\epsilon}_i^\top | \mathbf{X}_i)\} &= \int \left(\int \mathbf{x} \boldsymbol{\epsilon}^\top dP_{\boldsymbol{\epsilon}|\mathbf{X}}(\boldsymbol{\epsilon}) \right) dP_{\mathbf{X}}(\mathbf{x}) \\ &= \int \mathbf{x} \left(\int \boldsymbol{\epsilon}^\top dP_{\boldsymbol{\epsilon}|\mathbf{X}}(\boldsymbol{\epsilon}) \right) dP_{\mathbf{X}}(\mathbf{x}) \\ &= E\{\mathbf{X}_i E(\boldsymbol{\epsilon}_i^\top | \mathbf{X}_i)\} \\ &= E\{\mathbf{X}_i \mathbf{0}^\top\} \\ &= E\{\mathbf{0}\} \\ &= \mathbf{0} \end{aligned}$$

Unconditional (random $\mathbf{X}$) regression models typically assume zero covariance between error terms and explanatory variables. It is now clear that conditional (fixed $\mathbf{x}$) regression models smuggle this same assumption in by making the seemingly reasonable and harmless assertion that $E(\boldsymbol{\epsilon}_i) = \mathbf{0}$.

Zero covariance between error terms and explanatory variables means that *any potential explanatory variable not in the model must have zero covariance with the explanatory variables that*

[9] These are Lebesgue integrals with respect to probability measures and conditional probability measures. They include multiple sums and Reimann integrals as special cases.

are in the model. Of course this is almost never realistic without random assignment to experimental conditions, so that almost every application of regression methods to non-experimental data makes an assumption that cannot be justified. Now we will see the consequences.

For a simple regression, both ordinary least squares and an unconditional regression model like the true model on Page 14 with $c = 0$ lead to the same standard formula:

$$\begin{aligned}\widehat{\beta}_1 &= \frac{\sum_{i=1}^n (X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^n (X_i - \overline{X})^2} \\ &= \frac{\sum_{i=1}^n (X_i - \overline{X})(Y_i - \overline{Y})/n}{\sum_{i=1}^n (X_i - \overline{X})^2/n} \\ &= \frac{\widehat{\sigma}_{x,y}}{\widehat{\sigma}_x^2},\end{aligned}$$

where $\widehat{\sigma}_{x,y}$ is the sample covariance between X and Y, and $\widehat{\sigma}_x^2$ is the sample covariance of X. These are maximum likelihood estimates of $Cov(X,Y)$ and $Var(X)$ respectively under the assumption of normality, and if the divisors were $n-1$ instead of n, they would be unbiased.

By the consistency of the sample variance and covariance (see Section **??** in Appendix **??**), $\widehat{\sigma}_{x,y}$ converges to $Cov(X,Y)$ and $\widehat{\sigma}_x^2$ converges to $Var(X)$ as $n \to \infty$. Under the true model,

$$Cov(X,Y) = Cov(X_i, \beta_0 + \beta_1 X_i + \epsilon_i) = \beta_1 \sigma_x^2 + c.$$

So by a continuity argument (Slutsky lemmas **??** and **??**) in Section **??**),

$$\widehat{\beta}_1 = \frac{\widehat{\sigma}_{x,y}}{\widehat{\sigma}_x^2} \stackrel{a.s.}{\to} \beta_1 + \frac{c}{\sigma_x^2}. \tag{17}$$

Since the estimator is converging to quantity that is off by a fixed amount, it is reasonable to call it *asymptotically biased*. Thus, while the usual teaching is that sample regression coefficients are unbiased estimators, we see here that $\widehat{\beta}_1$ is biased as $n \to \infty$. Regardless of the true value β_1, the estimate $\widehat{\beta}_1$ could be absolutely anything, depending on the value of c, the covariance between X_i and ϵ_i. The only time $\widehat{\beta}_1$ behaves properly is when $c = 0$.

What's going on here is that the calculation of $\widehat{\beta}_1$ is based on a model that is *mis-specified*. That is, it's not the right model. The right model is what we've been calling the *true model*. And to repeat, the true model is the most reasonable model for simple regression, at least for most non-experimental data.

The lesson is this. *When a regression model fails to include all the explanatory variables that contribute to the response variable, and those omitted explanatory variables have non-zero covariance with variables that are in the model, the regression coefficients are inconsistent*. In other words, with more and more data they do not approach the right answer. Instead, they get closer and closer to a specific wrong answer.

If you think about it, this fits with what happens frequently in practical regrssion analysis. When you add a new explanatory variable to a regression equation, the coefficients of the variables that are already in the equation do not remain the same. Almost anything can happen. Positive coefficients can turn negative, negative ones can turn positive, statistical significance can appear where it was previously absent or disappear where it was previously present. Now you know why.

Notice that if the values of one or more explanatory variables are randomly assigned, the random assignment guarantees that these variables are independent of any and all variables that are omitted from the regression equation. Thus, the variables in the equation have zero covariance with those that are omitted, and all the trouble disappears. So, *well-controlled experimental studies are not subject to the kind of problems described here.*

Actually, the calculations in this section support a familiar point, the *correlation-causation* issue, which is often stated more or less as follows. If A and B are related to one another, one cannot necessarily infer that A affects B. It could be that B affects A, or that some third variable C is affecting both A and B. To this we can now add the possibility that the third variable C

affects B and is merely correlated with A.

Variables like C are often called *confounding variables*, or more rarely, *lurking variables*. The usual advice is that the only way to completely rule out their action is to randomly assign subjects in the study to the various values of A, and then assess the relationship of A to B. Again, now you know why.

It should be pointed out that while the correlation-causation issue presents grave obstacles to interpreting the results of observational studies, there is no problem with pure prediction. If you have a data set with x and y values and your interest is predicting y from the x values for a new set of data, a regression equation will be useful, provided that there is a reasonably strong relationship between x and y. From the standpoint of prediction, it does not really matter whether y is related to x directly, or indirectly through unmeasured variables that are related to x. You have x and not the unmeasured variables, so use it. An example would be an insurance company that seeks to predict the amount of money that you will claim next year (so they can increase your premiums accordingly now). If it turns out that this is predictable from the type of music you download, they will cheerfully use the information, and not care why it works.

Also, the convergence of $\widehat{\beta}_1$ to the wrong answer in (17) may be misleading, but it does not necessarily yield the wrong conclusion. In much of the social and biological sciences, the theories are not detailed and sophisticated enough to make predictions about the actual values of regression coefficients, just whether they should be positive, negative or zero. So, if the variable being tested and the omitted variables are pulling in the same direction (that is, if β_1 and c in Model (16) on Page 14 are either both positive or both negative), the study will come to the "right" conclusion. The trouble is that you can't tell, because you don't even know what the omitted variables are. All you can do is hope, and that's not a recipe for good science.

Trying to fit the true model We have seen that serious trouble arises from adopting a misspecified model with $c = Cov(X_i, \epsilon_i) = 0$, when in fact because of omitted variables, $c \neq 0$. It is natural, therefore, to attempt estimation and inference for the true model $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$ (see Page 14) in the case where $c = Cov(X_i, \epsilon_i)$ need not equal zero. For simplicity, assume that X_i and ϵ_i have a bivariate normal distribution, so that the observable data pairs (X_i, Y_i) for $i = 1, \ldots, n$ are a random sample from a bivariate normal distribution with mean vector $\boldsymbol{\mu}$ and variance-covariance matrix $\boldsymbol{\Sigma}$.

It is straightforward to calculate $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ from the equation and assumptions of the true model (16). The result is

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} = E\begin{pmatrix} X_i \\ Y_i \end{pmatrix} = \begin{pmatrix} \mu_x \\ \beta_0 + \beta_1\mu_x \end{pmatrix} \tag{18}$$

and

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{pmatrix} = cov\begin{pmatrix} X_i \\ Y_i \end{pmatrix} = \begin{pmatrix} \sigma_x^2 & \beta_1\sigma_x^2 + c \\ \beta_1\sigma_x^2 + c & \beta_1^2\sigma_x^2 + 2\beta_1 c + \sigma_\epsilon^2 \end{pmatrix}. \tag{19}$$

This shows the way in which the parameter vector $\boldsymbol{\theta} = (\mu_x, \sigma_x^2, \beta_0, \beta_1, \sigma_\epsilon^2, c)$ determines $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, and hence the probability distribution of the data.

Our primary interest is in β_1. Because the data pairs (X_i, Y_i) come from a bivariate normal distribution, all you can ever learn from the data are the approximate values of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. With larger and larger samples, all you get is better and better approximations of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. That's all there is to know. But even if you knew $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ exactly, could you know β_1? Formulas (18)

and (19) yield a system of five equations in six unknown parameters.

$$\begin{aligned} \mu_1 &= \mu_x \\ \mu_2 &= \beta_0 + \beta_1 \mu_x \\ \sigma_{11} &= \sigma_x^2 \\ \sigma_{12} &= \beta_1 \sigma_x^2 + c \\ \sigma_{22} &= \beta_1^2 \sigma_x^2 + 2\beta_1 c + \sigma_\epsilon^2 \end{aligned} \tag{20}$$

The problem of recovering the parameter values from $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ is exactly the problem of solving these five equations in six unknowns. $\mu_x = \mu_1$ and $\sigma_x^2 = \sigma_{11}$ are easy. The remaining 3 equations in 4 unknowns have infinitely many solutions. That is, infinitely many sets of parameter values yield *exactly the same distribution of the sample data*. Distinguishing among them based on sample data is impossible in principle.

To see this in detail, substitute μ_1 for μ_x and σ_{11} for σ_x^2 in (20), obtaining

$$\begin{aligned} \mu_2 &= \beta_0 + \beta_1 \mu_1 \\ \sigma_{12} &= \beta_1 \sigma_{11} + c \\ \sigma_{22} &= \beta_1^2 \sigma_{11} + 2\beta_1 c + \sigma_\epsilon^2 \end{aligned} \tag{21}$$

Letting the moments μ_j and σ_{ij} remain fixed, we will now write the other parameters as functions of c, the covariance between X_i and ϵ_i. Then, moving c will move the other parameters (except for $\mu_x = \mu_1$ and $\sigma_x^2 = \sigma_{11}$), tracing out a one-dimensional subset of the 6-dimensional parameter space where

- All the equations in (20) are satisfied,
- The values of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ remain constant, and
- The distribution of $(X_i, Y_i)^\top$ is $N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

First solve for β_1 in the second equation, obtaining $\beta_1 = \frac{\sigma_{12} - c}{\sigma_{11}}$. Substituting this expression for β_1 and simplifying, we are able to write all the other model parameters in terms of c, as follows.

$$\begin{aligned} \mu_x &= \mu_1 \\ \sigma_x^2 &= \sigma_{11} \\ \beta_0 &= \mu_2 - \mu_1 \left(\frac{\sigma_{12} - c}{\sigma_{11}} \right) \\ \beta_1 &= \frac{\sigma_{12} - c}{\sigma_{11}} \\ \sigma_\epsilon^2 &= \sigma_{22} + \frac{c^2 - \sigma_{12}^2}{\sigma_{11}} \end{aligned} \tag{22}$$

The parameters μ_x and σ_x^2 are constant functions of c, while β_0 and β_1 are linear functions, and σ_ϵ^2 is a quadratic function. The equations (22) define a one-dimensional surface in the six-dimensional parameter space, a kind of curved thread in $\mathbb{R}^6$. Moving c from $-\infty$ to ∞ traces out the points on the thread. Importantly, as c ranges from $-\infty$ to $+\infty$ the regression coefficient β_1 ranges from $+\infty$ to $-\infty$. This means that β_1 might be positive, it might be negative, or it might be zero. But you really can't tell, because all real values of β_1 on the surface yield the same population mean and population variance-covariance matrix, and hence the same distribution of the sample data. There is no way to distinguish between the possible values of β_1 based on sample data.

One technical detail needs to be resolved. Can c really range from $-\infty$ to ∞? If not, the possible values of β_1 would be restricted as well. Two conditions need to be checked. First, the covariance matrix of $(X_i, \epsilon_i)^\top$ has a non-negative determinant by the Cauchy-Schwarz inequality. For the bivariate normal density to exist (not a bad assumption), the determinant must be non-zero, and hence strictly positive. Second, σ_ϵ^2 must be greater than zero. For points on the thread,

the first condition is

$$\begin{vmatrix} \sigma_x^2 & c \\ c & \sigma_\epsilon^2 \end{vmatrix} = \sigma_x^2\sigma_\epsilon^2 - c^2$$
$$= \sigma_{11}\left(\sigma_{22} + \frac{c^2 - \sigma_{12}^2}{\sigma_{11}}\right) - c^2$$
$$= \sigma_{11}\sigma_{22} + c^2 - \sigma_{12}^2 - c^2$$
$$= \sigma_{11}\sigma_{22} - \sigma_{12}^2$$
$$= |\boldsymbol{\Sigma}| > 0.$$

This imposes no restriction on c at all. We also need to check whether $\sigma_\epsilon^2 > 0$ places any restriction on c for points on the thread.

$$\sigma_\epsilon^2 > 0$$
$$\Leftrightarrow \quad \sigma_{22} + \frac{c^2 - \sigma_{12}^2}{\sigma_{11}} > 0$$
$$\Leftrightarrow \quad \sigma_{11}\sigma_{22} + c^2 - \sigma_{12}^2 > 0$$
$$\Leftrightarrow \quad |\boldsymbol{\Sigma}| + c^2 > 0,$$

which is true since $|\boldsymbol{\Sigma}| > 0$. Again, the inequality places no restriction on c.

Let me beat this point into the ground a bit, because it is important. Since the data are bivariate normal, their probability distribution corresponds uniquely to the pair $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. All you can *ever* learn from *any* set of sample data is the probability distribution from which they come. So all you can ever get from bivariate normal data, no matter what the sample size, is a closer and closer approximation of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. If you cannot find out whether β_1 is positive, negative or zero from $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, you will *never* be able to make reasonable estimates or inferences about it from any set of sample data.

What would happen if you tried to estimate the parameters by maximum likelihood? For every $\boldsymbol{\mu} \in \mathbb{R}^2$ and every 2×2 symmetric positive definite $\boldsymbol{\Sigma}$, there is a surface (thread) in $\mathbb{R}^6$ defined by (22). This includes $(\widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Sigma}})$. On that particular thread, the likelihood is highest. Picture a surface with a curvy ridge at the top. The surface has infinitely many maxima, all at the same height, forming a connected set. If you take partial derivatives of the log likelihood and set them all equal to zero, there will be infinitely many solutions. If you do numerical maximum likelihood, good software will find a point on the ridge, stop, detect that the surface is not fully concave down there, and complain. Less sophisticated software will just find a point on the ridge, and stop. The stopping place, that is, the maximum likelihood estimate, depends entirely on where the numerical search starts.

To summarize, if explanatory variables are omitted from a regression equation and those variables have non-zero covariance c with explanatory variables that are *not* omitted, the result is non-zero covariance between explanatory variables and the error term. And, if there is a non-zero covariance between the error term an an explanatory variable in a regression equation, the false assumption that $c = 0$ can easily lead to false results. But allowing c to be non-zero means that infinitely many parameter estimates will be equally plausible, given any set of sample data. In particular, no set of data will be able to provide a basis for deciding whether regression coefficients are positive, negative or zero. The problem is fatal if all you have is X_i and Y_i.

The trouble here is lack of parameter identifiability. If a parameter is a function of the distribution of the observable data, it is said to be *identifiable*. The idea is that the parameter is potentially knowable if you knew the distribution of the observable data. If the parameter is not knowable based on the data, they naturally there will be trouble with estimation and inference. Parameter identifiability is a central theme of this book, and will be taken up again in Section 10 on Page 33.

6 Instrumental Variables as a Solution to Omitted Variables

The method of instrumental variables was introduced by the economist Phillip Wright in the appendix a 1928 book *The Tariff on Animal and Vegetable Oils* [**?**]. See also the historical account by Stock and Trebbi [**?**]. An instrumental variable is a variable that is correlated with an explanatory variable, but is not correlated with any error terms and has no direct connection to the response variable. In Econometrics, the instrumental variable usually *influences* the explanatory variable. An instrumental variable is usually not the main focus of attention; it's just a tool.

Example 6.1

Suppose we want to know the contribution of income to credit card debt. Because of omitted variables, the model

$$Y_i = \alpha + \beta X_i + \epsilon_i,$$

is guaranteed to fail. Many things influence both income and credit card debt, such as personal style of money management, education, number of children, expenses caused by illness .... The list goes on. As a result, X_i and ϵ_i have non-zero covariance. The least squares estimate of β is inconsistent, and so is every other possible estimate[10]. We can't possibly measure all the variables that affect both income and debt; we don't even know what they all are. Instead, let's add an instrumental variable.

Definition 6.1 *An instrumental variable for an explanatory variable is another random variable that has non-zero covariance with the explanatory variable, and* no direct connection with any other variable in the model.

Focus the study on real estate agents in many cities, and include median price of resale home for each agent along with income and credit card debt. Median price of resale home qualifies an an instrumental variable according to the definition. Since real estate agents typically receive a percentage of the selling price, it is definitely related to income. Also, housing prices are determined by external economic forces that have little to do with all the personal, individual-level variables that affect income and debt. So, we have the following:

- W_i is median price of resale home in agent i's district.
- X_i is annual income of real estate agent i.
- Y_i is agent i's credit card debt.

The model equations are

$$\begin{aligned} X_i &= \alpha_1 + \beta_1 W_i + \epsilon_{i1} \\ Y_i &= \alpha_2 + \beta_2 X_i + \epsilon_{i2}, \end{aligned}$$

and Figure 3 shows the path diagram. The main interest is in β_2, the link between income and credit card debt. The covariance between ϵ_1 and ϵ_2 represents all the omitted variables that affect income and credit card debt.

[10] This is strictly true if the data are normal. For non-normal data something might be possible, but one would have to know the specific non-normal distribution.

Figure 3: W is median price of resale home, X is income, Y is credit card debt

includegraphics[width=5in]Pictures/InstruVar

Denoting the expected value of the data vector $\mathbf{D}_i = (W_i, X_i, Y_i)^\top$ by $\boldsymbol{\mu} = [\mu_j]$ and its covariance matrix by $\boldsymbol{\Sigma} = [\sigma_{ij}]$, we have

$$\boldsymbol{\Sigma} = \begin{array}{c|ccc} & W & X & Y \\ \hline W & \sigma_w^2 & \beta_1\sigma_w^2 & \beta_1\beta_2\sigma_w^2 \\ X & \cdot & \beta_1^2\sigma_w^2 + \sigma_1^2 & \beta_2(\beta_1^2\sigma_w^2 + \sigma_1^2) + c \\ Y & \cdot & \cdot & \beta_1^2\beta_2^2\sigma_w^2 + \beta_2^2\sigma_1^2 + 2\beta_2 c + \sigma_2^2 \end{array} \tag{23}$$

The lower triangle of the covariance matrix is omitted to make it less cluttered. The notation in (23) is self-explanatory except possibly for $Var(\epsilon_{i1}) = \sigma_1^2$ and $Var(\epsilon_{i2}) = \sigma_2^2$. It is immediately apparent that the critical parameter β_2 can be recovered from $\boldsymbol{\Sigma}$ by $\beta_2 = \frac{\sigma_{13}}{\sigma_{12}}$, provided $\beta_1 \neq 0$. A nice Method of Moments estimator in terms of the sample covariances is $\widehat{\beta}_2 = \frac{\widehat{\sigma}_{13}}{\widehat{\sigma}_{12}}$.

The requirement that $\beta_1 \neq 0$ is no problem, because W is a good instrumental variable. Median resale price is certainly related to the income of real estate agents, and furthermore the relationship is guaranteed to be positive. This is a feature of good a instrumental variable. Its relationship to the explanatory variable should be clear, and so obvious that it is hardly worth investigating. The usefulness of the instrumental variable is in the light it casts on relationships that are not so obvious.

In this example, the instrumental variable worked beautifully. All the model parameters that appear in $\boldsymbol{\Sigma}$ can be recovered by simple substitution, $\mu_z = \mu_1$, and then α_1 and α_2 can be recovered from $\mu_2 = E(X_i)$ and $\mu_3 = E(Y_i)$ respectively. The function from $(\alpha_1, \alpha_2, \beta_1, \beta_2, \mu_w, \sigma_w^2, \sigma_1^2, \sigma_2^2, c)$ to $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is one-to one. Method of Moments estimates are readily available, and they are consistent by the continuity of the functions involved. Under the additional assumption of multivariate normality, the Method of Moments estimates are also maximum likelihood by the invariance principle.

To test the central null hypothesis $H_0 : \beta_2 = 0$, fancy software is not required. The covariance σ_{13} equals zero if and only if $\beta_2 = 0$, and they have the same sign because we are convinced that $\beta_1 > 0$. So it is necessary only to test the correlation between housing price and real estate agents' credit card debt. Under the normal assumption, the usual test is exact and a large sample is not required. If the normal assumption is worrisome, the non-parametric test associated with the Spearman rank correlation coefficient is a permutation test carried out on ranks, and an exact small-sample p-value is available even though some software produces a large-sample approximation by default.

The instrumental variable method saved the day in this example, but it does not solve the problem of omitted variables in every case, or even in most cases. This is because good instrumental variables are not easy to find. They will not just happen to be in the data set, except by a miracle. They really have to come from another universe, and still have a strong, clear connection to the explanatory variable. Data collection has to be *planned*, with a model that admits the existence of omitted variables explicitly in mind.

Measurement Error All models are inexact representations of reality, but I must admit that the model in Figure 3 is seriously wrong. Our interest is in how *true* income affects *true* credit card debt. But these variables are not observed. What we have in the data file are *reported* income

and *reported* credit card debt. For various reasons that the reader can easily supply, the truth and what people report about financial details are not the same thing. When we record median price of a resale home, that's unlikely to be perfectly accurate either. As we will see later in this chapter, measurement error in the explanatory variables presents serious problems for regression analysis in general. We will also see that instrumental variables can help with measurement error as well as with omitted variables, but first it is helpful to introduce the topic of measurement error in an organized way.

7 The Idea of Measurement Error

In a survey, suppose that a respondent's annual income is "measured" by simply asking how much he or she earned last year. Will this measurement be completely accurate? Of course not. Some people will lie, some will forget and give a reasonable guess, and still others will suffer from legitimate confusion about what constitutes income. Even physical variables like height, weight and blood pressure are subject to some inexactness of measurement, no matter how skilled the personnel doing the measuring. In fact, very few of the variables in the typical data set are measured completely without error.

One might think that for experimentally manipulated variables like the amount of drug administered in a biological experiment, laboratory procedures would guarantee that for all practical purposes, the amount of drug a subject receives is exactly what you think it is. But Alison Fleming (University of Toronto Psychology department) pointed out to me that when hormones are injected into a laboratory rat, the amount injected is exactly right, but due to tiny variations in needle placement, the amount actually reaching the animal's bloodstream can vary quite a bit. The same thing applies to clinical trials of drugs with humans. We will see later, though, that the statistical consequences of measurement error are not nearly as severe with experimentally manipulated variables, assuming the study is well-controlled in other respects.

Random variables that cannot be directly observed are called *latent variables*. The ones we can observe are sometimes called "manifest," but here they will be called "observed" or "observable," which is also a common usage. Upon reflection, it is clear that most of the time, we are interested in relationships among latent variables, but at best our data consist only of their imperfect, observable counterparts. One is reminded of the allegory of the cave in Plato's *Republic*, where human beings are compared to prisoners in a cave, with their heads chained so that they can only look at a wall. Behind them is a fire, which casts flickering shadows on the wall. They cannot observe reality directly; all they can see are the shadows.

A simple additive model for measurement error

Measurement error can take many forms. For categorical variables, there is *classification error*. Suppose a data file indicates whether or not each subject in a study has ever had a heart attack. Clearly, the latent Yes-No variable (whether the person has *truly* had a heart attack) does not correspond perfectly to what is in the data file, no matter how careful the assessment is. Misclassification can and does occur, in both directions.

Here, we will put classification error aside because it is technically difficult, and focus on a very simple form of measurement error that applies to continuous variables. There is a latent random variable X that cannot be observed, and a little random shock e that pushes X up or down, producing an observable random variable W. That is,

$$W = X + e \tag{24}$$

Let's say $E(X) = \mu$, $E(e) = 0$, $Var(X) = \sigma_x^2$, $Var(e) = \sigma_e^2$, and $Cov(X, e) = 0$. Because X and

e are uncorrelated,

$$Var(W) = Var(X) + Var(e) = \sigma_x^2 + \sigma_e^2.$$

Without further information, it is impossible to tell how much of the variance in the observable variable W comes from variation in the true quantity of interest, and how much comes from random noise.

In psychometric theory[11], the *reliability*[12] of a measurement is defined as the squared correlation of the true score with the observed score. Here the "true score" is X and the "observed score" is W. Recalling the definition of a correlation,

$$Corr(X,Y) = \frac{Cov(X,Y)}{SD(X)SD(Y)},$$

we have the reliability of the measurement W equal to

$$\begin{aligned}
\rho^2 &= \left(\frac{Cov(X,W)}{SD(X)SD(W)}\right)^2 \\
&= \left(\frac{\sigma_x^2}{\sqrt{\sigma_x^2}\sqrt{\sigma_x^2+\sigma_e^2}}\right)^2 \\
&= \frac{\sigma_x^4}{\sigma_x^2(\sigma_x^2+\sigma_e^2)} \\
&= \frac{\sigma_x^2}{\sigma_x^2+\sigma_e^2}. \qquad (25)
\end{aligned}$$

That is, *the reliability of a measurement is the proportion of the measurement's variance that comes from the true quantity being measured*, rather than from measurement error[13].

A reliability of one means there is no measurement error at all, while a reliability of zero means the measurement is pure noise. In the social sciences, reliabilities above 0.9 could be called excellent, from 0.8 to 0.9 good, and from 0.7 to 0.8 acceptable. Frequently, responses to single questions have reliabilities that are much less than this. To see why reliability depends on the number of questions that measure the latent variable, see Exercise **??** at the end of this section.

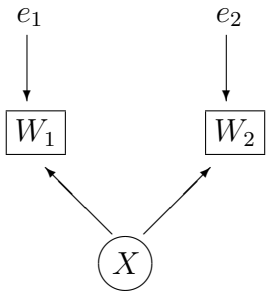
Since reliability represents quality of measurement, estimating it is an important goal. Using the definition directly is seldom possible. Reliability is the squared correlation between a latent variable and its observable counterpart, but by definition, values of the latent variable cannot be observed. On rare occasions and perhaps with great expense, it may be possible to obtain perfect or near-perfect measurements on a subset of the sample; the term *gold standard* is sometimes applied to such measurements. In that case, the reliability of the usual measurement can be estimated by a squared sample correlation between the usual measurement and the gold standard measurement. But even measurements that are called gold standard are seldom truly free of measurement error. Consequently, reliabilities that are estimated by correlating imperfect gold standards and ordinary measurements are biased downward: See Exercise **??** at the end of this section. It is clear that another approach is needed.

[11] Psychometric theory is the statistical theory of psychological measurement. The bible of psychometric theory is Lord and Novick's (1968) classic *Statistical theories of mental test scores* [**?**]. It is not too surprising that measurement error would be acknowledged and studied by psychologists. A large sector of psychological research employs "measures" of hypothetical constructs like neuroticism or intelligence (mostly paper-and-pencil tests), but no sensible person would claim that true value of such a trait is exactly the score on the test. It's true there is a famous quote "Intelligence is whatever an intelligence test measures." I have tried unsuccessfully to track down the source of this quote, and I now suspect that it is just an illustration of a philosophic viewpoint called Logical Positivism (which is how I first heard it), and not a serious statement about intelligence measurement.

[12] Reliability has a completely unrelated meaning in survival analysis, and I believe yet another meaning in statistical quality control.

[13] It's like the proportion of variance in the response variable explained by a regression, except that here the explanatory variable is the latent true score. Compare Expression (7) on Page 8.

Figure 4: Two independent measurements of a latent variable



Test-retest reliability Suppose that it is possible to make the measurement of W twice, in such a way that the errors of measurement are independent on the two occasions. We have

$$\begin{aligned} W_1 &= X + e_1 \\ W_2 &= X + e_2, \end{aligned}$$

where $E(X) = \mu$, $Var(X) = \sigma_x^2$, $E(e_1) = E(e_2) = 0$, $Var(e_1) = Var(e_2) = \sigma_e^2$, and X, e_1 and e_2 are all independent. Because $Var(e_1) = Var(e_2)$, W_1 and W_2 are called *equivalent measurements*. That is, they are contaminated by error to the same degree. Figure 4 is a path diagram of this model.

It turns out that the correlation between W_1 and W_2 is exactly equal to the reliability, and this opens the door to reasonable methods of estimation. The calculation (like many in this book) is greatly simplified by using the Centering Rule of Section 3 (see Page 4).

$$\begin{aligned} Corr(W_1, W_2) &= \frac{Cov(W_1, W_2)}{SD(W_1)SD(W_2)} \\ &= \frac{E(\overset{c}{W_1}\overset{c}{W_2})}{\sqrt{\sigma_x^2 + \sigma_e^2}\sqrt{\sigma_x^2 + \sigma_e^2}} \\ &= \frac{E(\overset{c}{X} + e_1)(\overset{c}{X} + e_2)}{\sigma_x^2 + \sigma_e^2} \\ &= \frac{E(\overset{c}{X}^2) + 0 + 0 + 0}{\sigma_x^2 + \sigma_e^2} \\ &= \frac{\sigma_x^2}{\sigma_x^2 + \sigma_e^2}, \end{aligned} \tag{26}$$

which is the reliability.

The calculation above is the basis of *test-retest reliability*[14], in which the reliability of a measurement such as an educational or psychological test is estimated by the sample correlation between two independent administrations of the test. That is, the test is given twice to the same sample of individuals, ideally with a short enough time between tests so that the trait does not really change, but long enough apart so they forget how they answered the first time.

[14]Closely related to test-retest reliability is *alternate forms reliability*, in which you correlate two equivalent versions of the test. In *split-half reliability*, you split the items of the test into two equivalent subsets and correlate them. There are also *internal consistency* estimates of reliability based on correlations among items. Assuming independent errors of measurement for split half reliability and internal consistency reliability is largely a fantasy, because both measurements are affected in the same way by short-term situational influences like mood, amount of sleep the night before, noise level, behaviour of the person administering the test, and so on.

Correlated measurement error Suppose participants remembered their wrong answers or lucky guesses from the first time they took a test, and mostly gave the same answer the second time. The result would be a positive correlation between the measurement errors e_1 and e_2. Omitted variables (see Section 5) like level of test anxiety for educational tests or desire to make a favourable impression for attitude questionnaires can also produce a positive covariance between errors of measurement. Whatever the source, positive covariance between e_1 and e_2 is an additional source of positive covariance between W_1 and W_2 that does *not* come from the latent variable X being measured. The result is an inflated estimate of reliability and an unduly rosy picture of the quality of measurement.

We will return more than once to the issue of correlated errors of measurement. For now, just notice how careful planning of the data collection (in this case, the time lag between the two administrations of the test) can eliminate or at least reduce the correlation between errors of measurement. In general, the best way to take care of correlated measurement error is with good research design.

Sample Test-retest Reliability Again, suppose it is possible to measure a variable of interest twice, in such a way that the errors of measurement are uncorrelated and have equal variance. Then the reliability may be estimated by doing this for a random sample of individuals. Let $X_1, \ldots, X_n$ be a random sample of latent variables (true scores), with $E(X_i) = \mu$ and $Var(X_i) = \sigma_x^2$. Independently for $i = 1, \ldots, n$, let

$$\begin{aligned} W_{i,1} &= X_i + e_{i,1} \\ W_{i,2} &= X_i + e_{i,2}, \end{aligned}$$

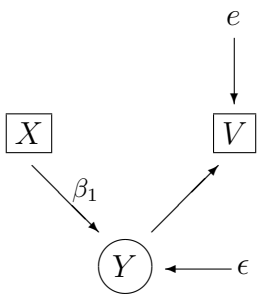
where $E(e_{i,1}) = E(e_{i,2}) = 0$, $Var(e_{i,1}) = Var(e_{i,2}) = \sigma_e^2$, and X_i, $e_{i,1}$ and $e_{i,2}$ are all independent for $i = 1, \ldots, n$. Then the sample correlation between the pairs of measurements is

$$\begin{aligned} R_n &= \frac{\sum_{i=1}^n (W_{i,1} - \overline{W}_1)(W_{i,2} - \overline{W}_2)}{\sqrt{\sum_{i=1}^n (W_{i,1} - \overline{W}_1)^2}\sqrt{\sum_{i=1}^n (W_{i,2} - \overline{W}_2)^2}} \\ &= \frac{\sum_{i=1}^n W_{i,1}W_{i,2} \; - \; n\overline{W}_1\overline{W}_2}{\sqrt{\sum_{i=1}^n W_{i,1}^2 - n\overline{W}_1^2}\sqrt{\sum_{i=1}^n W_{i,2}^2 - n\overline{W}_2^2}} \\ &= \frac{(\frac{1}{n}\sum_{i=1}^n W_{i,1}W_{i,2}) \; - \; \overline{W}_1\overline{W}_2}{\sqrt{(\frac{1}{n}\sum_{i=1}^n W_{i,1}^2) - \overline{W}_1^2}\sqrt{(\frac{1}{n}\sum_{i=1}^n W_{i,2}^2) - \overline{W}_2^2}}, \end{aligned} \tag{27}$$

where the subscript on the sample correlation coefficient R_n emphasizes that it is a function of the sample size n. By the Strong Law of Large Numbers (see Appendix **??**), we have the following:

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^n W_{i,1}W_{i,2} &\overset{a.s.}{\to} E(W_{i,1}W_{i,2}) = Cov(W_{i,1}, W_{i,2}) + E(W_{i,1})E(W_{i,2}) = \sigma_x^2 + \mu^2 \\ \overline{W}_1 &\overset{a.s.}{\to} E(W_{i,1}) = \mu \\ \overline{W}_2 &\overset{a.s.}{\to} E(W_{i,2}) = \mu \\ \frac{1}{n}\sum_{i=1}^n W_{i,1}^2 &\overset{a.s.}{\to} E(W_{i,1}^2) = Var(W_{i,1}) + (E\{W_{i,1}\})^2 = \sigma_x^2 + \sigma_e^2 + \mu^2 \\ \frac{1}{n}\sum_{i=1}^n W_{i,2}^2 &\overset{a.s.}{\to} E(W_{i,2}^2) = Var(W_{i,2}) + (E\{W_{i,2}\})^2 = \sigma_x^2 + \sigma_e^2 + \mu^2. \end{aligned}$$

Figure 5: Measurement error in the response variable



Now, since R_n is a continuous function of the various sample moments in (27) and almost sure convergence can be treated like an ordinary limit,

$$\begin{aligned} R_n &\stackrel{a.s.}{\to} \frac{\sigma_x^2 + \mu^2 - \mu^2}{\sqrt{\sigma_x^2 + \sigma_e^2 + \mu^2 - \mu^2}\sqrt{\sigma_x^2 + \sigma_e^2 + \mu^2 - \mu^2}} \\ &= \frac{\sigma_x^2}{\sigma_x^2 + \sigma_e^2} = \rho^2. \end{aligned}$$

So R_n is a strongly consistent estimator of the reliability. That is, for a large enough sample size, R_n will get arbitrarily close to the true reliability, and this happens with probability one.

8 Ignoring measurement error

Standard regression models make no provision at all for measurement error, so when such models are applied to real data, we are effectively ignoring any measurement error that may be present – pretending it's not there. This section will show that the result can be a real disaster, featuring incorrect estimates of regression parameters and Type I error probabilities approaching one as the sample size increases. Much of this material, including the history of the topic (warnings go back to at least 1936) can be found in a 2009 paper by Brunner and Austin [**?**].

Measurement error in the response variable

While ignoring measurement error in the explanatory variables can have very bad consequences, it turns out that under some conditions, measurement error in the response variable is a less serious problem.

Example 8.1

Independently for $i = 1, \ldots, n$, let

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_i + \epsilon_i \\ V_i &= \nu + Y_i + e_i, \end{aligned}$$

where $Var(X_i) = \sigma_x^2$, $Var(e_i) = \sigma_e^2$, $Var(\epsilon_i) = \sigma_\epsilon^2$, and X_i, e_i, ϵ_i are all independent. Figure 5 is a path diagram of this model.

In Example 8.1, the explanatory variable X_i is observable, but the response variable Y_i is latent. Instead of Y_i, we can see V_i, which is Y_i plus a piece of random noise, and also plus a constant ν that represents the difference between the expected value of the latent random variable and the expected value of its observable counterpart. This constant term could be called *measurement bias*. For example, if Y is true amount of exercise in minutes and V is reported exercise, the measurement bias ν is population mean exaggeration, in minutes.

Since Y_i cannot be observed, V_i is used in its place, and the data analyst fits the *naive* model

$$V_i = \beta_0 + \beta_1 X_i + \epsilon_i.$$

Studying Mis-specified Models The "naive model" above is an example of a model that is *mis-specified*. That is, the model says that the data are being generated in a particular way, but this is not how the data are actually being produced. Generally speaking, correct models will usually yield better results than incorrect models, but it's not that simple. In reality, most statistical models are imperfect. The real question is how much any given imperfection really matters. As Box and Draper (1987, p. 424) put it, "Essentially all models are wrong, but some are useful." [**?**]

So, it is not enough to complain that a statistical model is incorrect, or unrealistic. To make the point convincingly, one must show that by being wrong in a particular way, the model can yield results that are misleading. To do this, it is necessary to have a specific *true model* in mind; typically the so-called true model is one that is obviously more believable than the model being challenged. Then, one can examine estimators or test statistics based on the mis-specified model, and see how they behave when the true model holds. We have already done this in Section 5 in connection with omitted variables; see Example 5.1 starting on Page 13.

Under the true model of Example 8.1 (measurement error in the response variable only), we have $Cov(X,Y) = \beta_1\sigma_x^2$ and $Var(X) = \sigma_x^2$. Then,

$$\begin{aligned}
\widehat{\beta}_1 &= \frac{\sum_{i=1}^n (X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^n (X_i - \overline{X})^2} \\
&= \frac{\widehat{\sigma}_{x,y}}{\widehat{\sigma}_x^2} \\
&\stackrel{a.s.}{\to} \frac{Cov(X,Y)}{Var(X)} \\
&= \frac{\beta_1\sigma_x^2}{\sigma_x^2} \\
&= \beta_1.
\end{aligned}$$

That is, measurement error in the response variable causes no asymptotic bias. Even when the model is mis-specified by assuming that the response variable is measured without error, the ordinary least squares estimate of the slope is consistent. There is a general lesson here about mis-specified models. Mis-specification (using the wrong model) is not always a problem; sometimes everything works out fine.

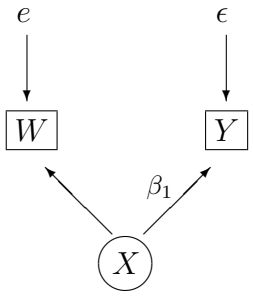
Let's see why the naive model works so well here. The response variable under the true model may be re-written

$$\begin{aligned}
V_i &= \nu + Y_i + e_i \\
&= \nu + (\beta_0 + \beta_1 X_i + \epsilon_i) + e_i \\
&= (\nu + \beta_0) + \beta_1 X_i + (\epsilon_i + e_i) \\
&= \beta_0' + \beta_1 X_i + \epsilon_i' \qquad (28)
\end{aligned}$$

What has happened here is a *re-parameterization* (not a one-to-one reparameterization), in which the pair (ν, β_0) is absorbed into β_0', and $Var(\epsilon_i + e_i) = \sigma_\epsilon^2 + \sigma_e^2$ is absorbed into a single unknown variance that will probably be called σ^2. It is true that ν and β_0 will never be knowable separately, and also σ_ϵ^2 and σ_e^2 will never be knowable separately. But that really doesn't matter, because the true interest is in β_1.

In this book and in standard statistical practice, there are many models in which the response variable appears to be measured without error. But error-free measurement is a rarity at best, so these models should be viewed as re-parameterized versions of models that do acknowledge the

Figure 6: Measurement error in the explanatory variable



reality of measurement error in the response variable. A critical feature of these re-parameterized models is that the measurement error is assumed independent of everything else in the model. When this fails, there is usually trouble.

Measurement error in the explanatory variable

Example 8.2

Independently for $i = 1, \ldots, n$, let

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_i + \epsilon_i \\ W_i &= X_i + e_i, \end{aligned}$$

where $Var(X_i) = \sigma_x^2$, $Var(e_i) = \sigma_e^2$, $Var(\epsilon_i) = \sigma_\epsilon^2$, and X_i, e_i, ϵ_i are all independent. Figure 6 is a path diagram of the model.

Unfortunately, the explanatory variable X_i cannot be observed; it is a *latent* variable. So instead W_i is used in its place, and the data analyst fits the *naive* model

$$Y_i = \beta_0 + \beta_1 W_i + \epsilon_i.$$

Under the naive model of Example 8.2, the ordinary least squares estimate of β_1 is

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^n (W_i - \overline{W})(Y_i - \overline{Y})}{\sum_{i=1}^n (W_i - \overline{W})^2} = \frac{\widehat{\sigma}_{w,y}}{\widehat{\sigma}_w^2}.$$

Now regardless of what model is correct, $\widehat{\sigma}_{w,y} \stackrel{a.s.}{\to} Cov(W,Y)$ and $\widehat{\sigma}_w^2 \stackrel{a.s.}{\to} Var(W)$[15], so that by the continuous mapping property of ordinary limits[16], $\widehat{\beta}_1 \stackrel{a.s.}{\to} \frac{Cov(W,Y)}{Var(W)}$.

Let us assume that the true model holds. In that case,

$$Cov(W,Y) = \beta_1 \sigma_x^2 \quad \text{and} \quad Var(W) = \sigma_x^2 + \sigma_e^2.$$

Consequently,

$$\begin{aligned} \widehat{\beta}_1 &= \frac{\sum_{i=1}^n (W_i - \overline{W})(Y_i - \overline{Y})}{\sum_{i=1}^n (W_i - \overline{W})^2} \\ &= \frac{\widehat{\sigma}_{w,y}}{\widehat{\sigma}_w^2} \\ &\stackrel{a.s.}{\to} \frac{Cov(W,Y)}{Var(W)} \\ &= \beta_1 \left(\frac{\sigma_x^2}{\sigma_x^2 + \sigma_e^2} \right). \end{aligned} \tag{29}$$

[15]This is true because sample variances and covariances are strongly consistent estimators of the corresponding population quantities; see Section **??** in Appendix **??**.

[16]Almost sure convergence acts like an ordinary limit, applying to all points in the underlying sample space, except possibly a set of probability zero. If you wanted to do this problem strictly in terms of convergence in probability, you could use the Weak Law of Large Numbers and then use Slutsky Lemma **??** of Appendix **??**.

Figure 7: Two explanatory variables measured with error

includegraphics[width=4.5in]Pictures/MeReg2Path

So when the fuzzy explanatory variable W_i is used instead of the real thing, $\widehat{\beta}_1$ converges not to the true regression coefficient, but to the true regression coefficient multiplied by the reliability of W_i. That is, it's biased, even as the sample size approaches infinity. It is biased toward zero, because reliability is between zero and one. The worse the measurement of X, the more the asymptotic bias.

What happens to $\widehat{\beta}_1$ in (29) is sometimes called *attenuation*, or weakening, and in this case that's what happens. The measurement error weakens the apparent relationship between X_1 and Y. If the reliability of W can be estimated from other data (and psychologists are always trying to estimate reliability), then the sample regression coefficient can be "corrected for attentuation." Sample correlation coefficients are sometimes corrected for attenuation too.

Now typically, social and biological sientists are not really interested in point estimates of regression coefficients. They only need to know whether they are positive, negative or zero. So the idea of attenuation sometimes leads to a false sense of security about measurement error. It's natural to think that all it does is to weaken what's really there, so if you can reject the null hypothesis and conclude that a relationship is present even with measurement error, you would have reached the same conclusion if the explanatory variables had not been measured with error.

Unfortunately, it's not so simple. The reasoning above is okay if there is just one explanatory variable, but we will see that with two or more explanatory variables the effects of measurement error are far more serious and potentially misleading.

Two Explanatory Variables

In Example 8.2, we saw that measurement error in the explanatory variable causes the estimated regression coefficient $\widehat{\beta}_1$ to be biased toward zero as $n \to \infty$. Bias toward zero weakens the apparent relationship between X and Y; and if $\beta_1 = 0$, there is no asymptotic bias. So for the case of a single explanatory variable measured with error, the sample relationships still reflect population relationships, with the sample relationships being weaker because of inexact measurement. But this only holds for regression with a single explanatory variable. Measurement error causes a lot more trouble for multiple regression. In this example, there are two explanatory variables, both measured with error.

Example 8.3

Independently for $i = 1, \ldots, n$,

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \epsilon_i \\ W_{i,1} &= X_{i,1} + e_{i,1} \\ W_{i,2} &= X_{i,2} + e_{i,2}, \end{aligned}$$

where where $E(X_{i,1}) = \mu_1$, $E(X_{i,2}) = \mu_2$, $E(\epsilon_i) = E(e_{i,1}) = E(e_{i,2}) = 0$, $Var(\epsilon_i) = \sigma^2$, $Var(e_{i,1}) = \omega_1$, $Var(e_{i,2}) = \omega_2$, the errors $\epsilon_i, e_{i,1}$ and $e_{i,2}$ are all independent, $X_{i,1}$ is independent of $\epsilon_i, e_{i,1}$ and $e_{i,2}$, $X_{i,2}$ is independent of $\epsilon_i, e_{i,1}$ and $e_{i,2}$, and

$$Var\begin{pmatrix} X_{i,1} \\ X_{i,1} \end{pmatrix} = \begin{pmatrix} \phi_{11} & \phi_{12} \\ \phi_{12} & \phi_{22} \end{pmatrix}.$$

Figure 7 shows the path diagram.

Again, because the actual explanatory variables $X_{i,1}$ and $X_{i,2}$ are latent variables that cannot be observed, $W_{i,1}$ and $W_{i,2}$ are used in their place. The data analyst fits the *naive* model

$$Y_i = \beta_0 + \beta_1 W_{i,1} + \beta_2 W_{i,2} + \epsilon_i.$$

An attractive feature of multiple regression is its ability to represent the relationship of one or more explanatory variables to the response variable, while *controlling for* other explanatory varables. In fact, this is the biggest appeal of multiple regression and similar methods for non-experimental data. In Example 8.3, our interest is in the relationship of X_2 to Y controlling for X_1. The main objective is to test $H_0 : \beta_2 = 0$, but we are also interested in the estimation of β_2.

We will try the same approach that worked for Example 8.2, calculating $\widehat{\beta}_2$ assuming the naive model, and then examining how $\widehat{\beta}_2$ behaves as $n \to \infty$ when the true model holds. We want to express $\widehat{\beta}_2$ in terms of sample variances and covariances, because they converge to the corresponding population variances and covariances as $n \to \infty$, and it is easy to calculate population variances and covariances under the true model. To keep the calculations fairly simple, it is helpful to center the explanatory variables and the response variable by subtracting off sample means. That is, $W_{i,1}$ is replaced by $(W_{i,1} - \overline{W}_1)$, $W_{i,2}$ is replaced by $(W_{i,2} - \overline{W}_2)$, and Y_i is replaced by $(Y_i - \overline{Y})$.

Think of fitting a plane to a 3-dimensional scatterplot, in such a way that the sum of squared vertical distances from the points to the plane is minimized. Clearly, subtracting off means does not alter the relative positions of the points, nor does it affect the orientation (slopes) of the best-fitting plane. All it does is to shift the axies, so that the origin is the point $(\overline{W}_1, \overline{W}_2, \overline{Y})$ and the equation of the best-fitting plane has no intercept. Then, the familar formula $\widehat{\boldsymbol{\beta}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y}$ (but with $\mathbf{W}$ instead of $\mathbf{X}$) will yield the desired regression coefficients.

Adopting a notation that will be used throughout the book, denote one of the n vectors of observable data by $\mathbf{D}_i$. Here,

$$\mathbf{D}_i = \begin{pmatrix} W_{i,1} \\ W_{i,2} \\ Y_i \end{pmatrix}.$$

Then, let $\boldsymbol{\Sigma} = [\sigma_{i,j}] = cov(\mathbf{D}_i)$. Corresponding to $\boldsymbol{\Sigma}$ is the sample variance covariance matrix $\widehat{\boldsymbol{\Sigma}} = [\widehat{\sigma}_{i,j}]$, with n rather than $n-1$ in the denominators. To make this setup completely explicit,

$$\boldsymbol{\Sigma} = cov \begin{pmatrix} W_{i,1} \\ W_{i,2} \\ Y_i \end{pmatrix} = \begin{pmatrix} \sigma_{1,1} & \sigma_{1,2} & \sigma_{1,3} \\ \sigma_{1,2} & \sigma_{2,2} & \sigma_{2,3} \\ \sigma_{1,3} & \sigma_{2,3} & \sigma_{3,3} \end{pmatrix}$$

Calculating the regression coefficients is straightforward.

$$\begin{aligned} \mathbf{W}^\top\mathbf{W} &= \begin{pmatrix} \sum_{i=1}^n (W_{i,1} - \overline{W}_1)^2 & \sum_{i=1}^n (W_{i,1} - \overline{W}_1)(W_{i,2} - \overline{W}_2) \\ \sum_{i=1}^n (W_{i,1} - \overline{W}_1)(W_{i,2} - \overline{W}_2) & \sum_{i=1}^n (W_{i,2} - \overline{W}_2)^2 \end{pmatrix} \\ &= n \begin{pmatrix} \widehat{\sigma}_{1,1} & \widehat{\sigma}_{1,2} \\ \widehat{\sigma}_{1,2} & \widehat{\sigma}_{2,2} \end{pmatrix} \\ \mathbf{W}^\top\mathbf{Y} &= \begin{pmatrix} \sum_{i=1}^n (W_{i,1} - \overline{W}_1)(Y_i - \overline{Y}) \\ \sum_{i=1}^n (W_{i,2} - \overline{W}_2)(Y_i - \overline{Y}) \end{pmatrix} \\ &= n \begin{pmatrix} \widehat{\sigma}_{1,3} \\ \widehat{\sigma}_{2,3} \end{pmatrix} \end{aligned}$$

Then with a bit of simplification[17],

$$\widehat{\boldsymbol{\beta}} = (\mathbf{W}^\top\mathbf{W})^{-1}\mathbf{W}^\top\mathbf{Y} = \begin{pmatrix} \widehat{\beta}_1 \\ \widehat{\beta}_2 \end{pmatrix} = \begin{pmatrix} \frac{\widehat{\sigma}_{22}\widehat{\sigma}_{13} - \widehat{\sigma}_{12}\widehat{\sigma}_{23}}{\widehat{\sigma}_{11}\widehat{\sigma}_{22} - \widehat{\sigma}_{12}^2} \\ \\ \frac{\widehat{\sigma}_{11}\widehat{\sigma}_{23} - \widehat{\sigma}_{12}\widehat{\sigma}_{13}}{\widehat{\sigma}_{11}\widehat{\sigma}_{22} - \widehat{\sigma}_{12}^2} \end{pmatrix}.$$

Because sample variances and covariances are strongly consistent estimators of the corresponding

[17] Okay, I admit it's brutal. I used Sage; see Appendix **??**

population quantities,

$$\widehat{\beta}_2 = \frac{\widehat{\sigma}_{11}\widehat{\sigma}_{23} - \widehat{\sigma}_{12}\widehat{\sigma}_{13}}{\widehat{\sigma}_{11}\widehat{\sigma}_{22} - \widehat{\sigma}_{12}^2} \xrightarrow{a.s.} \frac{\sigma_{11}\sigma_{23} - \sigma_{12}\sigma_{13}}{\sigma_{11}\sigma_{22} - \sigma_{12}^2}. \tag{30}$$

This convergence holds provided that the denominator $\sigma_{11}\sigma_{22} - \sigma_{12}^2 \neq 0$. The denominator is a determinant:

$$\sigma_{11}\sigma_{22} - \sigma_{12}^2 = \left| V \begin{pmatrix} W_{i,1} \\ W_{i,2} \end{pmatrix} \right|.$$

It will be non-zero provided at least one of

$$cov \begin{pmatrix} X_{i,1} \\ X_{i,2} \end{pmatrix} \quad \text{and} \quad cov \begin{pmatrix} e_{i,1} \\ e_{i,2} \end{pmatrix}$$

is positive definite – not a lot to ask.

The convergence of $\widehat{\beta}_2$ in Expression 30 applies regardless of what model is correct. To see what happens when the true model of Example 8.3 holds, we calculate the $\boldsymbol{\Sigma}$, the common variance-covariance matrix of the observable data vectors.

$$\begin{aligned}
\boldsymbol{\Sigma} &= cov \begin{pmatrix} W_{i,1} \\ W_{i,2} \\ Y_i \end{pmatrix} \\
&= \begin{pmatrix} \sigma_{1,1} & \sigma_{1,2} & \sigma_{1,3} \\ \sigma_{1,2} & \sigma_{2,2} & \sigma_{2,3} \\ \sigma_{1,3} & \sigma_{2,3} & \sigma_{3,3} \end{pmatrix} \\
&= \begin{pmatrix} \omega_1 + \phi_{11} & \phi_{12} & \beta_1\phi_{11} + \beta_2\phi_{12} \\ \phi_{12} & \omega_2 + \phi_{22} & \beta_1\phi_{12} + \beta_2\phi_{22} \\ \beta_1\phi_{11} + \beta_2\phi_{12} & \beta_1\phi_{12} + \beta_2\phi_{22} & \beta_1^2\phi_{11} + 2\,\beta_1\beta_2\phi_{12} + \beta_2^2\phi_{22} + \psi \end{pmatrix}
\end{aligned}$$

Subsituting into expression 30 and simplifying[18], we obtain

$$\begin{aligned}
\widehat{\beta}_2 &\xrightarrow{a.s.} \frac{\sigma_{11}\sigma_{23} - \sigma_{12}\sigma_{13}}{\sigma_{11}\sigma_{22} - \sigma_{12}^2} \\
&= \frac{(\beta_1\omega_1\phi_{12} + \beta_2\omega_1\phi_{22} + \beta_2\phi_{11}\phi_{22} - \beta_2\phi_{12}^2)}{(\omega_1\omega_2 + \omega_1\phi_{22} + \omega_2\phi_{11} + \phi_{11}\phi_{22} - \phi_{12}^2)} \\
&= \beta_2 + \frac{\beta_1\omega_1\phi_{12} + \beta_2\omega_2(\phi_{11} - \omega_1)}{(\phi_{1,1} + \omega_1)(\phi_{2,2} + \omega_2) - \phi_{12}^2}
\end{aligned} \tag{31}$$

By the asymptotic normality of the sample variance-covariance matrix (see Appendix **??**), $\widehat{\beta}_2$ has a distribution that is approximately normal for large samples, with approximate mean given by expression (31). Thus, it makes sense to call the second term in (31) the *asymptotic bias*. It is also the amount by which the estimate of β_2 will be wrong as $n \to \infty$.

Clearly, this situation is much more serious than the bias toward zero detected for the case of one explanatory variable. With two explanatory variables, the bias can be positive, negative or zero depending on the values of other unknown parameters.

In particular, consider the problems associated with testing $H_0 : \beta_2 = 0$. The purpose of this test is to determine whether, controlling for X_1, X_2 has any relationship to Y. The supposed ability of multiple regression to answer questions like this is the one of the main reasons it is so widely used in practice. So when measurement error makes this kind of inference invalid, it is a real problem.

Suppose that the null hypothesis is true, so $\beta_2 = 0$. In this case, Expression (31) becomes

$$\widehat{\beta}_2 \xrightarrow{a.s.} \frac{\beta_1\omega_1\phi_{12}}{(\phi_{1,1} + \omega_1)(\phi_{2,2} + \omega_2) - \phi_{12}^2}.$$

Recall that β_1 is the link between X_1 and Y, $\omega_1 = Var(e_1)$ is the variance of measurement error in X_1, and ϕ_{12} is the covariance between X_1 and X_2. Thus, when $H_0 : \beta_2 = 0$ is true, $\widehat{\beta}_2$ converges

[18] It's a lot easier with Sage or some other computer algebra software

to a non-zero quantity unless

- There is no relationship between X_1 and Y, or
- There is no measurement error in W_1, or
- There is no correlation between X_1 and X_2.

Brunner and Austin [**?**] have shown that whether H_0 is true or not, the standard error of $\widehat{\beta}_2$ goes to zero, and when the large-sample target of $\widehat{\beta}_2$ is non-zero, the p-value goes almost surely to zero. That is, the probability of making a Type I error goes to one because of measurement error in an explanatory variable — not the one being tested, but the one for which one is "controlling."

9 Modeling measurement error

It is clear that ignoring measurement error in regression can yield conclusions that are very misleading. But as soon as we try building measurement error into the statistical model, we encounter a technical issue that will occupy a central role in this book: parameter identifiability. For comparison, first consider a regression model without measurement error, where everything is nice. This is not quite the standard model, because the explanatory variables are random variables. General principles arise right away, so definitions will be provided as we go.

A first try at including measurement error

The following is basically the true model of Example 8.2, with everything normally distributed. Independently for $i = 1, \ldots, n$, let

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_i + \epsilon_i \\ W_i &= \nu + X_i + e_i, \end{aligned} \tag{32}$$

where

- X_i is normally distributed with mean μ_x and variance $\phi > 0$
- ϵ_i is normally distributed with mean zero and variance $\psi > 0$
- e_i is normally distributed with mean zero and variance $\omega > 0$
- X_i, e_i, ϵ_i are all independent.

The intercept term ν could be called "measurement bias." If X_i is true amount of exercise per week and W_i is reported amount of exercise per week, ν is the average amount by which people exaggerate.

Data from Model (32) are just the pairs (W_i, Y_i) for $i = 1, \ldots, n$. The true explanatory variable X_i is a latent variable whose value cannot be known exactly. The model implies that the (W_i, Y_i) are independent bivariate normal with

$$E\begin{pmatrix} W_i \\ Y_i \end{pmatrix} = \boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} = \begin{pmatrix} \mu_x + \nu \\ \beta_0 + \beta_1 \mu_x \end{pmatrix},$$

and variance covariance matrix

$$V\begin{pmatrix} W_i \\ Y_i \end{pmatrix} = \boldsymbol{\Sigma} = [\sigma_{i,j}] = \begin{pmatrix} \phi + \omega & \beta_1 \phi \\ \beta_1 \phi & \beta_1^2 \phi + \psi \end{pmatrix}.$$

There is a big problem here, and the moment structure equations reveal it.

$$\begin{aligned} \mu_1 &= \mu_x + \nu \\ \mu_2 &= \beta_0 + \beta_1\mu_x \\ \sigma_{1,1} &= \phi + \omega \\ \sigma_{1,2} &= \beta_1\phi \\ \sigma_{2,2} &= \beta_1^2\phi + \psi. \end{aligned} \tag{33}$$

It is impossible to solve these five equations for the seven model parameters[19]. That is, even with perfect knowledge of the probability distribution of the data (for the multivariate normal, that means knowing $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, period), it would be impossible to know the model parameters.

To make the problem clearer, look at the table below. It shows two diferent set of parameter values $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ that both yield the same mean vector and covariance matrix, and hence the exact same distribution of the observable data.

	μ_x	β_0	ν	β_1	ϕ	ω	ψ
$\boldsymbol{\theta}_1$	0	0	0	1	2	2	3
$\boldsymbol{\theta}_2$	0	0	0	2	1	3	1

Both $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ imply a bivariate normal distribution with mean zero and covariance matrix

$$\boldsymbol{\Sigma} = \begin{pmatrix} 4 & 2 \\ 2 & 5 \end{pmatrix},$$

and thus the same distribution of the sample data.

No matter how large the sample size, it will be impossible to decide between $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$, because they imply exactly the same probability distribution of the observable data. The problem here is that the parameters of Model (32) are not *identifiable*. This calls for a brief discussion of identifiability, a topic of central importance in structural equation modeling.

10 Parameter Identifiability

The Basic Idea Suppose we have a vector of observable data $\mathbf{D} = (D_1, \ldots, D_n)$, and a statistical model (a set of assertions implying a probability distribution) for $\mathbf{D}$. The model depends on a parameter θ, which is usually a vector. If the probability distribution of $\mathbf{D}$ corresponds uniquely to θ, then we say that the parameter vector is *identifiable*. But if any two different parameter values yield the same probability distribution, then the parameter vector is not identifiable. In this case, the data cannot be used to decide between the two parameter values, and standard methods of parameter estimation will fail. Even an infinite amount of data cannot tell you the true parameter values.

Definition 10.1 *A* Statistical Model *is a set of assertions that partly*[20] *specify the probability distribution of a set of observable data.*

Definition 10.2 *Suppose a statistical model implies* $\mathbf{D} \sim P_{\boldsymbol{\theta}}, \boldsymbol{\theta} \in \Theta$. *If no two points in* Θ *yield the same probability distribution, then the parameter* $\boldsymbol{\theta}$ *is said to be* identifiable. *On the other hand, if there exist* $\boldsymbol{\theta}_1$ *and* $\boldsymbol{\theta}_2$ *in* Θ *with* $P_{\boldsymbol{\theta}_1} = P_{\boldsymbol{\theta}_2}$, *the parameter* $\boldsymbol{\theta}$ *is* not identifiable.

A good example of non-identifiability appears in Section 5 on omitted variables in regression. There, the correct model has a set of infinitely many parameter values leading to exactly the same probability distribution for the observed data.

[19] That's a strong statement, and a strong Theorem is coming to justify it.

[20] Suppose that the distribution is assumed known except for the value of a parameter vector $\boldsymbol{\theta}$. So the distribution is "partly" specified.

Figure 8: Two parameters values yielding the same probability distribution

Theorem 1 *If the parameter vector is not identifiable, consistent estimation for all points in the parameter space is impossible.*

In Figure 8, θ_1 and θ_2 are two distinct sets of parameter values for which the distribution of the observable data is the same.
Let T_n be a estimator that is consistent for both θ_1 and θ_2. What this means is that if θ_1 is the correct parameter value, eventually as n increases, the probability distribution of T_n will be concentrated in the circular neighborhood around θ_1. And if θ_1 is the correct parameter value, it the probability distribution will be concentrated around θ_2.

But the probability distribution of the data, and hence of T_n (a function of the data) is identical for θ_1 and θ_2. This means that for a large enough sample size, most of T_n's probability distribution must be concentrated in the neighborhood around θ_1, and at the same time it must be concentrated in the neighborhood around θ_2. This is impossible, since the two regions do not overlap. Hence there can be no such consistent estimator T_n.

Theorem 1 says why parameter identifiability is so important. Without it, even an infinite amount of data cannot reveal the values of the parameters.

Surprisingly often, whether a set of parameter values can be recovered from the moments depends on where in the parameter space those values are located. That is, the parameter vector may be identifiable at some points but not others.

Definition 10.3 *The parameter is said to be* identifiable *at a point $\boldsymbol{\theta}_0$ if no other point in Θ yields the same probability distribution as $\boldsymbol{\theta}_0$.*

If the parameter is identifiable at at every point in Θ, it is identifiable, or *globally* (as opposed to locally) identifiable.

Definition 10.4 *The parameter is said to be* locally identifiable *at a point $\boldsymbol{\theta}_0$ if there is a neighbourhood of points surrounding $\boldsymbol{\theta}_0$, none of which yields the same probability distribution as $\boldsymbol{\theta}_0$.*

Obviously, local identifiability at a point is a necessary condition for global identifiability there.

It is possible for individual parameters (or other functions of the parameter vector) to be identifiable even when the entire parameter vector is not.

Definition 10.5 *Let $g(\boldsymbol{\theta})$ be a function of the parameter vector. If $g(\boldsymbol{\theta}_0) \neq g(\boldsymbol{\theta})$ implies $P_{\boldsymbol{\theta}_0} \neq P_{\boldsymbol{\theta}}$ for all $\boldsymbol{\theta} \in \Theta$, then the function $g(\boldsymbol{\theta})$ is said to be identifiable at the point $\boldsymbol{\theta}_0$.*

For example, let $D_1, \ldots, D_n$ be i.i.d. Poisson random variables with mean $\lambda_1 + \lambda_2$, where $\lambda_1 > 0$ and $\lambda_1 > 0$. The parameter is the pair $\boldsymbol{\theta} = (\lambda_1, \lambda_2)$. The parameter is not identifiable because any pair of λ values satisfying $\lambda_1 + \lambda_2 = c$ will produce exactly the same probability distribution. Notice also how maximum likelihood estimation will fail in this case; the likelihood

function will have a ridge, a non-unique maximum along the line $\lambda_1 + \lambda_2 = \overline{D}$, where $\overline{D}$ is the sample mean. The function $g(\boldsymbol{\theta}) = \lambda_1 + \lambda_2$, of course, is identifiable.

The failure of maximum likelihood for the Poisson example is very typical of situations where the parameter is not identifiable. Collections of points in the parameter space yield the same probability distribution of the observable data, and hence identical values of the likelihood. Usually these form connected sets of infinitely many points, and when a numerical likelihood search reaches such a higher-dimensional ridge or plateau, the software checks to see if it's a maximum, and (if it's good software) complains loudly because the maximum is not unique. The complaints might take unexpected forms, like a statement that the Hessian has negative eigenvalues. But in any case, maximum likelihood estimation fails.

The idea of a *function* of the parameter vector covers a lot of territory. It includes individual parameters and sets of parameters, as well as things like products and ratios of parameters. Look at the moment structure equations (33) that come from the regression Model (32). If $\sigma_{1,2} = 0$, this means $\beta_1 = 0$, because ϕ is a variance, and is greater than zero. Also in this case $\psi = \sigma_{2,2}$ and $\beta_0 = \mu_2$. So, the function $g(\boldsymbol{\theta}) = (\beta_0, \beta_1, \psi)$ is identifiable at all points in the parameter space where $\beta_1 = 0$.

Recall how for the regression Model (32), the moment structure equations (33) consist of five equations in seven unknown parameters. It was shown by a numerical example that there were two different sets of parameter values that produced the same mean vector and covariance matrix, and hence the same distribution of the observable data. Actually, infinitely many parameter values produce the same distribution, and it happens because there are more unknowns than equations. Theorem 2 is a strictly mathematical theorem[21] that provides the necessary details.

Theorem 2 *Let*

$$\begin{aligned} y_1 &= f_1(x_1, \ldots, x_p) \\ y_2 &= f_2(x_1, \ldots, x_p) \\ \vdots & \qquad \vdots \\ y_q &= f_q(x_1, \ldots, x_p), \end{aligned}$$

If the functions $f_1, \ldots, f_q$ are analytic (posessing a Taylor expansion) and $p > q$, the set of points $(x_1, \ldots, x_p)$ where the system of equations has a unique solution occupies at most a set of volume zero in $\mathbb{R}^p$.

The following corollary to Theorem 2 is the fundamental necessary condition for parameter identifiability. It will be called the **Parameter Count Rule**.

Rule 1 *Suppose identifiability is to be decided based on a set of moment structure equations. If there are more parameters than equations, the parameter vector is identifiable on at most a set of volume zero in the parameter space.*

When the data are multivariate normal (and this will frequently be assumed), then the distribution of the sample data corresponds exactly to the mean vector and covariance matrix, and to say that a parameter value is identifiable means that is can be recovered from elements of the mean vector and covariance matrix. Most of the time, that involves trying to solve the moment structure equations or covariance structure equations for the model parameters.

Even when the data are not assumed multivariate normal, the same process makes sense. Classical structural equation models, including models for regression with measurement error, are based on systems of simultaneous linear equations. Assuming simple random sampling from a large population, the observable data are independent and identically distributed, with a mean vector $\boldsymbol{\mu}$ and a covariance matrix $\boldsymbol{\Sigma}$ that may be written as functions of the model parameters

[21] The core of the proof may be found in Appendix 5 of Fisher (1966).

in a straightforward way. If it is possible to solve uniquely for a given model parameter in terms of the elements of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, then that parameter is a function of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, which in turn are functions of the probability distribution of the data. A function of a function is a function, and so the parameter is a function of the probability distribution of the data. Hence, it is identifiable.

Another way to reach this conclusion is to observe that if it is possible to solve for the parameters in terms of moments, simply "putting hats on everything" yields Method of Moments estimator. These estimators, though they may be less than ideal in some ways, will still usually be consistent by the Law of Large Numbers and continuous mapping. Theorem 1 tells us consistency would be impossible if the parameters were not identifiable.

To summarize, we have arrived at the standard way to check parameter identifiability for any linear simultaneous equation model, not just measurement error regression. *First, calculate the expected value and covariance matrix of the observable data, as a function of the model parameters. If it is possible to solve uniquely for the model parameters in terms of the means, variances and covariances of the observable data, then the model parameters are identifiable.*

If two distinct parameter vectors yield the same pair $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and the distribution is multivariate normal, the parameter vector is clearly not identifiable. When the distribution is *not* multivariate normal this conclusion does not necessarily follow; the parameters might be recoverable from higher moments, or possibly from the moment-generating function or characteristic function.

But this would require knowing exactly what the non-normal distribution of the data might be. When it comes to analyzing actual data using linear models like the ones in this book, there are really only two alternatives. Either the distribution is assumed[22] normal, or it is acknowledged to be completely unknown. In both cases, parameters will either be identifiable from the mean and covariance matrix (usually just the covariance matrix), or they will not be identifiable at all.

The conclusion is that in practice, "identifiable" means identifiable from the moments. This explains why the Parameter Count Rule (Rule 1) is frequently used to label parameters "not identifiable" even when there is no assumption of normality.

11 Double measurement

Consider again the model of Expression (32), a simple regression with measurement error in the single explanatory variable. This is a tiny example of something that occurs all too frequently in practice. The statistician or scientist has a data set that seems relevant to a particular topic, and a model for the observable data that is more or less reasonable. But the parameters of the model cannot be identified from the distribution of the data. In such cases, valid inference is very challenging, if indeed it is possible at all.

The best way out of this trap is to avoid getting trapped in the first place. Plan the statistical analysis in advance, and ensure identifiability by collecting the right kind of data. Double measurement is a straightforward way to get the job done. The key is to measure the explanatory variables twice, preferably using different methods or measuring instruments.

A scalar example

Instead of measuring the explanatory variable only once, suppose we had a second, independent measurement; "independent" means that the measurement errors are statistically independent of one another. Perhaps the two measurements are taken at different times, using different

[22] Even when the the data are clearly not normal, methods – especially likelihood ratio tests – based on a normal model can work quite well. For example,

instruments or methods. Then we have the following model. Independently for $i = 1, \ldots, n$, let

$$\begin{aligned} W_{i,1} &= \nu_1 + X_i + e_{i,1} \\ W_{i,2} &= \nu_2 + X_i + e_{i,2} \\ Y_i &= \beta_0 + \beta_1 X_i + \epsilon_i, \end{aligned} \tag{34}$$

where

- X_i is normally distributed with mean μ_x and variance $\phi > 0$
- ϵ_i is normally distributed with mean zero and variance $\psi > 0$
- $e_{i,1}$ is normally distributed with mean zero and variance $\omega_1 > 0$
- $e_{i,2}$ is normally distributed with mean zero and variance $\omega_2 > 0$
- $X_i, e_{i,1}, e_{i,2}$ and ϵ_i are all independent.

The model implies that the triples $\mathbf{D}_i = (W_{i,1}, W_{i,2}, Y_i)^\top$ are multivarate normal with

$$E(\mathbf{D}_i) = E\begin{pmatrix} W_{i,1} \\ W_{i,1} \\ Y_i \end{pmatrix} = \begin{pmatrix} \mu_x + \nu_1 \\ \mu_x + \nu_2 \\ \beta_0 + \beta_1\mu_x \end{pmatrix},$$

and variance covariance matrix

$$cov(\mathbf{D}_i) = \mathbf{\Sigma} = [\sigma_{i,j}] = \begin{pmatrix} \phi + \omega_1 & \phi & \beta_1\phi \\ & \phi + \omega_2 & \beta_1\phi \\ & & \beta_1^2\phi + \psi \end{pmatrix}. \tag{35}$$

Here are some comments.

- There are now nine moment structure equations in nine unknown parameters. This model passes the test of the Parameter Count Rule, meaning that identifiability is possible, but not guaranteed.
- Notice that the model dictates $\sigma_{1,3} = \sigma_{2,3}$. This *model-induced constraint* upon $\mathbf{\Sigma}$ is testable. If $H_0 : \sigma_{1,3} = \sigma_{2,3}$ were rejected, the correctness of the model would be called into question[23]. Thus, the study of parameter identifiability leads to a useful test of model fit.
- The constraint $\sigma_{1,3} = \sigma_{2,3}$ allows two solutions for β_1 in terms of the moments: $\beta_1 = \sigma_{13}/\sigma_{12}$ and $\beta_1 = \sigma_{23}/\sigma_{12}$. Does this mean the solution for β_1 is not "unique?" No; everything is okay. Because $\sigma_{1,3} = \sigma_{2,3}$, the two solutions are actually the same. If a parameter can be recovered from the moments in any way at all, it is identifiable.
- For the other model parameters appearing in the covariance matrix, the additional measurement of the explanatory variable also appears to have done the trick. It is easy to solve for ϕ, ω_1, ω_2 and ψ in terms of $\sigma_{i,j}$ values. Thus, these parameters are identifiable.
- On the other hand, the additional measurement did not help with the means and intercepts *at all*. Even assuming β_1 known because it can be recovered from $\mathbf{\Sigma}$, the remaining three linear equations in four unknowns have infinitely many solutions. There are still infinitely many solutions if $\nu_1 = \nu_2$.

[23] Philosophers of science agree that *falsifiability* – the possibility that a scientific model can be challenged by empirical data – is a very desirable property. The Wikipedia has a good discussion under *Falsifiability* — see http://en.wikipedia.org/wiki/Falsifiable. Statistical models may be viewed as primitive scientific models, and should be subject to the same scrutiny. It would be nice if scientists who use statistical methods would take a cold, clear look at the statistical models they are using, and ask "Is this a reasonable model for my data?"

Maximum likelihood for the parameters in the covariance matrix would work up to a point, but the lack of unique values for μ_x, ν_1, ν_2 and β_0 would cause numerical problems. A good solution is to *re-parameterize* the model, absorbing $\mu_x+\nu_1$ into a parameter called μ_1, $\mu_x+\nu_2$ into a parameter called μ_2, and $\beta_0 + \beta_1\mu_x$ into a parameter called μ_3. The parameters in $\boldsymbol{\mu} = (\mu_1, \mu_2, \mu_3)^\top$ lack meaning and interest[24], but we can estimate them with the vector of sample means $\overline{\mathbf{D}}$ and focus on the parameters in the covariance matrix.

Here is the multivariate normal likelihood from Appendix **??**, simplified so that it's clear that the likelihood depends on the data only through the MLEs $\overline{\mathbf{D}}$ and $\widehat{\boldsymbol{\Sigma}}$. This is just a reproduction of expression (**??**).

$$L(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = |\boldsymbol{\Sigma}|^{-n/2}(2\pi)^{-np/2} \exp -\frac{n}{2}\left\{tr(\widehat{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}^{-1}) + (\overline{\mathbf{D}}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\overline{\mathbf{D}}-\boldsymbol{\mu})\right\}$$

Notice that if $\boldsymbol{\Sigma}$ is positive definite then so is $\boldsymbol{\Sigma}^{-1}$, and so for *any* positive definite $\boldsymbol{\Sigma}$ the likelihood is maximized when $\boldsymbol{\mu} = \overline{\mathbf{D}}$. In that case, the last term just disappears. So, re-parameterizing and then letting $\widehat{\boldsymbol{\mu}} = \overline{\mathbf{D}}$ leaves us free to conduct inference on the model parameters in $\boldsymbol{\Sigma}$.

Just to clarify, after re-parameterization and estimation of $\boldsymbol{\mu}$ with $\overline{\mathbf{D}}_n$, the likelihood function may be written

$$L(\boldsymbol{\theta}) = |\boldsymbol{\Sigma}(\boldsymbol{\theta})|^{-n/2}(2\pi)^{-np/2} \exp -\frac{n}{2}\left\{tr(\widehat{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1})\right\}, \tag{36}$$

where $\boldsymbol{\theta}$ is now a vector of just those parameters appearing in the covariance matrix. This formulation is general. For the specific case of the double measurement Model (34), $\boldsymbol{\theta} = (\phi, \omega_1, \omega_2, \beta_1, \psi)^\top$, and $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ is given by Expression (35). Maximum likelihood estimation is numerical, and the full range of large-sample likelihood methods described in Section **??** of Appendix **??** is available.

The Double Measurement Design in Matrix Form

Now consider the general case of regression with measurement error in both the explanatory variables and the response variables, beginning with a model in which all random variables have expected value zero and there are no intercepts. One can think of this as writing the model in centered form; imagine the letter c over all the random vectors that are not error terms. Centering the model makes it easier to calculate variances and covariances, and *imagining* the letter c over the random vectors saves a lot of typesetting.

Independently for $i = 1, \ldots, n$, let

$$\begin{aligned} \mathbf{W}_{i,1} &= \mathbf{X}_i + \mathbf{e}_{i,1} \\ \mathbf{V}_{i,1} &= \mathbf{Y}_i + \mathbf{e}_{i,2} \\ \mathbf{W}_{i,2} &= \mathbf{X}_i + \mathbf{e}_{i,3}, \\ \mathbf{V}_{i,2} &= \mathbf{Y}_i + \mathbf{e}_{i,4}, \\ \mathbf{Y}_i &= \boldsymbol{\beta}\mathbf{X}_i + \boldsymbol{\epsilon}_i \end{aligned} \tag{37}$$

where

> $\mathbf{Y}_i$ is a $q \times 1$ random vector of latent response variables. Because q can be greater than one, the regression is multivariate.
>
> $\boldsymbol{\beta}$ is an $q \times p$ matrix of unknown constants. These are the regression coefficients, with one row for each response variable and one column for each explanatory variable.
>
> $\mathbf{X}_i$ is a $p \times 1$ random vector of latent explanatory variables, with expected value zero and variance-covariance matrix $\boldsymbol{\Phi}$, a $p \times p$ symmetric and positive definite matrix of unknown

[24] If X_i is true amount of exercise, μ_x is the average amount of exercise in the population; it's very meaningful. Also, the quantity ν_1 is interesting; it's the average amount people exaggerate how much they exercise using Questionnaire One. But when you add these two interesting quantities together, you get garbage. The parameter $\boldsymbol{\mu}$ in the re-parametrerized model is a garbage can.

constants.

$\boldsymbol{\epsilon}_i$ is the error term of the latent regression. It is a $q \times 1$ random vector with expected value zero and variance-covariance matrix $\boldsymbol{\Psi}$, a $q \times q$ symmetric and positive definite matrix of unknown constants.

$\mathbf{W}_{i,1}$ and $\mathbf{W}_{i,2}$ are $p \times 1$ observable random vectors, each representing $\mathbf{X}_i$ plus random error.

$\mathbf{V}_{i,1}$ and $\mathbf{V}_{i,2}$ are $q \times 1$ observable random vectors, each representing $\mathbf{Y}_i$ plus random error.

$\mathbf{e}_{i,1}, \ldots, \mathbf{e}_{i,1}$ are the measurement errors in $\mathbf{W}_{i,1}, \mathbf{V}_{i,1}, \mathbf{W}_{i,2}$ and $\mathbf{V}_{i,2}$ respectively. Joining the vectors of measurement errors into a single long vector $\mathbf{e}_i$, its covariance matrix may be written as a partitioned matrix

$$cov(\mathbf{e}_i) = V \begin{pmatrix} \mathbf{e}_{i,1} \\ \mathbf{e}_{i,2} \\ \mathbf{e}_{i,3} \\ \mathbf{e}_{i,4} \end{pmatrix} = \left(\begin{array}{c|c|c|c} \boldsymbol{\Omega}_{11} & \boldsymbol{\Omega}_{12} & \mathbf{0} & \mathbf{0} \\ \hline \boldsymbol{\Omega}_{12}^\top & \boldsymbol{\Omega}_{22} & \mathbf{0} & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{0} & \boldsymbol{\Omega}_{33} & \boldsymbol{\Omega}_{34} \\ \hline \mathbf{0} & \mathbf{0} & \boldsymbol{\Omega}_{34}^\top & \boldsymbol{\Omega}_{44} \end{array}\right) = \boldsymbol{\Omega}.$$

In addition, the matrices of covariances between $\mathbf{X}_i, \boldsymbol{\epsilon}_i$ and $\mathbf{e}_i$ are all zero.

The main idea of the Double Measurement Design is that every variable is measured by two different methods. Errors of measurement may be correlated within measurement methods[25], but not between methods. So for example, farmers who overestimate their number of pigs may also overestimate their number of cows. On the other hand, if the number of pigs is counted once by the farm manager at feeding time and on another occasion by a research assistant from an areal photograph, then it would be fair to assume that the errors of measurement for the different methods are uncorrelated.

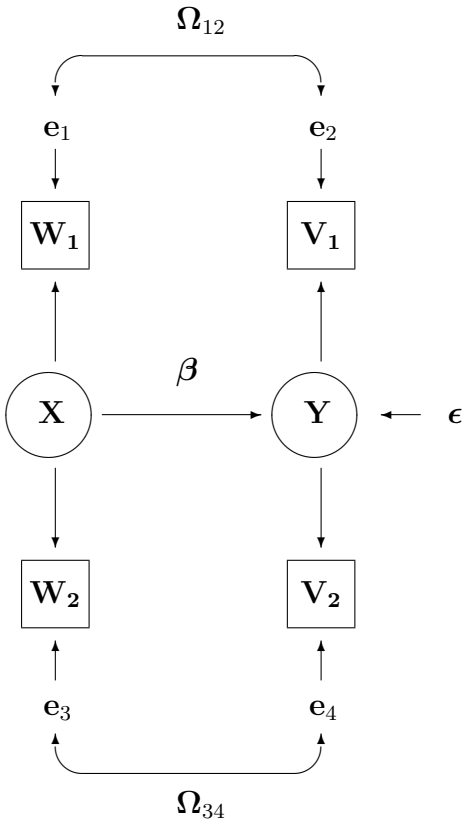
In symbolic terms, $\mathbf{e}_{i,1}$ is error in measuring the explanatory variables by method one, and $\mathbf{e}_{i,2}$ is error in measuring the response variables by method one. $cov(\mathbf{e}_{i,1}) = \boldsymbol{\Omega}_{11}$ need not be diagonal, so method one's errors of measurement for the explanatory variables may be correlated with one another. Similarly, $cov(\mathbf{e}_{i,2}) = \boldsymbol{\Omega}_{22}$ need not be diagonal, so method one's errors of measurement for the response variables may be correlated with one another. And, errors of measurement using the same method may be correlated between the explanatory and response variables. For method one, this is represented by the matrix $cov(\mathbf{e}_{i,1}, \mathbf{e}_{i,2}) = \boldsymbol{\Omega}_{12}$. The same pattern holds for method two. On the other hand, $\mathbf{e}_{i,1}$ and $\mathbf{e}_{i,2}$ are each uncorrelated with both $\mathbf{e}_{i,3}$ and $\mathbf{e}_{i,4}$.

To emphasize an important practical point, the matrices $\boldsymbol{\Omega}_{11}$ and $\boldsymbol{\Omega}_{33}$ must be of the same dimension, just as $\boldsymbol{\Omega}_{22}$ and $\boldsymbol{\Omega}_{44}$ must be of the same dimension – but none of the corresponding elements have to be equal. In particular, the corresponding diagonal elements may be unequal. This means that measurements of a variable by two different methods do not need to be equally precise.

The model is depicted in Figure 9. It follows the usual conventions for path diagrams of structural equation models. Straight arrows go from *exogenous* variables (that is, explanatory variables, those on the right-hand side of equations) to *endogenous* varables (response variables, those on the left side). Correlations among exogenous variables are represented by two-headed curved arrows. Observable variables are enclosed by rectangles or squares, while latent variables are enclosed by ellipses or circles. Error terms are not enclosed by anything.

[25]This is almost unavoidable anyway. The ability of the double measurement model to admit the existence of correlated measurement error and still be identifiable is a great virtue.

Figure 9: The Double Measurement Model



Proof of parameter identifiability The following is typical of easier proofs for structural equation models. The goal is to solve for the model parameters in terms of elements of the variance-covariance matrix of the observable data. This shows the parameters are functions of the distribution, so that no two distinct parameter values could yield the same distribution of the observed data.

Collecting $\mathbf{W}_{i,1}$, $\mathbf{V}_{i,1}$, $\mathbf{W}_{i,2}$ and $\mathbf{V}_{i,2}$ into a single long data vector $\mathbf{D}_i$, we write its variance-covariance matrix as a partitioned matrix:

$$\boldsymbol{\Sigma} = \left(\begin{array}{c|c|c|c} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} & \boldsymbol{\Sigma}_{13} & \boldsymbol{\Sigma}_{14} \\ \hline & \boldsymbol{\Sigma}_{22} & \boldsymbol{\Sigma}_{23} & \boldsymbol{\Sigma}_{24} \\ \hline & & \boldsymbol{\Sigma}_{33} & \boldsymbol{\Sigma}_{34} \\ \hline & & & \boldsymbol{\Sigma}_{44} \end{array}\right),$$

where the covariance matrix of $\mathbf{W}_{i,1}$ is $\boldsymbol{\Sigma}_{11}$, the covariance matrix of $\mathbf{V}_{i,1}$ is $\boldsymbol{\Sigma}_{22}$, the matrix of covariances between $\mathbf{W}_{i,1}$ and $\mathbf{V}_{i,1}$ is $\boldsymbol{\Sigma}_{12}$, and so on.

Now we express all the $\boldsymbol{\Sigma}_{ij}$ sub-matrices in terms of the parameter matrices of Model (37) by straightforward variance-covariance calculations. Students may be reminded that things go smoothly if one substitutes for everything in terms of explanatory variables and error terms before

actually starting to calculate covariances. For example,

$$
\begin{aligned}
\boldsymbol{\Sigma}_{12} &= cov(\mathbf{W}_{i,1}, \mathbf{V}_{i,1}) \\
&= E(\mathbf{W}_{i,1}\mathbf{V}_{i,1}^\top) \\
&= E\left((\mathbf{X}_i + \mathbf{e}_{i,1})(\mathbf{Y}_i + \mathbf{e}_{i,2})^\top\right) \\
&= E\left((\mathbf{X}_i + \mathbf{e}_{i,1})(\boldsymbol{\beta}\mathbf{X}_i + \boldsymbol{\epsilon}_i + \mathbf{e}_{i,2})^\top\right) \\
&= E\left((\mathbf{X}_i + \mathbf{e}_{i,1})(\mathbf{X}_i^\top\boldsymbol{\beta}^\top + \boldsymbol{\epsilon}_i^\top + \mathbf{e}_{i,2}^\top)\right) \\
&= E(\mathbf{X}_i\mathbf{X}_i^\top\boldsymbol{\beta}^\top + \mathbf{X}_i\boldsymbol{\epsilon}_i^\top + \mathbf{X}_i\mathbf{e}_{i,2}^\top + \mathbf{e}_{i,1}\mathbf{X}_i^\top\boldsymbol{\beta}^\top + \mathbf{e}_{i,1}\boldsymbol{\epsilon}_i^\top + \mathbf{e}_{i,1}\mathbf{e}_{i,2}^\top) \\
&= E(\mathbf{X}_i\mathbf{X}_i^\top)\boldsymbol{\beta}^\top + E(\mathbf{X}_i)E(\boldsymbol{\epsilon}_i^\top) + E(\mathbf{X}_i)E(\mathbf{e}_{i,2}^\top) + E(\mathbf{e}_{i,1})E(\mathbf{X}_i^\top)\boldsymbol{\beta}^\top + E(\mathbf{e}_{i,1})E(\boldsymbol{\epsilon}_i^\top) + E(\mathbf{e}_{i,1}\mathbf{e}_{i,2}^\top) \\
&= \boldsymbol{\Phi}\boldsymbol{\beta}^\top + 0 + 0 + 0 + 0 + \boldsymbol{\Omega}_{12}.
\end{aligned}
$$

In this manner, we obtain the partitioned covariance matrix of the observable data $\mathbf{D}_i = (\mathbf{W}_{i,1}^\top, \mathbf{V}_{i,1}^\top, \mathbf{W}_{i,2}^\top, \mathbf{V}_{i,}^\top$ as

$$
\begin{aligned}
\boldsymbol{\Sigma} &= \left(\begin{array}{c|c|c|c}
\boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} & \boldsymbol{\Sigma}_{13} & \boldsymbol{\Sigma}_{14} \\ \hline
 & \boldsymbol{\Sigma}_{22} & \boldsymbol{\Sigma}_{23} & \boldsymbol{\Sigma}_{24} \\ \hline
 & & \boldsymbol{\Sigma}_{33} & \boldsymbol{\Sigma}_{34} \\ \hline
 & & & \boldsymbol{\Sigma}_{44}
\end{array}\right) \qquad (38) \\
&= \left(\begin{array}{c|c|c|c}
\boldsymbol{\Phi} + \boldsymbol{\Omega}_{11} & \boldsymbol{\Phi}\boldsymbol{\beta}^\top + \boldsymbol{\Omega}_{12} & \boldsymbol{\Phi} & \boldsymbol{\Phi}\boldsymbol{\beta}^\top \\ \hline
 & \boldsymbol{\beta}\boldsymbol{\Phi}\boldsymbol{\beta}^\top + \boldsymbol{\Psi} + \boldsymbol{\Omega}_{22} & \boldsymbol{\beta}\boldsymbol{\Phi} & \boldsymbol{\beta}\boldsymbol{\Phi}\boldsymbol{\beta}^\top + \boldsymbol{\Psi} \\ \hline
 & & \boldsymbol{\Phi} + \boldsymbol{\Omega}_{33} & \boldsymbol{\Phi}\boldsymbol{\beta}^\top + \boldsymbol{\Omega}_{34} \\ \hline
 & & & \boldsymbol{\beta}\boldsymbol{\Phi}\boldsymbol{\beta}^\top + \boldsymbol{\Psi} + \boldsymbol{\Omega}_{44}
\end{array}\right)
\end{aligned}
$$

The equality (38) corresponds to a system of ten matrix equations in nine matrix unknowns. The unknowns are the parameter matrices of Model (37): $\boldsymbol{\Phi}$, $\boldsymbol{\beta}$, $\boldsymbol{\Psi}$, $\boldsymbol{\Omega}_{11}$, $\boldsymbol{\Omega}_{22}$, $\boldsymbol{\Omega}_{33}$, $\boldsymbol{\Omega}_{44}$, $\boldsymbol{\Omega}_{12}$, and $\boldsymbol{\Omega}_{34}$. In the solution below, notice that once a parameter has been identified, it may be used to solve for other parameters without explicitly substituting in terms of $\boldsymbol{\Sigma}_{ij}$ quantities. Sometimes a full explicit solution is useful, but to show identifiability all you need to do is show that the moment structure equations *can* be solved.

$$
\begin{aligned}
\boldsymbol{\Phi} &= \boldsymbol{\Sigma}_{13} \qquad (39) \\
\boldsymbol{\beta} &= \boldsymbol{\Sigma}_{23}\boldsymbol{\Phi}^{-1} = \boldsymbol{\Sigma}_{14}^\top\boldsymbol{\Phi}^{-1} \\
\boldsymbol{\Psi} &= \boldsymbol{\Sigma}_{24} - \boldsymbol{\beta}\boldsymbol{\Phi}\boldsymbol{\beta}^\top \\
\boldsymbol{\Omega}_{11} &= \boldsymbol{\Sigma}_{11} - \boldsymbol{\Phi} \\
\boldsymbol{\Omega}_{22} &= \boldsymbol{\Sigma}_{22} - \boldsymbol{\beta}\boldsymbol{\Phi}\boldsymbol{\beta}^\top - \boldsymbol{\Psi} \\
\boldsymbol{\Omega}_{33} &= \boldsymbol{\Sigma}_{33} - \boldsymbol{\Phi} \\
\boldsymbol{\Omega}_{44} &= \boldsymbol{\Sigma}_{44} - \boldsymbol{\beta}\boldsymbol{\Phi}\boldsymbol{\beta}^\top - \boldsymbol{\Psi} \\
\boldsymbol{\Omega}_{12} &= \boldsymbol{\Sigma}_{12} - \boldsymbol{\Phi}\boldsymbol{\beta}^\top \\
\boldsymbol{\Omega}_{34} &= \boldsymbol{\Sigma}_{34} - \boldsymbol{\Phi}\boldsymbol{\beta}^\top
\end{aligned}
$$

This shows that the parameters of Model (37) are identifiable, so that if data are collected following the double measurement recipe, then the data analysis may proceed with no worries about parameter identifiability.

Notice in the covariance structure equations (38), that $\boldsymbol{\Sigma}_{14} = \boldsymbol{\Sigma}_{23}^\top$. As in the scalar example of Section 11 (see page 36), this constraint on the covariance matrix $\boldsymbol{\Sigma}$ arises from the model, and provides a way to test whether the model is correct. These pq equalities are not the only ones implied by the model. Because $\boldsymbol{\Sigma}_{13} = \boldsymbol{\Phi}$, the $p \times p$ matrix of covariances $\boldsymbol{\Sigma}_{13}$ is actually a covariance matrix, so it is symmetric. This implies $p(p-1)/2$ more equalities.

Intercepts

Now Model (37) is expanded to include intercepts and non-zero expected values. We will see that this leads to complications that are seldom worth the trouble, and the classical centered models with zero expected value and no intercepts are usually preferable. Let

$$\begin{aligned}
\mathbf{W}_{i,1} &= \boldsymbol{\nu_1} + \mathbf{X}_i + \mathbf{e}_{i,1} \\
\mathbf{V}_{i,1} &= \boldsymbol{\nu_2} + \mathbf{Y}_i + \mathbf{e}_{i,2} \\
\mathbf{W}_{i,2} &= \boldsymbol{\nu_3} + \mathbf{X}_i + \mathbf{e}_{i,3} \\
\mathbf{V}_{i,2} &= \boldsymbol{\nu_4} + \mathbf{Y}_i + \mathbf{e}_{i,4}, \\
\mathbf{Y}_i &= \boldsymbol{\alpha} + \boldsymbol{\beta}\mathbf{X}_i + \boldsymbol{\epsilon}_i
\end{aligned}$$

where $\boldsymbol{\alpha}$, $\boldsymbol{\nu}_1$, $\boldsymbol{\nu}_2$, $\boldsymbol{\nu}_3$ and $\boldsymbol{\nu}_4$ are vectors of constants, and $E(\mathbf{X}_i) = \boldsymbol{\mu}_x$. Everything else is as in Model (37).The terms $\boldsymbol{\nu}_1 \ldots, \boldsymbol{\nu}_4$ are called *measurement bias*. For example, of one of the elements of $\mathbf{W}_{i,1}$ is reported amount of exercise, the corresponding element of $\boldsymbol{\nu}_1$ would be the average amount by which people exaggerate how much they exercise.

Again, the observable data $\mathbf{W}_{i,1}$, $\mathbf{V}_{i,1}$, $\mathbf{W}_{i,2}$ and $\mathbf{V}_{i,2}$ are collected into a data vector $\mathbf{D}_i$, with expected value $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. The pair $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is a function of the probability distribution of $\mathbf{D}_i$. If the parameter matrices of Model (40) are functions of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, then they are also functions of the distribution of $\mathbf{D}_i$, and thus they are identifiable.

Since the addition of constants has no effect on variances or covariances, the contents of $\boldsymbol{\Sigma}$ are given by (38), as before. The expected value $\boldsymbol{\mu}$ is the partitioned vector

$$\boldsymbol{\mu} = \left(\begin{array}{c} \boldsymbol{\mu}_1 \\ \hline \boldsymbol{\mu}_2 \\ \hline \boldsymbol{\mu}_3 \\ \hline \boldsymbol{\mu}_4 \end{array}\right) = \left(\begin{array}{c} E(\mathbf{W}_{i,1}) \\ \hline E(\mathbf{V}_{i,1}) \\ \hline E(\mathbf{W}_{i,2}) \\ \hline E(\mathbf{V}_{i,2}) \end{array}\right) = \left(\begin{array}{c} \boldsymbol{\nu}_1 + \boldsymbol{\mu}_x \\ \hline \boldsymbol{\nu}_2 + \boldsymbol{\alpha} + \boldsymbol{\beta}\boldsymbol{\mu}_x \\ \hline \boldsymbol{\nu}_3 + \boldsymbol{\mu}_x \\ \hline \boldsymbol{\nu}_4 + \boldsymbol{\alpha} + \boldsymbol{\beta}\boldsymbol{\mu}_x \end{array}\right). \tag{40}$$

To demonstrate the identification of Model (40), one would need to solve the equations in (40) uniquely for $\boldsymbol{\nu}_1$, $\boldsymbol{\nu}_2$, $\boldsymbol{\mu}_x$ and $\boldsymbol{\alpha}$. Even with $\boldsymbol{\beta}$ considered known and fixed because it is identified in (39), this is impossible in most of the parameter space, because (40) specifies $2m+2p$ additional equations in $3m + 3p$ additional unknowns.

It is tempting to assume the measuremant bias terms $\boldsymbol{\nu}_1 \ldots, \boldsymbol{\nu}_4$ to be zero; this would allow identification of $\boldsymbol{\alpha}$ and $\boldsymbol{\mu}_x$. Unfortunately, it is doubtful that such an assumption could be justified very often in practice. Most of the time, all we can do is identify the parameter matrices that appear in the covariance matrix, and also the *functions* $\boldsymbol{\mu_1} \ldots, \boldsymbol{\mu_4}$ of the parameters as given in equation (40). This can be viewed as a re-parameterization of the model.

Estimation and testing

Normal model As in the scalar example of Section 11, the (collapsed) expected values are estimated by the corresponding vector of sample means, and then set aside. With multivariate normal distributions for all the random vectors in the model, the resulting likelihood is again (36) on page 38. The full range of large-sample likelihood methods is then available. Maximum likelihood estimates are asymptotically normal, and asymptotic standard errors are convenient by-products of the numerical minimization as described in Section **??** of Appendix **??**; most software produces them by default. Dividing an estimated regression coefficient by its standard error gives a Z-test for whether the coefficient is different from zero. My experience is that likelihood ratio tests can substantially outperform both these Z-tests and the Wald tests that are their generalizations, especially when there is a lot of measurement error, the explanatory variables are strongly related to one another, and the sample size is not huge.

Distribution-free In presenting models for regression with measurement error, it is often convenient to assume that everything is multivariate normal. This is especially true when giving examples of models where the parameters are *not* identifiable. But normality is not necessary. Suppose Model (37) holds, and that the distributions of of the latent explanatory variables and error terms are unknown, except that they possess covariance matrices, with $\mathbf{e}_{i,1}$ and $\mathbf{e}_{i,2}$ having zero covariance with $\mathbf{e}_{i,3}$ and $\mathbf{e}_{i,4}$. In this case the parameter of the model could be expressed as $\theta = (\boldsymbol{\beta}, \boldsymbol{\Phi}, \boldsymbol{\Psi}, \boldsymbol{\Omega}, F_{\mathbf{X}}, F_{\boldsymbol{\epsilon}}, F_{\mathbf{e}})$, where $F_{\mathbf{X}}$, $F_{\boldsymbol{\epsilon}}$ and $F_{\mathbf{e}}$ are the (joint) cumulative distribution functions of $\mathbf{X}_i$, $\boldsymbol{\epsilon}_i$ and $\mathbf{e}_i$ respectively.

Note that the parameter in this "non-parametric" problem is of infinite dimension, but that presents no conceptual difficulty. The probability distribution of the observed data is still a function of the parameter vector, and to show identifiability, we would have to be able to recover the parameter vector from the probability distribution of the data. While in general we cannot recover the whole thing, we certainly can recover a useful *function* of the parameter vector, namely $\boldsymbol{\beta}$. In fact, $\boldsymbol{\beta}$ is the only quantity of interest; the remainder of the parameter vector consists only of nuisance parameters, whether it is of finite dimension or not.

To make the reasoning explicit, the covariance matrix $\boldsymbol{\Sigma}$ is a function of the probability distribution of the observed data, whether that probability distribution is normal or not. The calculations leading to (39) still hold, showing that $\boldsymbol{\beta}$ is a function of $\boldsymbol{\Sigma}$, and hence of the probability distribution of the data. Therefore, $\boldsymbol{\beta}$ is identifiable.

This is all very well, but can we actually *do* anything without knowing what the distributions are? Certainly! Looking at (39), one is tempted to just put hats on everything to obtain Method-of-Moments estimators. However, we can do a little better. Note that while $\boldsymbol{\Phi} = \boldsymbol{\Sigma}_{12}$ is a symmetric matrix in the population and $\widehat{\boldsymbol{\Sigma}}_{12}$ *converges* to a symmetric matrix, $\widehat{\boldsymbol{\Sigma}}_{12}$ will be non-symmetric for any finite sample size (with probability one if the distributions involved are continuous). A better estimator is obtained by averaging pairs of off-diagonal elements:

$$\widehat{\boldsymbol{\Phi}}_M = \frac{1}{2}(\widehat{\boldsymbol{\Sigma}}_{13} + \widehat{\boldsymbol{\Sigma}}_{13}^\top),$$

where the subscript M indicates a Method-of-Moments estimator. Using the second line of (39), a reasonable though non-standard estimator of $\boldsymbol{\beta}$ is

$$\widehat{\boldsymbol{\beta}}_M = \frac{1}{2}\left(\widehat{\boldsymbol{\Sigma}}_{14}^\top + \widehat{\boldsymbol{\Sigma}}_{23}\right)\widehat{\boldsymbol{\Phi}}_M^{-1} \tag{41}$$

Consistency follows from the Law of Large Numbers and a continuity argument. All this assumes the existence only of second moments and cross-moments. With the assumption of fourth moments (so that sample variances possess variances), the multivariate Central Limit Theorem provides a routine[26] basis for large-sample interval estimation and testing.

However, there is no need to bother. Research on the robustness of the normal model for structural equation models (Amemiya, Fuller and Pantula, 1987; Anderson and Rubin, 1956; Anderson and Amemiya, 1988; Anderson, 1989; Anderson and Amemiya, 1990; Browne, 1988; Browne and Shapiro, 1988; Satorra and Bentler, 1990) shows that procedures for (such as likelihood ratio and Wald tests) based on a multivariate normal model are asymptotically valid even when the normal assumption is false. And Satorra and Bentler (1990) describe Monte Carlo work suggesting that normal-theory methods generally perform better than at least one method (Browne, 1984) that is specifically designed to be distribution-free. Since the methods suggested by the estimator (41) are similar to Browne's weighted least squares approach, they are also likely to be inferior to the standard normal-theory tools.

It is important to note that while the normal-theory tests and confidence intervals for $\boldsymbol{\beta}$ can be trusted when the data are not normal, this does not extend to the other model parameters. For example, if the vector of latent variables $\mathbf{X}_i$ is not normal, then normal-theory inference about its covariance matrix will be flawed. In any event, the method of choice is maximum likelihood, with

[26] Okay, I admit there is a fairly long story here.

interpretive focus on the regression coefficients in $\boldsymbol{\beta}$ rather than on the other model parameters.