Chapter 0

Regression with measurement error

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.

0.1 Regression: Conditional or Unconditional?

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 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,

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

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

[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.

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 A.4.4 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 great 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.

0.2 The Centering Rule

In this book, we are focusing on *unconditional* regression models, in which the explanatory variables are random. Mostly, the models 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. This means there will be a lot

of variance and covariance calculations, and anything that makes it easier will be very welcome.

This section presents a theorem that is another way of expressing the *Centering Rule* given on page 91 of Appendix A. 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 do the calculation. The centered version of a random vector will be denoted[2] by $\overset{c}{\mathbf{X}} = \mathbf{X} - E(\mathbf{X})$, so that $E(\overset{c}{\mathbf{X}}) = \mathbf{0}$ and $V(\overset{c}{\mathbf{X}}) = E(\overset{c}{\mathbf{X}}\overset{c}{\mathbf{X}}') = E[(\mathbf{X} - E(\mathbf{X}))(\mathbf{X} - E(\mathbf{X}))'] = V(\mathbf{X})$.

Theorem 1 *Let* $\mathbf{X}_1, \ldots, \mathbf{X}_m$ *and* $\mathbf{Y}_1, \ldots, \mathbf{Y}_k$ *be random vectors, and*

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

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 $V(\mathbf{L}_1) = V(\overset{c}{\mathbf{L}}_1)$, $V(\mathbf{L}_2) = V(\overset{c}{\mathbf{L}}_2)$, *and* $C(\mathbf{L}_1, \mathbf{L}_2) = C(\overset{c}{\mathbf{L}}_1, \overset{c}{\mathbf{L}}_2)$.

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

$$\begin{aligned} V(\mathbf{X} + \mathbf{Y}) &= V(\overset{c}{\mathbf{X}} + \overset{c}{\mathbf{Y}}) \\ &= E(\overset{c}{\mathbf{X}} + \overset{c}{\mathbf{Y}})(\overset{c}{\mathbf{X}} + \overset{c}{\mathbf{Y}})' \\ &= E(\overset{c}{\mathbf{X}} + \overset{c}{\mathbf{Y}})(\overset{c}{\mathbf{X}}' + \overset{c}{\mathbf{Y}}') \\ &= E(\overset{c}{\mathbf{X}}\overset{c}{\mathbf{X}}') + E(\overset{c}{\mathbf{Y}}\overset{c}{\mathbf{Y}}') + E(\overset{c}{\mathbf{X}}\overset{c}{\mathbf{Y}}') + E(\overset{c}{\mathbf{Y}}\overset{c}{\mathbf{X}}') \\ &= V(\mathbf{X}) + V(\mathbf{Y}) + C(\mathbf{X}, \mathbf{Y}) + C(\mathbf{Y}, \mathbf{X}) \end{aligned}$$

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, $C(\mathbf{X}, \mathbf{Y})$ is not in general equal to $C(\mathbf{Y}, \mathbf{X})$, though $C(\mathbf{Y}, \mathbf{X}) = C(\mathbf{X}, \mathbf{Y})'$.

The centering rule is useful in scalar variance-covrince 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,

[2] This notation is very non-standard. Let's see if it helps.

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

This valuable calculation looks worse than it is at first glance, becuse every little step is show. It is significantly messier without centering.

0.3 Unconditional regression without measurement error

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

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

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 and ϵ_i are independent.

Under this model the pairs (X_i, Y_i) are bivariate normal, with

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

and variance-covariance matrix

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

The Centering Rule (Theorem 1 on page 5) is useful for calculating the covariance:

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

Here is some useful terminology:

Definition 0.3.1 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 0.3.2 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 regression Model (4), the moments structure equations are

$$\begin{aligned} \mu_1 &= \mu_x \\ \mu_2 &= \beta_0 + \beta_1\mu_x \\ \sigma_{1,1} &= \phi \\ \sigma_{1,2} &= \beta_1\phi \\ \sigma_{2,2} &= \beta_1^2\phi + \psi. \end{aligned} \tag{5}$$

Here, the moments are the elements of the mean vector $\boldsymbol{\mu}$, and the unique elements of the covariance matrix $\boldsymbol{\Sigma}$. This is a system of 5 equations in five unknowns, and may be

readily be solved to yield

$$\begin{aligned} \beta_0 &= \mu_2 - \frac{\sigma_{1,2}}{\sigma_{1,1}}\mu_1 \\ \mu_x &= \mu_1 \\ \phi &= \sigma_{1,1} \\ \beta_1 &= \frac{\sigma_{1,2}}{\sigma_{1,1}} \\ \psi &= \sigma_{2,2} - \frac{\sigma_{1,2}^2}{\sigma_{1,1}}. \end{aligned} \tag{6}$$

The existence of this nice solution is quite revealing. It tells us that the parameters of the normal regression Model (4) 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 Appendix A) says that the MLE of a function is just that function of the MLE. So, to obtain the maximum likelihood estimators of the regression model from the maximum likelihood estimators of the bivariate normal distribution, one may just put hats on the parameters in Expression 6, as follows:

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

Thus there is no need to re-derive the maximum likelihood likelihood estimators for the regression model.

These calculations 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, the solution provides a way to estimate the parameters. But it turns out that for some models a unique solution for the parameters is mathematically impossible. In such cases, successful parameter estimation by any method is usually 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[3].

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

$$\mathbf{Y}_i = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 \mathbf{X}_i + \boldsymbol{\epsilon}_i \tag{7}$$

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 $\boldsymbol{\theta} = (\boldsymbol{\beta}_0, \boldsymbol{\mu}_x, \boldsymbol{\Phi}, \boldsymbol{\beta}_1, \boldsymbol{\Psi})$, where it is understood that the symbols for the matrices really refer to their unique elements[4].

The observable data are the random vectors $\mathbf{D}_i = (\mathbf{X}_i', \mathbf{Y}_i')'$, for $i = 1, \ldots, n$. Because the data vectors are linear combinations of multivariate normals, they are also multivariate normal. That is, $\mathbf{D}_i \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, We write $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ as partitioned matrices (matrices of matrices).

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

[3] Of course it is no surprise that estimating the parameters of a regression model is technically possible.

[4] In the present case, this informal notation is probably clearer than the *vech* notation defined in Appendix A.

and

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

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

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

Thus, the moment structure equations are

$$\begin{aligned}\boldsymbol{\mu}_1 &= \boldsymbol{\mu}_x \\ \boldsymbol{\mu}_2 &= \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1\boldsymbol{\mu}_x \\ \boldsymbol{\Sigma}_{11} &= \boldsymbol{\Phi} \\ \boldsymbol{\Sigma}_{12} &= \boldsymbol{\Phi}\boldsymbol{\beta}_1' \\ \boldsymbol{\Sigma}_{22} &= \boldsymbol{\beta}_1\boldsymbol{\Phi}\boldsymbol{\beta}_1' + \boldsymbol{\Psi}.\end{aligned} \tag{8}$$

Solving for the parameter matrices is routine.

$$\begin{aligned}\boldsymbol{\beta}_0 &= \boldsymbol{\mu}_2 - \boldsymbol{\Sigma}_{12}'\boldsymbol{\Sigma}_{11}^{-1}\,\boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_x &= \boldsymbol{\mu}_1 \\ \boldsymbol{\Phi} &= \boldsymbol{\Sigma}_{11} \\ \boldsymbol{\beta}_1 &= \boldsymbol{\Sigma}_{12}'\boldsymbol{\Sigma}_{11}^{-1} \\ \boldsymbol{\Psi} &= \boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{12}'\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\end{aligned} \tag{9}$$

As in the univariate case, the invariance principle may be used to obtain the maximum likelihood estimators for the parameters of the regression model. Just put hats on all the parameters in Expression (9).

0.4 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 0.1, 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

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

we are saying that $X_{i,1}$ contributes to Y_i, but there are also other, unspecified influences. All these other influences are rolled together into ϵ_i. It is common practice to assume that $X_{i,1}$ and ϵ_i are independent, or at least uncorrelated, and that is the kind of assumption that will be made throughout this book. In fact, without this assumption everything usually falls apart on a technical level. But that does not mean the assumption can be justified in practice. Prepare yourself for a strong and bitter dose of reality.

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:

$$Y_i = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \beta_2 X_{i,3} + \epsilon_i, \tag{10}$$

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

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

where ϵ_i is independent of $X_{i,1}$, $X_{i,2}$ and $X_{i,3}$.

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

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

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.

$$\begin{aligned} Cov(X_{i,1}, \epsilon_i') &= Cov(X_{i,1}, \beta_2 X_{i,2} + \beta_3 X_{i,3} - \beta_2\mu_2 - \beta_3\mu_3 + \epsilon_i) \\ &= Cov(\overset{c}{X}_{i,1}, \beta_2 \overset{c}{X}_{i,2} + \beta_3 \overset{c}{X}_{i,3} + \epsilon_i) \\ &= \beta_2\phi_{12} + \beta_3\phi_{13} \end{aligned} \tag{11}$$

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[5].

Response variables are almost always affected by more than one explanatory variable, and in non-experimental data[6], 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.

We have arrived at the following model, which will be called the *true model* in the discussion that follows. Independently for $i = 1, \ldots, n$,

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \tag{12}$$

where $E(X_i) = \mu_x$, $Var(X_i) = \sigma_x^2$, $E(\epsilon_i) = 0$, $Var(\epsilon_i) = \sigma_\epsilon^2$, and $Cov(X_i, \epsilon_i) = c$.

Consider a data set consisting 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. 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 is independent of the variables that are in the equation.* Of course this is almost never true, and now we will see the consequences.

Both ordinary least squares and an unconditional regression model like the true model

[5]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 for.

[6]Values of variables are just observed, and not experimentally manipulated.

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 A.5 in Appendix A), $\widehat{\sigma}_{x,y}$ 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 7c and 7a) in Section A.5),

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

Thus, while the usual teaching is that sample regression coefficients are unbiased estimators, we see here that $\widehat{\beta}_1$ is biased, even 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 *misspecified*. 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 nonzero covariance with variables that are in the model, the regression coefficients are biased and inconsistent.* In other words, they give the wrong answer, and do not approach the right answer even for very large samples.

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 bias described here.*

Actually, the calculations in this section are the technical counterpart of a familiar point, the *correlation-causation* issue, which is often stated more or less as follows. If A and B are correlated, 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 the action of potential confounding variables 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.

Trying to fit the true model We have seen that terrible trouble arises from adopting a mis-specified model with $c = 0$, when in fact because of omitted variables, $c \neq 0$. It is natural, therefore, to attempt estimation and inference for the case where $c \neq 0$. For simplicity, assume that the observable variables are normally distributed. Then 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 (12). The result is

$$\boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} = E \begin{bmatrix} X_i \\ Y_i \end{bmatrix} = \begin{bmatrix} \mu_x \\ \beta_0 + \beta_1 \mu_x \end{bmatrix} \tag{13}$$

and

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{bmatrix} = V \begin{bmatrix} X_i \\ Y_i \end{bmatrix} = \begin{bmatrix} \sigma_x^2 & \beta_1 \sigma_x^2 + c \\ \beta_1 \sigma_x^2 + c & \beta_1 \sigma_x^2 + \sigma_\epsilon^2 \end{bmatrix}. \tag{14}$$

This shows the way in which the parameter vector $\boldsymbol{\theta} = (\mu_x, \sigma_x^2, \beta_0, \beta_1, \sigma_\epsilon^2, c)$ determine $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, and hence the probability distribution of the data. Now we will see that every such probability distribution can arise from infinitely many sets of parameter values.

For any given pair $\boldsymbol{\mu} = (\mu_1, \mu_2) \in \mathbb{R}^2$ and any 2×2 symmetric, positive definite matrix $\boldsymbol{\Sigma} = [\sigma_{i,j}]$, let the values of the other parameters depend on the value of β_1, as follows.

$$\begin{aligned} \beta_0 &= \mu_2 - \beta_1 \mu_1 \\ c &= \sigma_{12} - \beta_1 \sigma_{11} \\ \sigma_\epsilon^2 &= \sigma_{22} - \beta_1^2 \sigma_{11} \\ \mu_x &= \mu_1 \\ \sigma_x^2 &= \sigma_{11} \end{aligned} \tag{15}$$

This defines a 5-dimensional surface in the 6-dimensional parameter space. The value of

β_1 is not completely arbirtary. Because variances are greater than zero,

$$\begin{aligned} & \sigma_\epsilon^2 = \sigma_{22} - \beta_1^2\sigma_{11} > 0 \\ \Leftrightarrow \quad & \beta_1^2\sigma_{11} < \sigma_{22} \\ \Leftrightarrow \quad & \beta_1^2 < \frac{\sigma_{22}}{\sigma_{11}} \\ \Leftrightarrow \quad & |\beta_1| < \sqrt{\frac{\sigma_{22}}{\sigma_{11}}}. \end{aligned}$$

So let the parameter β_1 (which determines the relationship between X and Y, if any) vary between plus and minus $\sqrt{\frac{\sigma_{22}}{\sigma_{11}}}$. For each β_1 in this range, substituting (15) into (13) and (14) returns

$$\boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} \quad \text{and} \quad \boldsymbol{\Sigma} = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{bmatrix}.$$

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 the values of β_1 between plus and minus $\sqrt{\frac{\sigma_{22}}{\sigma_{11}}}$ yield the same population mean and population variance-covariance matrix for the parameter sets defined by the surface (15)[7].

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.

In particular, maximum likelihood estimation of the parameter vecctor $\boldsymbol{\theta}$ will fail. 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, the search will take you to a flat place defined by the surface (15). There, you will find the infinitely many parameter values corresponding to the generic MLE $(\overline{\mathbf{x}}, \widehat{\boldsymbol{\Sigma}})$.

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 $c = 0$ leads 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.

Is there any way out of this mess? What should be conclued from this discussion? Is regression completely useless when applied to non-experimental data, unless every

[7]Could you estimate β_1 if it were big enough? What would happen if $\beta_1^2 > \frac{\sigma_{22}}{\sigma_{11}}$? This would imply $\sigma_\epsilon^2 < 0$, which is impossible if the true model is correct.

conceivable explanatory variable is included in the model? The answer is no, it's not quite useless. But one must talk about the results very carefully.

First of all, 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.

When it comes to *interpreting* results, it is possible to say something useful as long as you are cautious. Suppose you have a regression with three explanatory variables. It is an observational study with plenty of potential omitted variables, so assuming the error terms are uncorrelated with the explanatory variables is not really supportable. Using the usual methods, you reject $H_0 : \beta_3 = 0$, with $\widehat{\beta}_3 > 0$. You can say something like the following: "Controlling for age and sex, intake of Vitamin D supplements is positively related to bone density among older adults. That is, older adults who take more Vitamin D tend to have denser bones, even when you allow for age and sex."

This is okay so far, but if the investigator assumes that higher bone density is actually *produced* by Vitamin D supplements (more or less as Y is produced by X_3 in the regression equation), then that is not justified by the statistical analysis. It could be that some other variable like amount of exercise is causing increased bone density, and is also associated with intake of Vitamin D supplements.

All this discussion may have the effect of obscuring a point that should not be hidden.

> In this book and elsewhere, we frequently consider models in which the covariance of the error term and the explanatory variables equals zero. If the model represents a belief that the explanatory variables *contribute* to the response variable, the assumption of zero covariance quietly makes a very big claim. The claim is that *if there are any unmeasured explanatory variables that also contribute to the response variable, they are unrelated to the explanatory variables that are in the model*. Such an assumption is impossible to believe most of the time, and it is not harmless. When it is wrong, the result can be biased parameter estimates, and tests that support false conclusions with high probability.

0.5 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.

0.5.1 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. Mis-classification 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,

$$W = X + e \tag{16}$$

Let's say $E(X) = \mu$, $E(e) = 0$, $Var(X) = \sigma^2_X$, $Var(e) = \sigma^2_e$, and $Cov(X, e) = 0$. Because X and e are uncorrelated,

$$Var(W) = Var(X) + Var(e) = \sigma^2_X + \sigma^2_e.$$

So, 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[8], the *reliability*[9] 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

$$\begin{aligned} \rho &= \left(\frac{Cov(X,W)}{SD(X)SD(W)}\right)^2 \\ &= \left(\frac{\sigma_X^2}{\sqrt{\sigma_X^2}\sqrt{\sigma_X^2+\sigma_e^2}}\right)^2 \\ &= \frac{\sigma_X^4}{\sigma_X^2(\sigma_X^2+\sigma_e^2)} \\ &= \frac{\sigma_X^2}{\sigma_X^2+\sigma_e^2}. \end{aligned} \tag{17}$$

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.

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 6 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. This means another approach is needed.

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

[8]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* [5]. 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.

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

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 4 at the end of this section.

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 $E(X) = \mu$, $Var(X) = \sigma^2_X$, $E(e_1) = E(e_2) = 0$, $Var(e_1) = Var(e_2) = \sigma^2_e$, 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.

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 throughout this course) is greatly simplified by using the *Centering Rule* on page 91 of Appendix A. Basically, the centering rule says it is safe to assume that all expected values are zero, even though they may not be. The answer will be the same.

So, assuming without loss of generality that $\mu = 0$,

$$\begin{aligned} Corr(W_1, W_2) &= \frac{Cov(W_1, W_2)}{SD(W_1)SD(W_2)} \\ &= \frac{E(W_1 W_2)}{\sqrt{\sigma^2_X + \sigma^2_e}\sqrt{\sigma^2_X + \sigma^2_e}} \\ &= \frac{E(X + e_1)(X + e_2)}{\sigma^2_X + \sigma^2_e} \\ &= \frac{E(X^2) + 0 + 0 + 0}{\sigma^2_X + \sigma^2_e} \\ &= \frac{\sigma^2_X}{\sigma^2_X + \sigma^2_e}, \end{aligned} \tag{18}$$

which is the reliability. Notice the final crucial step, in which σ^2_X is substituted for $E(X^2)$.

The calculation above is the basis of *test-retest reliability*[10], 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

[10]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.

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.

Correlated measurement error Notice that if participants remembered their wrong answers or lucky guesses from the first time they took an educational test and just gave the same answer the second time, the result would be a positive correlation between the measurement errors e_1 and e_2. This would mess everything up. Throughout this course we will return again and again 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.

The 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 $Var(X_i) = \sigma^2_X$. 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 $E(e_{i,1}) = E(e_{i,2}) = 0$, $Var(e_{i,1}) = Var(e_{i,2}) = \sigma^2_e$, 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

$$\begin{aligned} R_n &= \frac{\sum_{i=1}^n (W_{i,1} - \overline{W}_1)(W_{i,2} - \overline{W}_2)}{\sqrt{\sum_{i=1}^n (W_{i,1} - \overline{W}_1)^2}\sqrt{\sum_{i=1}^n (W_{i,2} - \overline{W}_2)^2}} \\ &= \frac{\sum_{i=1}^n W_{i,1}W_{i,2} \; - \; n\overline{W}_1\overline{W}_2}{\sqrt{\sum_{i=1}^n W^2_{i,1} - n\overline{W}^2_1}\sqrt{\sum_{i=1}^n W^2_{i,2} - n\overline{W}^2_2}} \\ &= \frac{(\frac{1}{n}\sum_{i=1}^n W_{i,1}W_{i,2}) \; - \; \overline{W}_1\overline{W}_2}{\sqrt{(\frac{1}{n}\sum_{i=1}^n W^2_{i,1}) - \overline{W}^2_1}\sqrt{(\frac{1}{n}\sum_{i=1}^n W^2_{i,2}) - \overline{W}^2_2}}, \end{aligned} \tag{19}$$

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 A.5), we

have the following:

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

Now, since R_n is a continuous function of the various sample moments in (19) and almost sure convergence can be treated like an ordinary limit,

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

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. Notice that this was a limits problem and not a variance-covariance computation, so there was no assumption of zero expected values – even though the limit calculation also works out for that restricted case.

0.6 Ignoring measurement error

This section will show what happens in multiple regression when measurement error in the explanatory variables is ignored. It turns out that under some conditions, measurement error in the response variable is a less serious problem.

0.6.1 Measurement error in the response variable

Example 0.6.1.1 *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.

Here, 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.

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." [3]

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 in a particular way. 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.

Under the true model of Example 0.6.1.1, we have $Cov(X,Y) = \beta_1\sigma_x^2$ and $Var(X) = \sigma_x^2$. Then,

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

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 disaster; 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

$$\begin{aligned} V_i &= \nu + Y_i + e_i \\ &= \nu + (\beta_0 + \beta_1 X_i + \epsilon_i) + e_i \\ &= (\nu + \beta_0) + \beta_1 X_i + (\epsilon_i + e_i) \\ &= \beta_0' + \beta_1 X_i + \epsilon_i' \end{aligned} \tag{20}$$

What has happened here is a *re-parameterization*, 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. So in quite a few of the examples that follow, it will appear that the response variable is being measured without error, but what it really means is that of course there is measurement error in Y_i, but the measurement error is absorbed into the error term. Similarly, the measurement bias ν is absorbed into the intercept, making the intercept a quantity of convenience more than an interpretable model parameter.

In this book and in standard statistical practice, there are many models in which the response variable appears to be measured without error. But of course error-free measurement is a rarity at best, so these models should be viewed as re-parameterized versions of models that acknowledge the reality of measurement error in the response variable. Two important features of these re-parameterized models are that the intercepts represent measurement bias as well as the intercepts of the original models, and that the measurement error is assumed independent of everything else in the model.

0.6.2 Measurement error in the explanatory variable

Example 0.6.2.1 *Independently for* $i = 1, \ldots, n$,

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

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.*

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 0.6.2.1, 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 $\widehat{\sigma}^2_w \stackrel{a.s.}{\to} Var(W)$[11], so that by the continuous mapping property of ordinary limits[12], $\widehat{\beta}_1 \stackrel{a.s.}{\to} \frac{Cov(W,Y)}{Var(W)}$.

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

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

Consequently,

$$\begin{aligned}
\widehat{\beta}_1 &= \frac{\sum_{i=1}^n (W_i - \overline{W})(Y_i - \overline{Y})}{\sum_{i=1}^n (W_i - \overline{W})^2} \\
&= \frac{\widehat{\sigma}_{w,y}}{\widehat{\sigma}^2_w} \\
&\stackrel{a.s.}{\to} \frac{Cov(W,Y)}{Var(W)} \\
&= \beta_1 \left(\frac{\sigma^2_x}{\sigma^2_x + \sigma^2_e} \right). \qquad (21)
\end{aligned}$$

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 (21) 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 smetimes 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 measuremnt 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.

[11]This is true because sample variances and covariances are strongly consistent estimators of the corresponding population quantities; see Section A.5.2 in Appendix A, problems 9 and 10 .

[12]This is true because 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 descend to the level of convergence in probability, you could observe that almost sure convergence implies convergence in probability, and then use Slutsky Lemma 7a of Appendix A.5.

0.6.3 Two Explanatory Variables

In Example 0.6.2.1, 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 measured with error.

Example 0.6.3.1 *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*

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

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*

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

A very nice 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 0.6.3.1, 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 0.6.2.1, estimating $\widehat{\beta}_2$ 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 $\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ (but with $\mathbf{W}$ instead of $\mathbf{X}$) will yield the desired regression coefficients.

Adopting a notation that will be used throughout the course, 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 $\boldsymbol{\Sigma} = [\sigma_{i,j}] = V(\mathbf{D}_i)$. 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,

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

Calculating the regression coefficients is straightforward.

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

Then with a bit of simplification,

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

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

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

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

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

is positive definite – not a lot to ask.

The convergence of $\widehat{\beta}_2$ in expression 22 applies regardless of what model is correct. To see what happens when the true model of Example 0.6.3.1 holds, we calculate the $\boldsymbol{\Sigma}$, the common variance-covariance matrix of the observable data vectors.

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

Subsituting into expression 22 and simplifying, we obtain

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

By the asymptotic normality of the sample variance-covariance matrix (see Appendix A.5), $\widehat{\beta}_2$ has a distribution that is approximately normal for large samples, with approximate mean given by expression (23). Thus, it makes sense to call the second term in (23) 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 $\beta_2 = 0$. Also, suppose that the conditions of Example 0.6.3.1 hold. The explanatory variables are measured with error, but the data analyst ignores it and tests $H_0 : \beta_2 = 0$ using ordinary regression methods. The test will

be either an F-test or t-test, and since $F = t^2$ in this case, the two tests are the same (assuming a 2-sided t-test). The numerator of the t statistic is (to be continued).

Combined with estimated standard error going almost surely to zero, Get t statistic for $H_0 : \beta_2 = 0$ going to plus/minus infinity, and p-value going almost Surely to zero, unless

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

And, anything that increases $Var(W_2)$ will decrease the bias.

0.6.4 A large scale simulation study

This was covered in lecture.

0.7 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 course: 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 prvided as we go.

0.7.1 A first try at including measurement error

The following is basically the true model of Example 0.6.2.1, with everything normally distributed. Independently for $i = 1, \ldots, n$, let

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_i + \epsilon_i \\ W_i &= \nu + X_i + e_i, \end{aligned} \tag{24}$$

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 (24) 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.

$$\begin{aligned} \mu_1 &= \mu_x + \nu \\ \mu_2 &= \beta_0 + \beta_1 \mu_x \\ \sigma_{1,1} &= \phi + \omega \\ \sigma_{1,2} &= \beta_1 \phi \\ \sigma_{2,2} &= \beta_1^2 \phi + \psi. \end{aligned} \tag{25}$$

It is impossible to solve these five equations for the seven model parameters[13]. 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

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

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 (24) are not *identifiable*. This calls for a brief discussion of identifiability, a topic of central importance in this course.

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

0.8 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 0.8.1 *A* Statistical Model *is a set of assertions that partly*[14] *specify the probability distribution of a set of observable data.*

Definition 0.8.2 *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 0.4 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.

Theorem 2 *If the parameter vector is not identifiable, consistent estimation for all points in the parameter space is impossible.*

In Figure 1, 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

Figure 1: Two parameters values yielding the same probability distribution

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.

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

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

In the discussion of model identification, the definitions are in terms of the *distribution* of the observable data. But we will be using a multivariate normal model, for which the distribution of the observable data corresponds exactly to the pair $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. That means that in practice, the parameter vector is identifiable if it can be recovered from $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, and most of the time it will be "recovered" by solving the moment structure equations, or at least verifying that a solution exists. Why does this work? Because if the parameter vector is a function of the moments (which correspond to the distribution of the data), then it is impossible for two different parameter values to yield the same distribution, because functions produce only one value of their arguments.

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 0.8.3 *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.

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 0.8.4 *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 (25) that come from the regression Model (24). 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 (24), the moment structure equations (25) 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 3 is a strictly mathematical theorem[15] that provides the necessary details.

Theorem 3 *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 3 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 be the assumption throughout most of the course), 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

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

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 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.

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 all the random vectors in the model are multivariate normal, this condition is necessary as well as sufficient.

0.9 Double measurement

Consider again the model of Expression (24), 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.

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

$$\begin{aligned} W_{i,1} &= \nu_1 + X_i + e_{i,1} \\ W_{i,2} &= \nu_2 + X_i + e_{i,2} \\ Y_i &= \beta_0 + \beta_1 X_i + \epsilon_i, \end{aligned} \tag{26}$$

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 $\omega_2 > 0$
- $X_i, e_{i,1}, e_{i,1}$ and ϵ_i are all independent.

The model implies that the triples $\mathbf{D}_i = (W_{i,1}, W_{i,2}, Y_i)'$ 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

$$V(\mathbf{D}_i) = \boldsymbol{\Sigma} = [\sigma_{i,j}] = \begin{bmatrix} \phi + \omega_1 & \phi & \beta_1\phi \\ & \phi + \omega_2 & \beta_1\phi \\ & & \beta_1^2\phi + \psi \end{bmatrix}. \tag{27}$$

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 $\boldsymbol{\Sigma}$ is testable. If $H_0 : \sigma_{1,3} = \sigma_{2,3}$ were rejected, the correctness of the model would be called into question[16]. 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.

[16] 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?"

- 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 $\boldsymbol{\Sigma}$, the remaining three linear equations in four unknowns have infinitely many solutions. There are still infinitely many solutions if $\nu_1 = \nu_2$.

Maximum likelihood for the parameters in the covariance matrix would workup 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 $\boldsymbol{\mu} = (\mu_1, \mu_2, \mu_3)'$ lack meaning and interest[17], 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 A.3.2, 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 (A.15).

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

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

$$L(\boldsymbol{\theta}) = |\boldsymbol{\Sigma}(\boldsymbol{\theta})|^{-n/2}(2\pi)^{-np/2} \exp -\frac{n}{2}\left\{tr(\widehat{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1})\right\}, \tag{28}$$

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 (43), $\boldsymbol{\theta} = (\phi, \omega_1, \omega_2, \beta_1, \psi)'$, and $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ is given by Expression (27). Maximum likelihood estimation is numerical, and the full range of large-sample likelihood methods described in Section A.4 of Appendix A is available.

[17] 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.

0.9.2 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 no intercepts. Independently for $i = 1, \ldots, n$, let

$$\begin{aligned} \mathbf{W}_{i,1} &= \mathbf{X}_i + \mathbf{e}_{i,1} \\ \mathbf{V}_{i,1} &= \mathbf{Y}_i + \mathbf{e}_{i,2} \\ \mathbf{W}_{i,2} &= \mathbf{X}_i + \mathbf{e}_{i,3}, \\ \mathbf{V}_{i,2} &= \mathbf{Y}_i + \mathbf{e}_{i,4}, \\ \mathbf{Y}_i &= \boldsymbol{\beta}\mathbf{X}_i + \boldsymbol{\epsilon}_i \end{aligned} \tag{29}$$

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 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

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

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, 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. $V(\mathbf{e}_{i,1}) = \mathbf{\Omega}_{11}$ need not be diagonal, so method one's errors of measurement for the explanatory variables may be correlated with one another. Similarly, $V(\mathbf{e}_{i,2}) = \mathbf{\Omega}_{22}$ 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 $C(\mathbf{e}_{i,1}, \mathbf{e}_{i,2}) = \mathbf{\Omega}_{12}$. 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 $\mathbf{\Omega}_{11}$ and $\mathbf{\Omega}_{33}$ must be of the same dimension, just as $\mathbf{\Omega}_{22}$ and $\mathbf{\Omega}_{44}$ must be of the same dimension – but none of the corresponding elements need be equal. In particular, the corresponding diagonal elements nee not be equal. This means that measurements of a variable by two different methods do not need to be equally precise.

The model is depicted in Figure 2. 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.

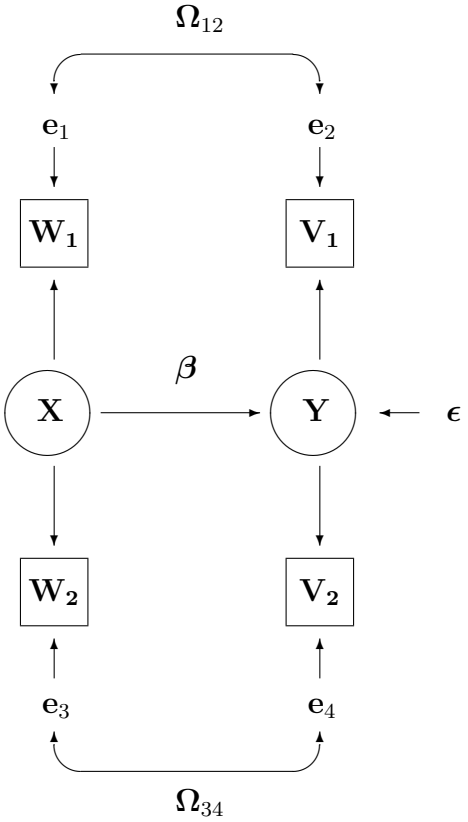
Proof of parameter identifiability The following is typical of easier proofs for structural equation models. The goal is to solve for the model parameters in terms of elements of the variance-covariance matrix of the observable data. This shows the parameters are functions of the distribution, so that 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:

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

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

Figure 2: The Double Measurement Model



Now we express all the $\boldsymbol{\Sigma}_{ij}$ sub-matrices in terms of the parameter matrices of Model (29) 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,

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

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

$$
\begin{aligned}
\boldsymbol{\Sigma} &= \left[\begin{array}{c|c|c|c} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} & \boldsymbol{\Sigma}_{13} & \boldsymbol{\Sigma}_{14} \\ \hline & \boldsymbol{\Sigma}_{22} & \boldsymbol{\Sigma}_{23} & \boldsymbol{\Sigma}_{24} \\ \hline & & \boldsymbol{\Sigma}_{33} & \boldsymbol{\Sigma}_{34} \\ \hline & & & \boldsymbol{\Sigma}_{44} \end{array}\right] \\
&= \left[\begin{array}{c|c|c|c} \boldsymbol{\Phi}+\boldsymbol{\Omega}_{11} & \boldsymbol{\Phi}\boldsymbol{\beta}'+\boldsymbol{\Omega}_{12} & \boldsymbol{\Phi} & \boldsymbol{\Phi}\boldsymbol{\beta}' \\ \hline & \boldsymbol{\beta}\boldsymbol{\Phi}\boldsymbol{\beta}'+\boldsymbol{\Psi}+\boldsymbol{\Omega}_{22} & \boldsymbol{\beta}\boldsymbol{\Phi} & \boldsymbol{\beta}\boldsymbol{\Phi}\boldsymbol{\beta}'+\boldsymbol{\Psi} \\ \hline & & \boldsymbol{\Phi}+\boldsymbol{\Omega}_{33} & \boldsymbol{\Phi}\boldsymbol{\beta}'+\boldsymbol{\Omega}_{34} \\ \hline & & & \boldsymbol{\beta}\boldsymbol{\Phi}\boldsymbol{\beta}'+\boldsymbol{\Psi}+\boldsymbol{\Omega}_{44} \end{array}\right]
\end{aligned} \tag{30}
$$

The equality (30) corresponds to a system of ten matrix equations in nine matrix unknowns. The unknowns are the parameter matrices of Model (29): $\boldsymbol{\Phi}$, $\boldsymbol{\beta}$, $\boldsymbol{\Psi}$, $\boldsymbol{\Omega}_{11}$, $\boldsymbol{\Omega}_{22}$, $\boldsymbol{\Omega}_{33}$, $\boldsymbol{\Omega}_{44}$, $\boldsymbol{\Omega}_{12}$, 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.

$$
\begin{aligned}
\boldsymbol{\Phi} &= \boldsymbol{\Sigma}_{13} \\
\boldsymbol{\beta} &= \boldsymbol{\Sigma}_{23}\boldsymbol{\Phi}^{-1} = \boldsymbol{\Sigma}_{14}'\boldsymbol{\Phi}^{-1} \\
\boldsymbol{\Psi} &= \boldsymbol{\Sigma}_{24} - \boldsymbol{\beta}\boldsymbol{\Phi}\boldsymbol{\beta}' \\
\boldsymbol{\Omega}_{11} &= \boldsymbol{\Sigma}_{11} - \boldsymbol{\Phi} \\
\boldsymbol{\Omega}_{22} &= \boldsymbol{\Sigma}_{22} - \boldsymbol{\beta}\boldsymbol{\Phi}\boldsymbol{\beta}' - \boldsymbol{\Psi} \\
\boldsymbol{\Omega}_{33} &= \boldsymbol{\Sigma}_{33} - \boldsymbol{\Phi} \\
\boldsymbol{\Omega}_{44} &= \boldsymbol{\Sigma}_{44} - \boldsymbol{\beta}\boldsymbol{\Phi}\boldsymbol{\beta}' - \boldsymbol{\Psi} \\
\boldsymbol{\Omega}_{12} &= \boldsymbol{\Sigma}_{12} - \boldsymbol{\Phi}\boldsymbol{\beta}' \\
\boldsymbol{\Omega}_{34} &= \boldsymbol{\Sigma}_{34} - \boldsymbol{\Phi}\boldsymbol{\beta}'
\end{aligned} \tag{31}
$$

This shows that the parameters of Model (29) 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 (30), that $\boldsymbol{\Sigma}_{14} = \boldsymbol{\Sigma}_{23}'$. As in the scalar example of Section 0.9.1 (see page 33), 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.

0.9.3 Intercepts

Now Model (29) 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 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 (29).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 (32) 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 (30), as before. The expected value $\boldsymbol{\mu}$ is the partitioned vector

$$\boldsymbol{\mu} = \left[\begin{array}{c} \boldsymbol{\mu}_1 \\ \hline \boldsymbol{\mu}_2 \\ \hline \boldsymbol{\mu}_3 \\ \hline \boldsymbol{\mu}_4 \end{array}\right] = \left[\begin{array}{c} E(\mathbf{W}_{i,1}) \\ \hline E(\mathbf{V}_{i,1}) \\ \hline E(\mathbf{W}_{i,2}) \\ \hline E(\mathbf{V}_{i,2}) \end{array}\right] = \left[\begin{array}{c} \boldsymbol{\nu}_1 + \boldsymbol{\mu}_x \\ \hline \boldsymbol{\nu}_2 + \boldsymbol{\alpha} + \boldsymbol{\beta}\boldsymbol{\mu}_x \\ \hline \boldsymbol{\nu}_3 + \boldsymbol{\mu}_x \\ \hline \boldsymbol{\nu}_4 + \boldsymbol{\alpha} + \boldsymbol{\beta}\boldsymbol{\mu}_x \end{array}\right]. \tag{32}$$

To demonstrate the identification of Model (32), one would need to solve the equations in (32) 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 (31), this is impossible in most of the parameter space, because (32) 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 (32). This can be viewed as a re-parameterization of the model.

0.9.4 Estimation and testing

Normal model As in the scalar example of Section 0.9.1, 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 (28) on page 35. 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 A.4 of Appendix A; 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 (29) 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 (31) 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 (31), 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:

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

where the subscript M indicates a Method-of-Moments estimator. Using the second line of (31), a reasonable though non-standard estimator of $\boldsymbol{\beta}$ is

$$\widehat{\boldsymbol{\beta}}_M = \frac{1}{2}\left(\widehat{\boldsymbol{\Sigma}}_{14}' + \widehat{\boldsymbol{\Sigma}}_{23}\right)\widehat{\boldsymbol{\Phi}}_M^{-1} \tag{33}$$

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 basis for large-sample interval estimation and testing.

0.10 Instrumental variables

Sometimes, double measurement is not a practical alternative. Usually, this happens because the data are already collected, and the study was designed without planning for a latent variable analysis. The guilty parties might be academic or private sector researchers who do not know what a parameter is, much less parameter identifiability. Or, the data might have been collected for some purpose other than research. For example, a paper mill might report the amount and concentrations of poisonous chemicals they dump into a nearby river. They take the measurements because they have agreed to do so, or because they are required to do it by law — but they certainly are not going to do it twice. Much economic data and public health data is of this kind. In such situations, all one can do is to use what information happens to be available. The instrumental variable method is a lovely trick from Econometrics[18]. It allows for measurement error in the explanatory variables not by measuring the explanatory variables more than once, but by including additional *response variables* in the model.

0.10.1 One explanatory variable

In a simple measurement error regression model like (24), suppose that we have access to data for a second response variables that depends on the latent explanatory variable X_i. Our main interest is still in the response variable Y_i; the second response variable is called an *instrumental* variable because it's just a tool.

Here is the expanded version of Model (24). The original response variable Y_i is now called $Y_{i,1}$. Independently for $i = 1, \ldots, n$.

$$\begin{aligned} W_i &= \nu + X_i + e_i \\ Y_{i,1} &= \alpha_1 + \beta_1 X_i + \epsilon_{i,1} \\ Y_{i,2} &= \alpha_2 + \beta_2 X_i + \epsilon_{i,2} \end{aligned} \tag{34}$$

where e_i, $\epsilon_{i,1}$ and $\epsilon_{i,2}$ are all independent, $Var(X_i) = \phi$, $Var(\epsilon_{i,1}) = \psi_1$, $Var(\epsilon_{i,2}) = \psi_2$, $Var(e_i) = \omega$, $E(X_i) = \mu_x$, and the expected values of all error terms are zero.

It is usually helpful to check the Parameter Count Rule (Rule 1 on page 32) before doing detailed calculations. For this model, the parameter vector is $\boldsymbol{\theta} = (\nu, \alpha_1, \alpha_2, \beta_1, \beta_2, \mu_x, \phi, \omega, \psi_1, \psi_2)$. Writing the vector of observable data for case i as $\mathbf{D}_i = (W_i, Y_{i,1}, Y_{i,2})'$, we see that $\boldsymbol{\mu} = E(\mathbf{D}_i)$ has three elements and $\boldsymbol{\Sigma} = V(\mathbf{D}_i)$ has $3(3+1)/2 = 6$ unique elements. Thus identifiability of the entire parameter vector is ruled out in most of the parameter space.

[18]The instrumental variable method appears for the first time in the appendix of a book published in 1928 by Phillip Wright, the *father* of Sewell Wright, the biologist whose work on path analysis led to modern structural equation modeling as well as much of Econometrics. The story is told in a 2003 paper by Stock and Trebbi [8].

However, it turns out that useful *functions* of the parameter vector are identifiable, and this includes β_1, the parameter of primary interest.

Based on our experience with the double measurement model, we pessimistic about identifying expected values and intercepts. So consider first the covariance matrix. Elements of $\mathbf{\Sigma} = V(\mathbf{D}_i)$ may be obtained by elementary one-variable calculations, like $Var(W_i) = Var(\nu + X_i + e_i) = Var(X_i) + Var(e_i) = \phi + \omega$, and (dropping the subscript i to reduce notational clutter)

$$\begin{aligned} Cov(W, Y_1) &= E(\overset{c}{W}\overset{c}{Y}_1) = E(\overset{c}{X} + e)(\beta_1 \overset{c}{X} + \epsilon_1) = E(\beta_1 \overset{c}{X^2} + \overset{c}{X}\,\epsilon_1 + \beta_1 e\,\overset{c}{X} + e\epsilon_1) \\ &= \beta_1 E(\overset{c}{X^2}) + E(\overset{c}{X}\,\epsilon_1) + \beta_1 E(e\,\overset{c}{X}) + E(e\epsilon_1) \\ &= \beta_1 Var(X) + E(\overset{c}{X})E(\epsilon_1) + \beta_1 E(e)E(\overset{c}{X}) + E(e)E(\epsilon_1) \\ &= \beta_1 \phi \end{aligned}$$

In this way we obtain

$$\mathbf{\Sigma} = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} \\ & \sigma_{22} & \sigma_{23} \\ & & \sigma_{33} \end{pmatrix} = \begin{pmatrix} \phi + \omega & \beta_1\phi & \beta_2\phi \\ & \beta_1^2\phi + \psi_1 & \beta_1\beta_2\phi \\ & & \beta_2^2\phi + \psi_2 \end{pmatrix},$$

which is a nice compact way to look at the six covariance structure equations in six unknown parameters. The fact that there are the same number of equations and unknowns does not guarantee the existence of a unique solution; it merely tells us that a unique solution is possible in most of the parameter space. In fact, identifiability depends on where the true parameter is located.

Since $\sigma_{12} = 0$ if and only if $\beta_1 = 0$, the parameter β_1 is identifiable whenever it equals zero. But then both $\sigma_{12} = 0$ and $\sigma_{23} = 0$, reducing the six equations in six unknowns to four equations in five unknowns, meaning the other parameters in the covariance matrix can't all be recovered.

But what if β_1 does not equal zero? At those points in the parameter space where β_2 is non-zero, $\beta_1 = \frac{\sigma_{23}}{\sigma_{13}}$. This means that adding the instrumental variable Y_2 to the model bought us what we need, which is the possibility of correct estimation and inference about β_1. Note that stipulating $\beta_2 \neq 0$ is not a lot to ask, because it just means that the instrumental variable is related to the response variable.

If both $\beta_1 \neq 0$ and $\beta_2 \neq 0$, all six parameters in the covariance matrix can be recovered

by simple substitutions as follows:

$$\begin{aligned}
\beta_1 &= \frac{\sigma_{23}}{\sigma_{13}} \\
\beta_2 &= \frac{\sigma_{23}}{\sigma_{12}} \\
\phi &= \frac{\sigma_{12}\sigma_{13}}{\sigma_{23}} \\
\omega &= \sigma_{11} - \frac{\sigma_{12}\sigma_{13}}{\sigma_{23}} \\
\psi_1 &= \sigma_{22} - \frac{\sigma_{12}\sigma_{23}}{\sigma_{13}} \\
\psi_2 &= \sigma_{33} - \frac{\sigma_{13}\sigma_{23}}{\sigma_{12}}
\end{aligned}$$

This is a success, but actually the job is not done yet. Four additional parameters appear only in the expected value of the data vector; they are the expected value and intercepts: ν, μ_x, α_1, and α_2. We have

$$\begin{aligned}
\mu_1 &= \nu + \mu_x \\
\mu_2 &= \alpha_1 + \beta_1\mu_x \\
\mu_3 &= \alpha_2 + \beta_2\mu_x
\end{aligned} \tag{35}$$

Even treating β_1 and β_2 as known because they can be identified from the covariance matrix, this system of three linear equations in four unknowns does not have a unique solution.

As in the double measurement case, this lack of identifiability is really not too serious, because our primary interest is in β_1. So we re-parameterize, absorbing the expected value and intercepts into $\boldsymbol{\mu}$ exactly as defined in the mean structure equations (35). The new parameters μ_1, μ_2 and μ_3 may not be too interesting in their own right, but they can be safely estimated by the vector of sample means and then disregarded.

To clarify, the original parameter was

$$\boldsymbol{\theta} = (\nu, \mu_x, \alpha_1, \alpha_2, \beta_1, \beta_2, \phi, \omega, \psi_1, \psi_2).$$

Now it's

$$\boldsymbol{\theta} = (\mu_1, \mu_2, \mu_3, \beta_1, \beta_2, \phi, \omega, \psi_1, \psi_2).$$

The dimension of the parameter space is now one less, and we haven't lost anything that is either accessible or important. This is all the more true because the model pretends that the response variables are measured without error. So the equations for $Y_{i,1}$ and $Y_{i,2}$ should be viewed as re-parameterizations like the one in Expression (20) on page 23, and the intercepts α_1 and α_2 are already the original intercepts plus un-knowable measurement bias terms.

To an important degree, this is the story of structural equation models in general. The models usually used in practice are not what the scientist or statistician originally

had in mind. Instead, they are the result of judicious re-parameterizations, in which the original parameter vector is collapsed into a vector of *functions* that is identifiable, and at the same time allows valid inference about the original parameters that are of primary interest.

Correlation between explanatory variables and error terms Recalling Section 0.4 on omitted variables in regression, it is remarkable that while the primary explanatory variable $X_{i,1}$ must not be correlated with the error term $\epsilon_{i,1}$, the instrumental variable $X_{i,2}$ is allowed to be correlated with the error term $\epsilon_{i,2}$, perhaps reflecting the operation of omitted explanatory variables that affect $Y_{i,2}$ and have non-zero covariance with $X_{i,2}$. Suppose $Cov(X_i, \epsilon_{i,2}) = \kappa$, which might be non-zero. Then the covariance matrix of $\mathbf{D}_i$ becomes

$$\begin{pmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} \\ & \sigma_{22} & \sigma_{23} \\ & & \sigma_{33} \end{pmatrix} = \begin{pmatrix} \phi + \omega & \beta_1\phi & \beta_2\phi + \kappa \\ & \beta_1^2\phi + \psi_1 & \beta_1\beta_2\phi + \beta_1\kappa \\ & & \beta_2^2\phi + \psi_2 + 2\beta_2\kappa \end{pmatrix}.$$

Assuming as before that Y_2 is a useful instrumental variable so that $\beta_2 \neq 0$,

$$\frac{\sigma_{23}}{\sigma_{13}} = \frac{\beta_1(\beta_2\phi + \kappa)}{\beta_2\phi + \kappa} = \beta_1. \tag{36}$$

In fact, if $\kappa \neq 0$, we don't even need $\beta_2 \neq 0$. That is, the instrumental variable need not even be influenced by the explanatory variable. It need only be influenced by some unknown varable that is *correlated* with the explanatory variable.

Testing $H_0 : \beta_1 = 0$ Since primary interest is in the relationship between X and Y_1, this is the null hypothesis we are most likely to try testing, and the most likely technique is a likelihood ratio test or a Wald Z-test. Now the parameter β_1 is identifiable, so a valid test is possible. But when $\beta_1 = 0$ the whole parameter *vector* is not identifiable, and the technical conditions of the likelihood ratio test are not satisfied. It becomes quite interesting; when $\kappa = 0$, the likelihood ratio statistic actually has 2 *df* even though H_0 appears to impose only one restriction on the parameter[19]. We can deal with this kind of complication if we really need to, but everything is much easier with more than one instrumental variable.

More than one instrumental variable Suppose that the data set contains another *two* variables that depend on the latent explanatory variable X_i. Our main interest is still in the response variable $Y_{i,1}$; the other two are instrumental variables. Now the model is,

[19]Notice that $H_0 : \beta_1 = 0$ imposes *two* restrictions on the covariance matrix. These correspond to the two degrees of freedom of the correct test.

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

$$\begin{aligned} W_i &= \nu + X_i + e_i \\ Y_{i,1} &= \alpha_1 + \beta_1 X_i + \epsilon_{i,1} \\ Y_{i,2} &= \alpha_2 + \beta_2 X_i + \epsilon_{i,2} \\ Y_{i,3} &= \alpha_3 + \beta_3 X_i + \epsilon_{i,3}, \end{aligned} \tag{37}$$

where e_i, e_i, $\epsilon_{i,1}$, $\epsilon_{i,2}$ and $\epsilon_{i,3}$ are all independent, $Var(X_i) = \phi$, $Var(\epsilon_{i,1}) = \psi_1$, $Var(\epsilon_{i,2}) = \psi_2$, $Var(\epsilon_{i,3}) = \psi_3$, $Var(e_i) = \omega$, $E(X_i) = \mu_x$ and the expected values of all error terms are zero.

Writing the vector of observable data for case i as $\mathbf{D}_i = (W_i, Y_{i,1}, Y_{i,2}, Y_{i,3})'$,

$$\boldsymbol{\mu} = E \begin{pmatrix} W_i \\ Y_{i,1} \\ Y_{i,2} \\ Y_{i,3} \end{pmatrix} = \begin{pmatrix} \nu + \mu_x \\ \alpha_1 + \beta_1 \mu_x \\ \alpha_2 + \beta_2 \mu_x \\ \alpha_3 + \beta_3 \mu_x \end{pmatrix}$$

and

$$\boldsymbol{\Sigma} = \begin{bmatrix} \phi + \omega & \beta_1 \phi & \beta_2 \phi & \beta_3 \phi \\ & \beta_1^2 \phi + \psi_1 & \beta_1 \beta_2 \phi & \beta_1 \beta_3 \phi \\ & & \beta_2^2 \phi + \psi_2 & \beta_2 \beta_3 \phi \\ & & & \beta_3^2 \phi + \psi_3 \end{bmatrix}. \tag{38}$$

To establish identifiability of the parameters that appear in the covariance matrix, the task is to solve the following ten equations in eight unknowns:

$$\begin{aligned} \sigma_{11} &= \phi + \omega \\ \sigma_{12} &= \beta_1 \phi \\ \sigma_{13} &= \beta_2 \phi \\ \sigma_{14} &= \beta_3 \phi \\ \sigma_{22} &= \beta_1^2 \phi + \psi_1 \\ \sigma_{23} &= \beta_1 \beta_2 \phi \\ \sigma_{24} &= \beta_1 \beta_3 \phi \\ \sigma_{33} &= \beta_2^2 \phi + \psi_2 \\ \sigma_{34} &= \beta_2 \beta_3 \phi \\ \sigma_{44} &= \beta_3^2 \phi + \psi_3 \end{aligned} \tag{39}$$

for ϕ, ω, β_1, β_2, β_3, ψ_1, ψ_2, and ψ_3. Assuming the instrumental variables are well-chosen, so that both β_2 and β_3 are both non-zero,

$$\frac{\sigma_{13}\sigma_{14}}{\sigma_{34}} = \frac{\beta_2 \beta_3 \phi^2}{\beta_2 \beta_3 \phi} = \phi. \tag{40}$$

Then, simple substitutions allow us to solve for the rest of the parameters, yielding the complete solution

$$
\begin{aligned}
\phi &= \frac{\sigma_{13}\sigma_{14}}{\sigma_{34}} \\
\omega &= \sigma_{11} - \frac{\sigma_{13}\sigma_{14}}{\sigma_{34}} \\
\beta_1 &= \frac{\sigma_{12}\sigma_{34}}{\sigma_{13}\sigma_{14}} \\
\beta_2 &= \frac{\sigma_{34}}{\sigma_{14}} \\
\beta_3 &= \frac{\sigma_{34}}{\sigma_{13}} \\
\psi_1 &= \sigma_{22} - \frac{\sigma_{12}^2\sigma_{34}}{\sigma_{13}\sigma_{14}} \\
\psi_2 &= \sigma_{33} - \frac{\sigma_{13}\sigma_{34}}{\sigma_{14}} \\
\psi_3 &= \sigma_{44} - \frac{\sigma_{14}\sigma_{34}}{\sigma_{13}}
\end{aligned} \tag{41}
$$

This proves identifiability. The solution is thorough but somewhat tedious, even for this simple example. The student may wonder how much work really needs to be shown. I would suggest showing the calculations leading to the covariance matrix (38), saying "Denote the i, j element of $\boldsymbol{\Sigma}$ by σ_{ij}," skipping the system of equations (39) because they are present in (38), and showing the solution for ϕ in (40), *including* the stipulation that β_2 and β_3 are both non-zero. Then, instead of the explicit solution (41), write something like this:

$$
\begin{aligned}
\omega &= \sigma_{11} - \phi \\
\beta_1 &= \frac{\sigma_{12}}{\phi} \\
\beta_2 &= \frac{\sigma_{13}}{\phi} \\
\beta_3 &= \frac{\sigma_{14}}{\phi} \\
\psi_1 &= \sigma_{22} - \beta_1^2\phi \\
\psi_2 &= \sigma_{33} - \beta_2^2\phi \\
\psi_3 &= \sigma_{44} - \beta_3^2\phi
\end{aligned}
$$

Notice how once we have solved for a model parameter, we use it to solve for other parameters without explicitly substituting in terms of σ_{ij}. The objective is to prove that a unique solution exists by showing how to get it. A full statement of the solution is not necessary unless you need it for some other purpose.

Turning to the mean structure equations, five additional parameters appear only in the expected value $\mathbf{D}_i$; they are $\nu, \mu_x, \alpha_1, \alpha_2$ and α_3. Even treating β_1, β_2 and β_3 as known

because they are identified from the covariance matrix, the resulting four linear equations in five unknowns does not have a unique solution.

$$\begin{aligned}\mu_1 &= \nu + \mu_x \\ \mu_2 &= \alpha_1 + \beta_1\mu_x \\ \mu_3 &= \alpha_2 + \beta_2\mu_x \\ \mu_4 &= \alpha_3 + \beta_3\mu_x\end{aligned}$$

As in the case of a single instrumental variable we re-parameterize, absorbing the expected value and intercepts into $\boldsymbol{\mu}$. The new parameters $\mu_1, \ldots, \mu_4$ may not be too interesting in their own right, but they can be safely estimated by the vector of sample means and then disregarded.

With two (or more) instrumental variables, the identifiability argument does not need to be as fussy about the locations in the parameter space where different functions of the parameter vector are identifiable. In particular, there is no loss of identifiability under the natural null hypothesis that $\beta_1 = 0$, and testing that null hypothesis presents no special difficulties.

Constraints on the covariance matrix Like the double measurement model, the model with one explanatory variable and two instrumental variables imposes equality constraints on the covariance matrix of the observable data. In the solution given by Expression (41), the critical parameter β_1 is recovered by $\beta_1 = \frac{\sigma_{12}\sigma_{34}}{\sigma_{13}\sigma_{14}}$, but a look at the covariance structure equations (39) shows that $\beta_1 = \frac{\sigma_{23}}{\sigma_{13}}$ and $\beta_1 = \frac{\sigma_{24}}{\sigma_{14}}$ are also correct. These seemingly different ways of solving for the parameter must be the same. That is,

$$\frac{\sigma_{12}\sigma_{34}}{\sigma_{13}\sigma_{14}} = \frac{\sigma_{23}}{\sigma_{13}} \text{ and } \frac{\sigma_{12}\sigma_{34}}{\sigma_{13}\sigma_{14}} = \frac{\sigma_{24}}{\sigma_{14}}.$$

Simplifying a bit yields

$$\sigma_{12}\sigma_{34} = \sigma_{14}\sigma_{23} = \sigma_{13}\sigma_{24}. \tag{42}$$

Since all three products equal $\beta_1\beta_2\beta_3\phi^2$, it is clear that the model implies the equality constraints (42) even where the identifiability conditions $\beta_2 \neq 0$ and $\beta_3 \neq 0$ do not hold.

What is happening geometrically is that the covariance structure equations are mapping a parameter space[20] of dimension eight into a moment space of dimension ten. The image of the parameter space is an eight-dimensional surface in the moment space, contained in the set defined by the relations (42). Ten minus eight equals two, the number of over-identifying restrictions.

Here are two more comments. First, we will see that even models with non-identifiable parameters can imply equality constraints. Second, models also frequently imply *inequality* constraints on the moments. For example, in (41), $\phi = \frac{\sigma_{13}\sigma_{14}}{\sigma_{34}}$. Because ϕ is a variance, we have the inequality restriction $\frac{\sigma_{13}\sigma_{14}}{\sigma_{34}} > 0$, something that is not automatically true

[20] Actually it's a subset of the parameter space, containing just those parameters that appear in the covariance matrix,

of covariance matrices in general. Most structural equation models imply quite a few inequality restrictions, and locating them all and listing them in non-redundant form can be challenging. But any fact that suggests a way of disconfirming a model can be a valuable tool.

0.10.2 Multiple explanatory variables

Most real-life models have more than one explanatory variable. No special difficulties arise for the method of instrumental variables. In fact, the presence of multiple explanatory variables only provides more ways to identify the parameters and more over-identifying restrictions.

Here is an example with just two explanatory variables and two instrumental variables. Independently for $i = 1, \ldots, n$,

$$\begin{aligned} W_{i,1} &= \nu_1 + X_{i,1} + e_{i,1} \\ Y_{i,1} &= \alpha_1 + \beta_1 X_{i,1} + \epsilon_{i,1} \\ Y_{i,2} &= \alpha_2 + \beta_2 X_{i,1} + \epsilon_{i,2} \\ W_{i,2} &= \nu_2 + X_{i,2} + e_{i,2} \\ Y_{i,3} &= \alpha_3 + \beta_3 X_{i,2} + \epsilon_{i,3} \\ Y_{i,4} &= \alpha_4 + \beta_4 X_{i,2} + \epsilon_{i,4} \end{aligned}$$

where $E(X_{i,j}) = \mu_j$, $e_{i,j}$ and $\epsilon_{i,j}$ are independent of one another and of $X_{i,j}$, $Var(e_{i,j}) = \omega_j$, $Var(\epsilon_{i,j}) = \psi_j$, and

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

As usual, intercepts and expected values can't be recovered individually. Eight parameters are intercepts and expected values of latent variables that appear in the expressions for only six expected values of the observable variables. So we re-parameterize, absorbing them into $\mu_1, \ldots, \mu_6$. Then we estimate $\boldsymbol{\mu}$ with the vector of 6 sample means and set it aside, forever.

Denoting the data vectors by $\mathbf{D}_i = (W_{i,1}, Y_{i,1}, Y_{i,2}, W_{i,2}, Y_{i,3}, Y_{i,4})'$, the covariance matrix $\boldsymbol{\Sigma} = V(\mathbf{D}_i)$ is

$$[\sigma_{ij}] = \begin{pmatrix} \phi_{11}+\omega_1 & \beta_1\phi_{11} & \beta_2\phi_{11} & \phi_{12} & \beta_3\phi_{12} & \beta_4\phi_{12} \\ & \beta_1^2\phi_{11}+\psi_1 & \beta_1\beta_2\phi_{11} & \beta_1\phi_{12} & \beta_1\beta_3\phi_{12} & \beta_1\beta_4\phi_{12} \\ & & \beta_2^2\phi_{11}+\psi_2 & \beta_2\phi_{12} & \beta_2\beta_3\phi_{12} & \beta_2\beta_4\phi_{12} \\ & & & \phi_{22}+\omega_2 & \beta_3\phi_{22} & \beta_4\phi_{22} \\ & & & & \beta_3^2\phi_{22}+\psi_3 & \beta_3\beta_4\phi_{22} \\ & & & & & \beta_4^2\phi_{22}+\psi_4 \end{pmatrix}$$

Disregarding the expected values, the parameter[21] is

$$\boldsymbol{\theta} = (\beta_1, \beta_2, \beta_3, \beta_4, \phi_{11}, \phi_{12}, \phi_{22}, \omega_1, \omega_2, \psi_1, \psi_2, \psi_3, \psi_4).$$

Since $\boldsymbol{\theta}$ has 13 elements and $\boldsymbol{\Sigma}$ has $\frac{6(6+1)}{2} = 21$ variances and non-redundant covariances, this problem easily passes the test of the parameter count rule. Provided the parameter vector is identifiable, the model will impose $21 - 13 = 8$ over-identifying restrictions on $\boldsymbol{\Sigma}$.

First notice that if $\phi_{12} \neq 0$, all the regression coefficients are immediately identifiable. Since the instrumental variables Y_2 and Y_4 are presumably well-chosen, it may be assumed that $\beta_2 \neq 0$ and $\beta_4 \neq 0$. In that case, the entire parameter vector is identifiable — for example identifying ϕ_{11} from σ_{12} and then ω_1 from σ_{11} ....

Since it is very common for explanatory variables to be related to one another in non-experimental studies, assumptions like $\phi_{12} \neq 0$ are very reasonable, and in any case are testable as part of an exploratory data analysis. So, extension of this design to data sets with more than two explanatory variables is straightforward, and identifiability follows without detailed calculations.

Be aware, though, that the instrumental variable models presented here are actually re-parameterizations of models with measurement error in the response variables. One must carefully consider the methods of data collection to rule out correlation between measurement error in the explanatory variables and measurement error in the response variables. Such correlations would appear as non-zero covariances between e_{ij} and ϵ_{ij} terms in the models, and it will be seen in homework how this can sink the ship on a technical level.

Just to be clear, when data are collected by a common method in a common setting, errors of measurement will naturally be correlated with one another. For example, in a study investigating the connection between diet and athletic accomplishment in children, suppose the data all came from questionnaires filled out by parents. It would be very natural for some parents to exaggerate the healthfulness of the food they serve and also to exaggerate their children's athletic achievements. On the other extreme, some parents would immediately figure out the purpose of the study, and tell the interviewers what they want to hear. "My kids eat junk (I can't control them) and they are terrible in sports." Both these tendencies would produce a positive covariance between the measurement errors in the explanatory and response variables. And in the absence of other information, it would be impossible to tell whether a positive relationship between observable diet and athletic performance came from this, or from an actual relationship between the latent variables.

[21] Since the distributions of the random variables in the model are unspecified, one could say that they are also unknown parameters. In this case, the quantity $\boldsymbol{\theta}$ is really a function of the full parameter vector, even after re-parameterization.