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. 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 Covariance and Relationship

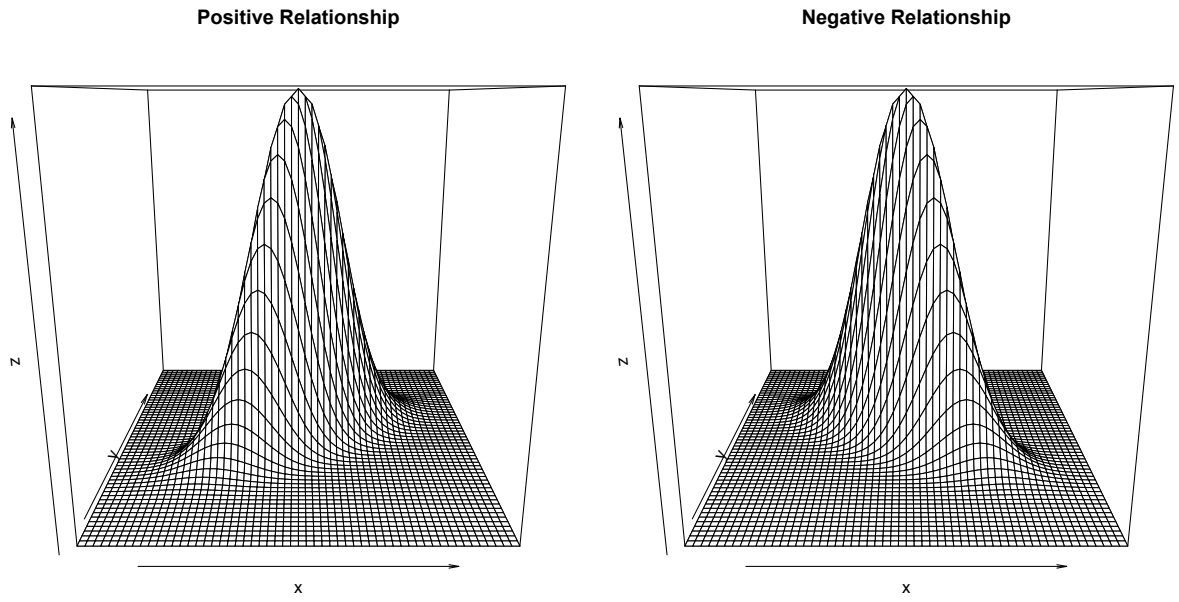
Most of the models we will consider are linear in the explanatory variables as well as the regression parameters, and so relationships between explanatory variables and response variables are represented by covariances. To clarify this fundamental point, first note that saying two random variables are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

More generally, consider two random variables X and Y whose joint distribution might not be bivariate normal. Suppose there is a tendency for higher values of X to go with higher values of Y, and for lower values of X to go with lower values of Y. This idea of a "positive" relationship is pictured in the left panel of Figure 1. Since the probability of an (x, y) pair is roughly proportional to the height of the surface, a large sample of points will be most dense where the surface is highest[1]. On a scatterplot, the best-fitting

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

line will have a positive slope. The right panel of Figure 1 shows a negative relationship. There, the best-fitting line will have a negative slope.

Figure 1: Relationship between X and Y



The word "covariance" suggests that it is a measure of how X and Y vary together. To see that positive relationships yield positive covariances and negative relationships yield negative covariances, look at Figure 2.

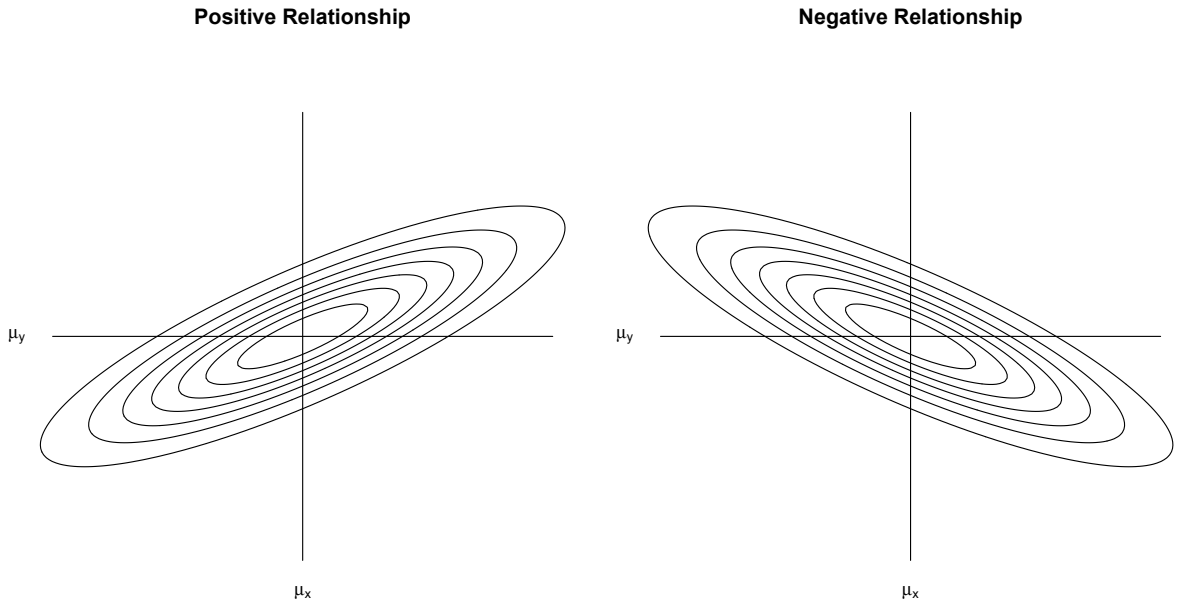
Figure 2 shows contour plots of the densities in Figure 1. Imagine you are looking down at a density from directly above, and that the density has been cut into slices that are parallel with the x, y plane. The ellipses are the cut marks. The outer ellipse is lowest, the next one in is a bit higher, and so on. All the points on an ellipse (contour) are at the same height. It's like a topographic map of a mountainous region, except that the contours on maps are not so regular.

The definition of covariance is

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

In the left panel of Figure 2, more of the probability is in the upper right and lower left, and that is where $(x-\mu_x)(y-\mu_y)$ is positive. The positive volume in these regions is greater than the negative volume in the upper left and lower right, so that the integral is positive. In the right-hand panel the opposite situation occurs, and the covariance is negative. The pictures are just of one example, but the rule is general. Positive covariances reflect positive relationships and negative covariances reflect negative relationships.

Figure 2: Contour Plots



In the study of linear structural equation models, one frequently needs to calculate covariances and matrices of covariances. Covariances of linear combinations are frequently required. The following rules are so useful that they are repeated from Sections A.1 and A.3 of Appendix A.

Let $X_1, \ldots, X_{n_1}$ and $Y_1, \ldots, Y_{n_2}$ be scalar random variables, and define the linear combinations L_1 and L_2 by

$$\begin{aligned} L_1 &= a_1 X_1 + \cdots + a_{n_1} X_{n_1} = \sum_{i=1}^{n_1} a_i X_i, \text{ and} \\ L_2 &= b_1 Y_1 + \cdots + b_{n_2} Y_{n_2} = \sum_{i=1}^{n_2} b_i Y_i, \end{aligned}$$

where the a_j and b_j are constants. Then

$$cov(L_1, L_2) = \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} a_i b_j Cov(X_i, Y_j). \tag{1}$$

In the matrix version, let $\mathbf{X}_1, \ldots, \mathbf{X}_{n_1}$ and $\mathbf{Y}_1, \ldots, \mathbf{Y}_{n_2}$ be random vectors, and define

the linear combinations $\mathbf{L}_1$ and $\mathbf{L}_2$ by

$$\begin{aligned}\mathbf{L}_1 &= \mathbf{A}_1\mathbf{X}_1 + \cdots + \mathbf{A}_{n_1}\mathbf{X}_{n_1} = \sum_{i=1}^{n_1}\mathbf{A}_i\mathbf{X}_i, \text{ and} \\ \mathbf{L}_2 &= \mathbf{B}_1\mathbf{Y}_1 + \cdots + \mathbf{B}_{n_2}\mathbf{Y}_{n_2} = \sum_{i=1}^{n_2}\mathbf{B}_i\mathbf{Y}_i,\end{aligned}$$

where the $\mathbf{A}_j$ and $\mathbf{B}_j$ are matrices of constants. Then

$$cov(\mathbf{L}_1, \mathbf{L}_2) = \sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\mathbf{A}_i\, cov(\mathbf{X}_i, \mathbf{Y}_j)\, \mathbf{B}_j^\top. \tag{2}$$

Both these results say that to calculate the covariance of two linear combinations, just take the covariance of each term in the first linear combination with each term in the second linear combination, and add. When simplifying the results of calculations, it can be helpful to recall that $Cov(X, X) = Var(X)$ and $cov(\mathbf{x}, \mathbf{x}) = cov(\mathbf{x})$.

0.2 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

$$\mathbf{y} \;=\; \mathbf{X}\boldsymbol{\beta} \;+\; \boldsymbol{\epsilon}, \tag{3}$$

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[2] data, it makes more sense to view the explanatory variables as random variables rather than constants. Why? Because if you took repeated

[2] *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.

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

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

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

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

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

The following calculation might make the double expectation a bit clearer. The outer expected value is with respect to the joint probability distribution of the explanatory variable values – all n vectors of them; think of the $n \times p$ matrix $\mathbf{X}$. To avoid unfamiliar

notation, suppose they are all continuous, with joint density $f(\mathbf{x})$. Then

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

Consistent Estimation It will now be shown that when the explanatory variable values are random, $\widehat{\boldsymbol{\beta}}_n \xrightarrow{p} \boldsymbol{\beta}$; see Section A.5 in Appendix A for a brief discussion of consistency. The demonstation is a bit lengthy; the details are shown because one of the intermediate results will be very useful later. The argument begins by establishing an alternative formula for the ordinary least-squares estimates. The explanatory variable values are fixed for now, but in the end, the formula will be applied to random X values.

A regression model can be "centered" by subtracting sample means from the values of the explanatory variables. Geometrically, what this does is to shift the cloud of points in a high-dimensional scatterplot left or right along each x axis – or equivalently, to adopt a shifted set of co-ordinate axes. Clearly, this will not affect the tilt (slopes) of the best-fitting hyperplane, but it will affect the intercept. Writing the regression model in scalar form and then centering, ...

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_{i,1} + \cdots + \beta_p x_{i,p} + \epsilon_i \\ &= \beta_0 + \beta_1 \overline{x}_1 + \cdots + \beta_p \overline{x}_p \\ &\quad + \beta_1(x_{i,1} - \overline{x}_1) + \cdots + \beta_p(x_{i,p} - \overline{x}_p) + \epsilon_i \\ &= \alpha_0 + \alpha_1(x_{i,1} - \overline{x}_1) + \cdots + \alpha_p(x_{i,p} - \overline{x}_p) + \epsilon_i, \end{aligned}$$

where the α parameters are the regression coefficients of the centered model. We have $\alpha_0 = \beta_0 + \beta_1\overline{x}_1 + \cdots + \beta_k\overline{x}_p$, and $\alpha_j = \beta_j$ for $j = 1, \ldots, p$. This re-parameterization is one-to-one. Since the least-squares and maximum likelihood estimates coincide for multiple regression with normal errors, the invariance principle of maximum likelihood estimation (See Section A.6.3 in Appendix A) says that $\widehat{\alpha}_j = \widehat{\beta}_j$ for $j = 1, \ldots, p$. That is, centering does not change the estimated slopes. In addition, the MLE of the intercept for the centered model is $\widehat{\alpha}_0 = \widehat{\beta}_0 + \widehat{\beta}_1\overline{x}_1 + \cdots + \widehat{\beta}_p\overline{x}_p$.

For any regression model with an intercept, the sum of residuals is zero. Thus,

$$\begin{aligned}\bar{y} &= \frac{1}{n}\sum_{i=1}^{n}\widehat{y}_i \\ &= \frac{1}{n}\sum_{i=1}^{n}\left(\widehat{\beta}_0 + \widehat{\beta}_1 x_{i,1} + \cdots + \widehat{\beta}_p x_{i,p}\right) \\ &= \widehat{\beta}_0 + \widehat{\beta}_1\overline{x}_1 + \cdots + \widehat{\beta}_p\overline{x}_p \\ &= \widehat{\alpha}_0\end{aligned}$$

That is, the least-squares estimate of the intercept is $\bar{y}$ for any centered regression model, regardless of the data.

We already know how to calculate the $\widehat{\beta}_j$, but we are working toward another formula for them. Suppose we start with the centered model

$$y_i = \alpha_0 + \beta_1(x_{i,1} - \overline{x}_1) + \cdots + \beta_p(x_{i,p} - \overline{x}_p) + \epsilon_i.$$

Because this is a centered model, we know that $\widehat{\alpha}_0 = \overline{y}$. To find the $\widehat{\beta}_j$, first substitute $\widehat{\alpha}_0 = \overline{y}$ and then minimize

$$Q(\boldsymbol{\beta}) = \sum_{i=1}^{n}\left(y_i - \overline{y}\ - \beta_1(x_{i,1} - \overline{x}_1) - \cdots - \beta_p(x_{i,p} - \overline{x}_p)\ \right)^2$$

over all $\boldsymbol{\beta}$. This is the same as centering y as well as x, and then fitting a regression through the origin. The usual formula $\widehat{\boldsymbol{\beta}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$ applies. We just need to remember that the columns of the $n \times p$ matrix $\mathbf{X}$ are centered, and so is the $n \times 1$ vector $\mathbf{y}$. For $p = 3$, the $\mathbf{X}$ matrix looks like this:

$$\begin{pmatrix} x_{11} - \bar{x}_1 & x_{12} - \bar{x}_2 & x_{13} - \bar{x}_3 \\ x_{21} - \bar{x}_1 & x_{22} - \bar{x}_2 & x_{23} - \bar{x}_3 \\ x_{31} - \bar{x}_1 & x_{32} - \bar{x}_2 & x_{33} - \bar{x}_3 \\ \vdots & \vdots & \vdots \\ x_{n1} - \bar{x}_1 & x_{n2} - \bar{x}_2 & x_{n3} - \bar{x}_3 \end{pmatrix}.$$

The $\mathbf{X}^\top\mathbf{X}$ matrix, the so-called the "sums of squares and cross products" matrix, is

$$\begin{aligned}\mathbf{X}^\top\mathbf{X} &= \begin{pmatrix} x_{11} - \bar{x}_1 & x_{21} - \bar{x}_1 & x_{31} - \bar{x}_1 & \cdots & x_{n1} - \bar{x}_1 \\ x_{12} - \bar{x}_2 & x_{22} - \bar{x}_2 & x_{32} - \bar{x}_2 & \cdots & x_{n2} - \bar{x}_2 \\ x_{13} - \bar{x}_3 & x_{23} - \bar{x}_3 & x_{33} - \bar{x}_3 & \cdots & x_{n3} - \bar{x}_3 \end{pmatrix} \begin{pmatrix} x_{11} - \bar{x}_1 & x_{12} - \bar{x}_2 & x_{13} - \bar{x}_3 \\ x_{21} - \bar{x}_1 & x_{22} - \bar{x}_2 & x_{23} - \bar{x}_3 \\ x_{31} - \bar{x}_1 & x_{32} - \bar{x}_2 & x_{33} - \bar{x}_3 \\ \vdots & \vdots & \vdots \\ x_{n1} - \bar{x}_1 & x_{n2} - \bar{x}_2 & x_{n3} - \bar{x}_3 \end{pmatrix} \\ &= \begin{pmatrix} \sum_{i=1}^n (x_{i1} - \overline{x}_1)^2 & \sum_{i=1}^n (x_{i1} - \overline{x}_1)(x_{i2} - \overline{x}_2) & \sum_{i=1}^n (x_{i1} - \overline{x}_1)(x_{i3} - \overline{x}_3) \\ \sum_{i=1}^n (x_{i2} - \overline{x}_2)(x_{i1} - \overline{x}_1) & \sum_{i=1}^n (x_{i2} - \overline{x}_2)^2 & \sum_{i=1}^n (x_{i2} - \overline{x}_2)(x_{i3} - \overline{x}_3) \\ \sum_{i=1}^n (x_{i3} - \overline{x}_3)(x_{i1} - \overline{x}_1) & \sum_{i=1}^n (x_{i3} - \overline{x}_3)(x_{i2} - \overline{x}_2) & \sum_{i=1}^n (x_{i3} - \overline{x}_3)^2 \end{pmatrix}.\end{aligned}$$

It's clear that larger examples would follow this same pattern. The entries in the matrix look like sample variances and covariances, except that they are not divided by n. Dividing and multiplying by n, we have $\mathbf{X}^\top\mathbf{X} = n\widehat{\boldsymbol{\Sigma}}_x$, where $\widehat{\boldsymbol{\Sigma}}_x$ is the sample variance-covariance matrix of the explanatory variables.

Still looking at the $p = 3$ case for simplicity,

$$\begin{aligned}
\mathbf{X}^\top\mathbf{y} &= \begin{pmatrix} x_{11}-\bar{x}_1 & x_{21}-\bar{x}_1 & x_{31}-\bar{x}_1 & \cdots & x_{n1}-\bar{x}_1 \\ x_{12}-\bar{x}_2 & x_{22}-\bar{x}_2 & x_{32}-\bar{x}_2 & \cdots & x_{n2}-\bar{x}_2 \\ x_{13}-\bar{x}_3 & x_{23}-\bar{x}_3 & x_{33}-\bar{x}_3 & \cdots & x_{n3}-\bar{x}_3 \end{pmatrix} \begin{pmatrix} y_1-\bar{y} \\ y_2-\bar{y} \\ y_3-\bar{y} \\ \vdots \\ y_n-\bar{y} \end{pmatrix} \\
&= \begin{pmatrix} \sum_{i=1}^n (x_{i1}-\overline{x}_1)(y_i-\overline{y}) \\ \sum_{i=1}^n (x_{i1}-\overline{x}_1)(y_i-\overline{y}) \\ \sum_{i=1}^n (x_{i1}-\overline{x}_1)(y_i-\overline{y}) \end{pmatrix} \\
&= n \begin{pmatrix} \frac{1}{n}\sum_{i=1}^n (x_{i1}-\overline{x}_1)(y_i-\overline{y}) \\ \frac{1}{n}\sum_{i=1}^n (x_{i1}-\overline{x}_1)(y_i-\overline{y}) \\ \frac{1}{n}\sum_{i=1}^n (x_{i1}-\overline{x}_1)(y_i-\overline{y}) \end{pmatrix} \\
&= n\widehat{\boldsymbol{\Sigma}}_{xy},
\end{aligned}$$

where $\widehat{\boldsymbol{\Sigma}}_{xy}$ is the $k \times 1$ vector of sample covariances between the explanatory variables and the response variable.

Putting the pieces together, the least squares estimator of $\boldsymbol{\beta}$ is

$$\begin{aligned}
\widehat{\boldsymbol{\beta}}_n &= (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y} \\
&= (n\widehat{\boldsymbol{\Sigma}}_x)^{-1} n\widehat{\boldsymbol{\Sigma}}_{xy} \\
&= \frac{1}{n}(\widehat{\boldsymbol{\Sigma}}_x)^{-1} n\widehat{\boldsymbol{\Sigma}}_{xy} \\
&= \widehat{\boldsymbol{\Sigma}}_x^{-1}\widehat{\boldsymbol{\Sigma}}_{xy}. \qquad (4)
\end{aligned}$$

Several comments are in order. First, recall that $\widehat{\boldsymbol{\beta}}_n$ is a vector of least-squares slopes only. It does not include the intercept. However, the intercept for a centered model is $\bar{y}$, and is easily computed. Second, because the slopes are the same for the centered model and the uncentered model, formula (4) applies equally to centered and uncentered models. Third, in spite of the suggestive $\widehat{\boldsymbol{\Sigma}}$ notation, expression (4) is just a computational formula. It applies whether the explanatory variable values are random or fixed. Only when the variables are random do $\widehat{\boldsymbol{\Sigma}}_x$ and $\widehat{\boldsymbol{\Sigma}}_{xy}$ actually estimate variances and covariances.

When the explanatory variables are random, the Strong Law of Large Numbers and continuous mapping yield

$$\widehat{\boldsymbol{\beta}}_n \stackrel{a.s.}{\to} \boldsymbol{\Sigma}_x^{-1}\boldsymbol{\Sigma}_{xy}. \tag{5}$$

Note that the convergence (5) requires only that $\boldsymbol{\Sigma}_x$ be positive definite, so that the inverse exists. Also, it applies regardless of whether the regression model is correct or not. For this reason, it can be a valuable tool for studying *mis-specified* regression models — that is, models that are assumed, but are not actually correct. If you can calculate $\widehat{\boldsymbol{\Sigma}}_x$ and $\widehat{\boldsymbol{\Sigma}}_{xy}$ under the true model, you can determine where the estimated regression coefficients are going as the sample size increases. This will often indicate whether the mis-specification is likely to cause mistaken conclusions.

For the present, suppose that the usual regression model is correct. Independently for $i = 1, \ldots, n$, let

$$y_i = \beta_0 + \boldsymbol{\beta}^\top \mathbf{X}_i + \epsilon_i$$

where

β_0 (the intercept) is an unknown scalar constant.

$\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown parameters.

$\mathbf{X}_i$ is a $p \times 1$ random vector with expected value $\boldsymbol{\mu}$ and positive definite covariance matrix $\boldsymbol{\Sigma}_x$.

ϵ_i is a scalar random variable with $E(\epsilon_i) = 0$ and $Var(\epsilon_i) = \sigma^2$.

$cov(\mathbf{X}_i, \epsilon_i) = \mathbf{0}$.

So,

$$\begin{aligned}
\boldsymbol{\Sigma}_{xy} &= cov(\mathbf{X}_i, y_i) \\
&= cov(\mathbf{X}_i, \beta_0 + \boldsymbol{\beta}^\top \mathbf{X}_i + \epsilon_i) \\
&= cov(\mathbf{X}_i, \boldsymbol{\beta}^\top \mathbf{X}_i + \epsilon_i) \\
&= cov(\mathbf{X}_i, \boldsymbol{\beta}^\top \mathbf{X}_i) + cov(\mathbf{X}_i, \epsilon_i) \\
&= cov(\mathbf{X}_i, \mathbf{X}_i)\boldsymbol{\beta} + \mathbf{0} \\
&= \boldsymbol{\Sigma}_x \boldsymbol{\beta}.
\end{aligned}$$

Then by (5)

$$\begin{aligned}
\widehat{\boldsymbol{\beta}}_n &\stackrel{a.s.}{\to} \boldsymbol{\Sigma}_x^{-1}\boldsymbol{\Sigma}_{xy} \\
&= \boldsymbol{\Sigma}_x^{-1}\boldsymbol{\Sigma}_x\boldsymbol{\beta} \\
&= \boldsymbol{\beta}.
\end{aligned}$$

Since almost sure convergence implies convergence in probability (see Section A.5 in Appendix A), we have $\widehat{\boldsymbol{\beta}}_n \stackrel{p}{\to} \boldsymbol{\beta}$. This is the standard definition of (weak) consistency. The meaning is that as the sample size increases, the probability that the usual least-squares estimate $\widehat{\boldsymbol{\beta}}_n$ is arbitrarily close to $\boldsymbol{\beta}$ – even though the explanatory variable values are random variables. This holds even though $\widehat{\boldsymbol{\beta}}_n$ was derived under the assumption that the x_{ij} were fixed constants.

Size α Tests Suppose Model (3) 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,

$$\begin{aligned} Pr\{F \in A\} &= \int \cdots \int \alpha f(\mathbf{x}) \, d\mathbf{x} \\ &= \alpha \int \cdots \int f(\mathbf{x}) \, d\mathbf{x} \\ &= \alpha. \end{aligned} \tag{6}$$

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 fear is unfounded.

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.6.5 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

$$\begin{aligned} L(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \mathbf{x}, \mathbf{y}) &= f_{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2}(\mathbf{x}, \mathbf{y}) \\ &= f_{\boldsymbol{\theta}_2}(\mathbf{y}|\mathbf{x}) \, f_{\boldsymbol{\theta}_1}(\mathbf{x}) \\ &= L_2(\boldsymbol{\theta}_2, \mathbf{x}, \mathbf{y}) \, L_1(\boldsymbol{\theta}_1, \mathbf{x}) \end{aligned} \tag{7}$$

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 (7) 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 (7), it's a fairly reasonable set of assumptions.

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

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

0.3 Unconditional regression with observed variables

Example 0.3.1 *Simple Regression*

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

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \tag{8}$$

where

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

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

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

and variance-covariance matrix

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

The linear property of the covariance (Expression 1 on page 9) is useful for calculating the covariance between the explanatory and response variables.

$$\begin{aligned} Cov(X_i, Y_i) &= Cov(X_i, \beta_0 + \beta_1 X_i + \epsilon_i) \\ &= Cov(X_i, \beta_1 X_i + \epsilon_i) \\ &= \beta_1 Cov(X_i, X_i) + Cov(X_i, \epsilon_i) \\ &= \beta_1 Var(X_i) + 0 \\ &= \beta_1\phi \end{aligned}$$

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

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

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

$$\begin{aligned} \rho_{xy} &= \frac{Cov(X,Y)}{SD(X)SD(Y)} \\ &= \frac{E\{(X-\mu_x)(Y-\mu_y)\}}{\sqrt{Var(X)}\sqrt{Var(Y)}} \\ &= E\left\{\left(\frac{X-\mu_x}{\sigma_x}\right)\left(\frac{Y-\mu_y}{\sigma_y}\right)\right\}. \end{aligned} \tag{9}$$

That is, the correlation between two random variables is the covariance between versions of the variables that have been standardized to have mean zero and variance one. Using (9),

the correlation for Example 0.3.1 is

$$\begin{aligned} \rho &= \frac{\beta_1\phi}{\sqrt{\phi}\sqrt{\beta_1^2\phi+\sigma^2}} \\ &= \frac{\beta_1\sqrt{\phi}}{\sqrt{\beta_1^2\phi+\sigma^2}}. \end{aligned} \tag{10}$$

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

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

Thus the variance of Y_i is separated into two parts[3], the part that comes from X_i and the part that comes from ϵ_i. The part that comes from X_i is $\beta_1^2\phi$, and the part that comes from ϵ_i (that is, everything else) is σ^2. From (10) the *squared* correlation between X_i and Y_i is

$$\rho^2 = \frac{\beta_1^2\phi}{\beta_1^2\phi+\sigma^2}, \tag{11}$$

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

Example 0.3.2 *Multiple Regression*

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

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

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

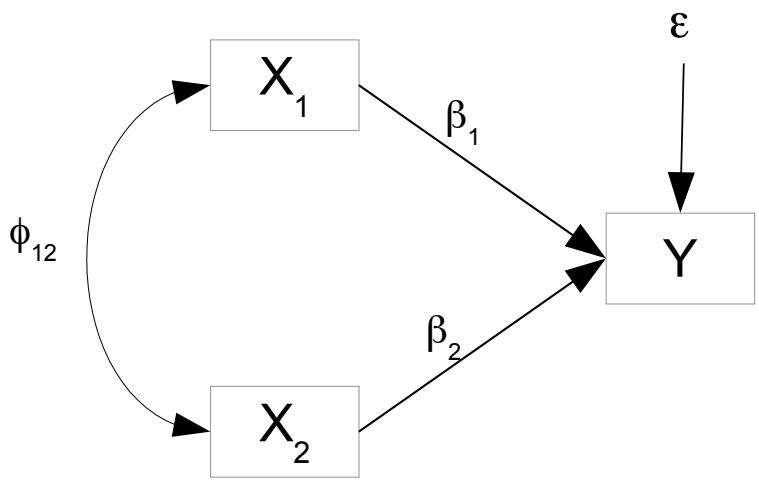
[3]The word "analysis" means splitting into parts, so this is literally analysis of variance.

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

Figure 3 shows a path diagram for this model. The explanatory and response variables are all observed, so they are enclosed in boxes. The double-headed curved arrow between the explanatory variables represents a possibly non-zero covariance. This covariance might arises from interesting and important processes including common influences on the X variables, but those processes are not part of the model. Curved double-headed arrows represent *unanalyzed* covariances between explanatory variables.

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

Figure 3: Unconditional multiple regression



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

$$\begin{aligned} Cov(X_{i,1}, Y_i) &= Cov(X_{i,1}, \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \epsilon_i) \\ &= Cov(X_{i,1}, \beta_1 X_{i,1} + \beta_2 X_{i,2} + \epsilon_i) \\ &= \beta_1 Cov(X_{i,1}, X_{i,1}) + \beta_2 Cov(X_{i,1}, X_{i,2}) + Cov(X_{i,1}, \epsilon_i) \\ &= \beta_1 Var(X_{i,1}) + \beta_2 Cov(X_{i,1}, X_{i,2}) + 0 \\ &= \beta_1 \phi_{11} + \beta_2 \phi_{12} \end{aligned}$$

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

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

between X_1 and Y may be positive even though β_1 is negative.

All this is true of the unconditional relationship between X_1 and Y, but what if you "control" for X_2 by holding it constant at some fixed value? When the explanatory variables are all random, the relationship between X_1 and Y controlling for X_2 simply refers to a conditional distribution — the joint distribution of X_1 and Y given $X_2 = x_2$. In this case the regression equation is

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

The constant is simply absorbed into the intercept. It's a little strange in that the intercept is potentially different for $i = 1, \ldots, n$, but that doesn't affect the covariance. Following the calculations in Example 0.3.1, the conditional covariance between $X_{i,1}$ and Y_i is $\beta_1\phi_{11}$. Thus to test whether X_1 is connected to Y controlling for X_2 (or correcting for it, or allowing for it or some such term), it is appropriate to test $H_0 : \beta_1 = 0$. If the null hypothesis is rejected, the sign of the estimated regression coefficient guides your conclusion as to whether the conditional relationship is positive or negative. These considerations extend immediately to multiple regression.

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

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

The explanatory variables contribute to the variance of the response individually through their variances and squared regression coefficients, and also jointly through their regression coefficients and their covariance. This joint effect is not an interaction in the ordinary sense of the term; the model of Example 0.3.2 has no product term. The null hypothesis $H_0 : \beta_1 = 0$ means that X_1 does not contribute at all to the variance of Y, either directly or through its covariance with X_2.

Estimation

Here is some useful terminology, repeated from Appendix A.

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 simple (one explanatory variable) regression model of Example 0.3.1, the moments are the elements of the mean vector $\boldsymbol{\mu} = E\begin{pmatrix} X_i \\ Y_i \end{pmatrix}$, and the unique elements of the covariance matrix $\boldsymbol{\Sigma} = cov\begin{pmatrix} X_i \\ Y_i \end{pmatrix}$. The moments structure equations are

$$\begin{aligned} \mu_1 &= \mu_x \\ \mu_2 &= \beta_0 + \beta_1\mu_x \\ \sigma_{1,1} &= \phi \\ \sigma_{1,2} &= \beta_1\phi \\ \sigma_{2,2} &= \beta_1^2\phi + \psi. \end{aligned} \tag{12}$$

In this model, the parameters are μ_x, ϕ, β_0, β_1, ψ, and also the unknown distribution functions of X_i and ϵ_i. Our interest is in the Greek-letter parameters, especially β_0 and β_1. Method of Moments estimates (See Section A.6.2 in Appendix A) can be obtained by solving the moment structure equations (12) for the unknown parameters and putting hats on the result. The moment structure equations form a system of 5 equations in five unknowns, and may be readily be solved to yield

$$\begin{aligned} \beta_0 &= \mu_2 - \frac{\sigma_{1,2}}{\sigma_{1,1}}\mu_1 \\ \mu_x &= \mu_1 \\ \phi &= \sigma_{1,1} \\ \beta_1 &= \frac{\sigma_{1,2}}{\sigma_{1,1}} \\ \psi &= \sigma_{2,2} - \frac{\sigma_{1,2}^2}{\sigma_{1,1}}. \end{aligned} \tag{13}$$

Thus, even though the distributions of X_i and ϵ_i are unknown, we have nice consistent estimators of the interesting part of the unknown parameter. Putting hats on the parameters in Expression 13,

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

It is very standard to assume that X_i and ϵ_i are normally distributed. In this case, the existence of the solution (13) tells us that the parameters of the normal version of this regression model stand in a one-to-one-relationship with the mean and covariance matrix

of the bivariate normal distribution posessed by the observable data. In fact, the two sets of parameter values are 100% equivalent; they are just different ways of expressing the same thing. For some purposes, the parameterization represented by the regression model may be more informative.

Furthermore, the Invariance Principle of maximum likelihood estimation (see Section A.6.3 in Appendix A) says that the MLE of a one-to-one function is just that function of the MLE. So, the Method of Moments estimates are also the Maximum Likelihood estimates in this case. Recognizing the formula for $\widehat{\beta}_1$ as a special case of Expression 4 on Page 14 (from the centered multiple regression model), we see that $\widehat{\beta}_1$ is also the ordinay least-squares estimate.

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

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

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

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

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

Example 0.3.3 *Multivariate Regression*

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

$$\mathbf{Y}_i = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1^\top \mathbf{X}_i + \boldsymbol{\epsilon}_i \tag{14}$$

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.

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

$\boldsymbol{\beta}_0$ 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.

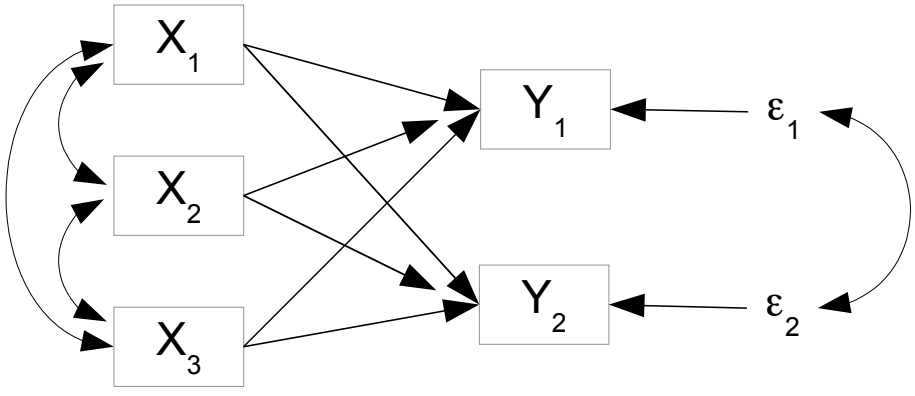
$\boldsymbol{\beta}_1$ is a $p \times q$ matrix of unknown constants. These are the regression coefficients, with one row for each explanatory variable and one column for each response variable.

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

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

Figure 4 depicts a model with three explanatory variables and two response variables. The explanatory and response variables are all observed, so they are enclosed in boxes. Double-headed curved arrows between the explanatory variable represent possible non-zero covariances. The straight arrows from the explanatory to response variables represent direct influence, or at least that we are interested in predicting y from x rather than the other way around. There is a regression coefficient $\beta_{j,k}$ on each arrow. The error terms ϵ_1 and ϵ_2 represent all other influences on Y_1 and Y_2. Since there could be common influences (omitted variables that affect both Y_1 and Y_2), the error terms are assumed to be correlated. This is the reason for the curved double-headed arrow joining ϵ_1 and ϵ_2.

Figure 4: Multivariate multiple regression



There is one regression equation for each response variable. In scalar form, the model

equations are

$$\begin{aligned} Y_{i,1} &= \beta_{0,1} + \beta_{1,1}X_{i,1} + \beta_{2,1}X_{i,2} + \beta_{3,1}X_{i,3} + \epsilon_{i,1} \\ Y_{i,2} &= \beta_{0.2} + \beta_{1,2}X_{i,1} + \beta_{2,2}X_{i,2} + \beta_{3,2}X_{i,3} + \epsilon_{i,2}. \end{aligned}$$

In matrix form,

$$\begin{array}{ccccccccc} \mathbf{Y}_i & = & \boldsymbol{\beta}_0 & + & \boldsymbol{\beta}_1^\top & & \mathbf{X}_i & + & \boldsymbol{\epsilon}_i \end{array}$$

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

Returning to the general case of Example 0.3.3, the observable data are the random vectors $\mathbf{D}_i = \left(\frac{\mathbf{X}_i}{\mathbf{Y}_i}\right)$, for $i = 1, \ldots, n$. The notation indicates that $\mathbf{D}_i$ is a partitioned random vector, with $\mathbf{X}_i$ stacked directly on top of $\mathbf{Y}_i$. Using the notation $E(\mathbf{D}_i) = \boldsymbol{\mu}$ and $cov(\mathbf{D}_i) = \boldsymbol{\Sigma}$, one may write $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ as partitioned matrices (matrices of matrices).

$$\boldsymbol{\mu} = \left(\frac{E(\mathbf{X}_i)}{E(\mathbf{Y}_i)}\right) = \left(\frac{\boldsymbol{\mu}_1}{\boldsymbol{\mu}_2}\right)$$

and

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

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

$$\begin{aligned} \boldsymbol{\Sigma}_{12} &= cov(\mathbf{X}_i, \mathbf{Y}_i) \\ &= cov\left(\mathbf{X}_i \ , \ \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1^\top \mathbf{X}_i + \boldsymbol{\epsilon}_i\right) \\ &= cov\left(\mathbf{X}_i \ , \ \boldsymbol{\beta}_1^\top \mathbf{X}_i + \boldsymbol{\epsilon}_i\right) \\ &= cov\left(\mathbf{X}_i, \mathbf{X}_i\right)\boldsymbol{\beta}_1 + cov(\mathbf{X}_i, \boldsymbol{\epsilon}_i) \\ &= cov(\mathbf{X}_i)\boldsymbol{\beta}_1 + \mathbf{0} \\ &= \boldsymbol{\Phi}\boldsymbol{\beta}_1 \end{aligned}$$

Thus, the moment structure equations are

$$\begin{aligned} \boldsymbol{\mu}_1 &= \boldsymbol{\mu}_x \\ \boldsymbol{\mu}_2 &= \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1^\top \boldsymbol{\mu}_x \\ \boldsymbol{\Sigma}_{11} &= \boldsymbol{\Phi} \\ \boldsymbol{\Sigma}_{12} &= \boldsymbol{\Phi}\boldsymbol{\beta}_1 \\ \boldsymbol{\Sigma}_{22} &= \boldsymbol{\beta}_1^\top \boldsymbol{\Phi}\boldsymbol{\beta}_1 + \boldsymbol{\Psi}. \end{aligned} \tag{15}$$

Solving for the parameter matrices is routine.

$$\begin{aligned} \boldsymbol{\beta}_0 &= \boldsymbol{\mu}_2 - \boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_x &= \boldsymbol{\mu}_1 \\ \boldsymbol{\Phi} &= \boldsymbol{\Sigma}_{11} \\ \boldsymbol{\beta}_1 &= \boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Psi} &= \boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{12}^{\top}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12} \end{aligned} \tag{16}$$

As in the univariate case, the Method of Moments estimates are obtained by putting hats on all the parameters in Expression (16). If the distributions of $\mathbf{X}_i$ and $\boldsymbol{\epsilon}_i$ are multivariate normal, the Invariance Principle reveals that Method of Moments estimates are also the maximum likelihood estimates.

Recall that in the proof of consistency for ordinary least squares with random explanatory variables, we centered the explanatory variables and obtained Formula 4 on Page 14: $\widehat{\boldsymbol{\beta}}_n = \widehat{\boldsymbol{\Sigma}}_x^{-1}\widehat{\boldsymbol{\Sigma}}_{xy}$. Compare this to the estimate of the slopes obtained from the solution (16) above: $\widehat{\boldsymbol{\beta}}_1 = \widehat{\boldsymbol{\Sigma}}_{11}^{-1}\widehat{\boldsymbol{\Sigma}}_{12}$. The formulas are almost the same. $\widehat{\boldsymbol{\Sigma}}_{11} = \widehat{\boldsymbol{\Sigma}}_x$, the sample variance-covariance matrix of the explanatory variables. $\widehat{\boldsymbol{\Sigma}}_{12}$ and $\widehat{\boldsymbol{\Sigma}}_{xy}$ are both matrices of sample covariances between explanatory and response variables, except that $\widehat{\boldsymbol{\Sigma}}_{12}$ is $p \times q$ while $\widehat{\boldsymbol{\Sigma}}_{xy}$ is $p \times 1$. $\widehat{\boldsymbol{\Sigma}}_{12}$ has one column for each response variable. So, in addition to being a method of moments estimate and a maximum likelihood estimate under normality $\widehat{\boldsymbol{\beta}}_1$ is a $p \times q$ matrix of least-squares estimates,

0.4 Omitted Variables

Some very serious problems can 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 typical applications can be misleading.

The trouble is not the explanatory variables are random. As we saw in Section 0.2, that's fine. But when 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, things can get ugly. In this section, we will see how omitting important explanatory variables from a regression equation can cause the error term to be correlated with the explanatory variables that remain, and how that can produce incorrect results.

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

$$Y_i = \beta_0 + \beta_1 X_{i,1} + \epsilon_i, \tag{17}$$

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

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

It could be argued that a *conditional* regression model (the usual model in which the explanatory variable values are fixed constants) is just a convenient way to represent dependence between X and Y by specifying a generic, more or less reasonable conditional distribution for Y given $X = x$. In this case, the correlation-causation issue can be set aside, and taken up when it is time to interpret the results. But if the explanatory variables are explicitly random, it is harder to avoid the obvious. In the simple regression model (17), the random variable Y_i is a function of the random variables X_i and ϵ_i. It is being directly produced by them. If this is taken seriously as a *scientific* model as well as a statistical model[7], it is inescapably causal; it is a model of what affects what. That's why the straight arrows in path diagrams are directional. The issue of whether X is influencing Y, or Y is influencing X or both is a modelling issue that will mostly be decided based on subject-matter theory.

It is natural to ask whether the data can be used to decide which way the arrows should be pointing. The answer is usually no, but it can be yes with certain other restrictions on the model. We will return to this issue later in the book. In the meantime, regression models with random explanatory variables, like the general structural equation models that are their extensions, will be recognized as causal models.

Again, Equation (17) says that X_i is influencing Y_i. All other influences are represented ϵ_i. It is common practice to assume that $X_{i,1}$ and ϵ_i are independent, or at least uncorrelated. But that does not mean the assumption can be justified in practice. Prepare yourself for a dose of reality.

Example 0.4.1 *Omitted Explanatory Variables*

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

$$Y_i = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \beta_3 X_{i,3} + \epsilon_i, \tag{18}$$

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

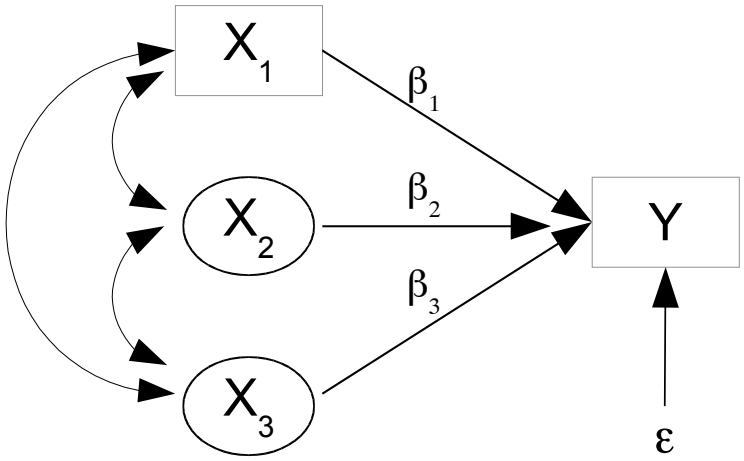
$$E\begin{pmatrix} X_{i,1} \\ X_{i,2} \\ X_{i,3} \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{pmatrix} \quad \text{and} \quad cov\begin{pmatrix} X_{i,1} \\ X_{i,2} \\ X_{i,3} \end{pmatrix} = \begin{pmatrix} \phi_{11} & \phi_{12} & \phi_{13} \\ & \phi_{22} & \phi_{23} \\ & & \phi_{33}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

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

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

Figure 5 shows a path diagram of this situation. Because the explanatory variables $X_{i,2}$ and $X_{i,3}$ are not observable, they are *latent* variables, and so they are enclsed by ovals in the path diagram. Their covariances with $X_{i,1}$ and each other are represented by two-headed curved arrows.

Figure 5: Omitted explanatory variables



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

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

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

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

$$\begin{aligned} Cov(X_{i,1}, \epsilon_i') &= Cov(X_{i,1}, \beta_2 X_{i,2} + \beta_3 X_{i,3} - \beta_2\mu_2 - \beta_3\mu_3 + \epsilon_i) \\ &= \beta_1 Cov(X_{i,1}, X_{i,2}) + \beta_3 Cov(X_{i,1}, X_{i,3}) + 0 \\ &= \beta_2\phi_{12} + \beta_3\phi_{13} \end{aligned} \tag{19}$$

So, when explanatory variables are omitted from the regression equation and those explanatory variables have non-zero covariance with variables that *are* in the equation, the

result is non-zero covariance between the error term and the explanatory variables in the equation[8].

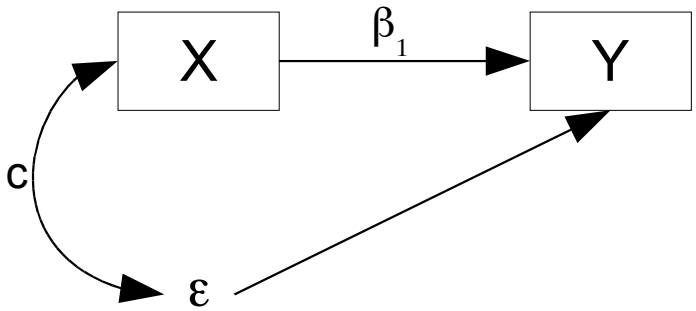
Response variables are almost always affected by more than one explanatory variable, and in observational data, explanatory variables usually have non-zero covariances with one another. So, the most realistic model for a regression with just one explanatory variable should include a covariance between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

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

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \tag{20}$$

where $E(X_i) = \mu_x$, $Var(X_i) = \sigma_x^2$, $E(\epsilon_i) = 0$, $Var(\epsilon_i) = \sigma_\epsilon^2$, and $Cov(X_i, \epsilon_i) = c$. A path diagram of the true model is given in Figure 6. The covariance c is indicated on the curved arrow connecting the explanatory variable and the error term. Consider a

Figure 6: Omitted explanatory variables have been swallowed by ϵ



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 textbooks and implemented in countless computer programs and even 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

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

obvious, but this model implies independence of the explanatory variable and the error term. It is a conditional model, and the distribution of the error terms is *the same* for every fixed set of values $x_1, \ldots, x_n$. Using a loose but understandable notation for densities and conditional densities,

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

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

Surprisingly, this does not depend on the assumption of any particular distribution for the error terms. All you need is the stipulation $E(\epsilon_i) = 0$ in a fixed-x regression model. It's worth doing this in generality, so consider the multivariate multiple regression model of Example 0.3.3 on page 23:

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

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

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

and

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

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

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

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

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

Zero covariance between error terms and explanatory variables means that *any potential explanatory variable not in the model must have zero covariance with the explanatory variables that are in the model*. Of course this is almost never realistic without random assignment to experimental conditions, so that almost every application of regression methods to non-experimental data makes an assumption that cannot be justified. Now we will see the consequences.

For a simple regression, both ordinary least squares and an unconditional regression model like the true model on Page 29 with $c = 0$ lead to the same standard formula:

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

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. 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,

$$\begin{aligned}Cov(X, Y) &= Cov(X_i, \beta_0 + \beta_1 X_i + \epsilon_i) \\ &= \beta_1 Cov(X_i, X_i) + Cov(X_i, \epsilon_i) \\ &= \beta_1 \sigma_x^2 + c\end{aligned}$$

So by continuity,

$$\widehat{\beta}_1 = \frac{\widehat{\sigma}_{x,y}}{\widehat{\sigma}_x^2} \xrightarrow{a.s.} \beta_1 + \frac{c}{\sigma_x^2}. \tag{21}$$

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

What's going on here is that the calculation of $\widehat{\beta}_1$ is based on a model that is *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 non-zero covariance with variables that are in the model, the regression coefficients are inconsistent*. In other words, with more and more data they do not approach the right answer. Instead, they get closer and closer to a specific wrong answer.

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

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

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

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

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

Also, the convergence of $\widehat{\beta}_1$ to the wrong answer in (21) may be misleading, but it does not necessarily yield the wrong conclusion. In much of the social and biological sciences, the theories are not detailed and sophisticated enough to make predictions about the actual values of regression coefficients, just whether they should be positive, negative or zero. So, if the variable being tested and the omitted variables are pulling in the same direction (that is, if β_1 and c in Model (20) on Page 29 are either both positive or both

negative), the study will come to the "right" conclusion. The trouble is that you can't tell, because you don't even know what the omitted variables are. All you can do is hope, and that's not a recipe for good science.

Trying to fit the true model We have seen that serious trouble arises from adopting a mis-specified model with $c = Cov(X_i, \epsilon_i) = 0$, when in fact because of omitted variables, $c \neq 0$. It is natural, therefore, to attempt estimation and inference for the true model $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$ (see Page 29) in the case where $c = Cov(X_i, \epsilon_i)$ need not equal zero. For simplicity, assume that X_i and ϵ_i have a bivariate normal distribution, so that the observable data pairs (X_i, Y_i) for $i = 1, \ldots, n$ are a random sample from a bivariate normal distribution with mean vector $\boldsymbol{\mu}$ and variance-covariance matrix $\boldsymbol{\Sigma}$.

It is straightforward to calculate $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ from the equation and assumptions of the true model (20). The result is

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} = E \begin{pmatrix} X_i \\ Y_i \end{pmatrix} = \begin{pmatrix} \mu_x \\ \beta_0 + \beta_1 \mu_x \end{pmatrix} \tag{22}$$

and

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{pmatrix} = cov \begin{pmatrix} X_i \\ Y_i \end{pmatrix} = \begin{pmatrix} \sigma_x^2 & \beta_1\sigma_x^2 + c \\ \beta_1\sigma_x^2 + c & \beta_1^2\sigma_x^2 + 2\beta_1 c + \sigma_\epsilon^2 \end{pmatrix}. \tag{23}$$

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

Our primary interest is in β_1. Because the data pairs (X_i, Y_i) come from a bivariate normal distribution, all you can ever learn from the data are the approximate values of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. With larger and larger samples, all you get is better and better approximations of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. That's all there is to know. But even if you knew $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ exactly, could you know β_1? Formulas (22) and (23) yield a system of five equations in six unknown parameters.

$$\begin{aligned} \mu_1 &= \mu_x \\ \mu_2 &= \beta_0 + \beta_1\mu_x \\ \sigma_{11} &= \sigma_x^2 \\ \sigma_{12} &= \beta_1\sigma_x^2 + c \\ \sigma_{22} &= \beta_1^2\sigma_x^2 + 2\beta_1 c + \sigma_\epsilon^2 \end{aligned} \tag{24}$$

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

To see this in detail, substitute μ_1 for μ_x and σ_{11} for σ_x^2 in (24), obtaining

$$\begin{aligned} \mu_2 &= \beta_0 + \beta_1\mu_1 \\ \sigma_{12} &= \beta_1\sigma_{11} + c \\ \sigma_{22} &= \beta_1^2\sigma_{11} + 2\beta_1 c + \sigma_\epsilon^2 \end{aligned} \tag{25}$$

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

- All the equations in (24) are satisfied,
- The values of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ remain constant, and
- The distribution of $(X_i, Y_i)^\top$ is $N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

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

$$\begin{aligned} \mu_x &= \mu_1 \\ \sigma_x^2 &= \sigma_{11} \\ \beta_0 &= \mu_2 - \mu_1 \left(\frac{\sigma_{12} - c}{\sigma_{11}} \right) \\ \beta_1 &= \frac{\sigma_{12} - c}{\sigma_{11}} \\ \sigma_\epsilon^2 &= \sigma_{22} + \frac{c^2 - \sigma_{12}^2}{\sigma_{11}} \end{aligned} \tag{26}$$

The parameters μ_x and σ_x^2 are constant functions of c, while β_0 and β_1 are linear functions, and σ_ϵ^2 is a quadratic function. The equations (26) define a one-dimensional surface in the six-dimensional parameter space, a kind of curved thread in $\mathbb{R}^6$. Moving c from $-\infty$ to ∞ traces out the points on the thread. Importantly, as c ranges from $-\infty$ to $+\infty$ the regression coefficient β_1 ranges from $+\infty$ to $-\infty$. This means that β_1 might be positive, it might be negative, or it might be zero. But you really can't tell, because all real values of β_1 on the surface yield the same population mean and population variance-covariance matrix, and hence the same distribution of the sample data. There is no way to distinguish between the possible values of β_1 based on sample data.

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

must be greater than zero. For points on the thread, the first condition is

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

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

$$\begin{aligned} & \sigma_\epsilon^2 > 0 \\ \Leftrightarrow\ & \sigma_{22} + \frac{c^2 - \sigma_{12}^2}{\sigma_{11}} > 0 \\ \Leftrightarrow\ & \sigma_{11}\sigma_{22} + c^2 - \sigma_{12}^2 > 0 \\ \Leftrightarrow\ & |\boldsymbol{\Sigma}| + c^2 > 0 \end{aligned}$$

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

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

What would happen if you tried to estimate the parameters by maximum likelihood? For every $\boldsymbol{\mu} \in \mathbb{R}^2$ and every 2×2 symmetric positive definite $\boldsymbol{\Sigma}$, there is a surface (thread) in $\mathbb{R}^6$ defined by (26). This includes $(\widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Sigma}})$. On that particular thread, the likelihood is highest. Picture a surface with a curvy ridge at the top. The surface has infinitely many maxima, all at the same height, forming a connected set. If you take partial derivatives of the log likelihood and set all six of them equal to zero, there will be infinitely many solutions. If you do numerical maximum likelihood, good software will find a point on the ridge, stop, detect that the surface is not fully concave down there, and complain. Less sophisticated software will just find a point on the ridge, and stop. The stopping place, that is, the maximum likelihood estimate, depends entirely on where the numerical search starts.

To summarize, if explanatory variables are omitted from a regression equation and those variables have non-zero covariance c with explanatory variables that are *not* omitted, the result is non-zero covariance between explanatory variables and the error term. And, if there is a non-zero covariance between the error term an an explanatory variable in a

regression equation, the false assumption that $c = 0$ can easily lead to false results. But allowing c to be non-zero means that infinitely many parameter estimates will be equally plausible, given any set of sample data. In particular, no set of data will be able to provide a basis for deciding whether regression coefficients are positive, negative or zero. The problem is fatal if all you have is X_i and Y_i.

The trouble here is lack of parameter identifiability. If a parameter is a function of the distribution of the observable data, it is said to be *identifiable*. The idea is that the parameter is potentially knowable if you knew the distribution of the observable data. If the parameter is not knowable based on the data, they naturally there will be trouble with estimation and inference. Parameter identifiability is a central theme of this book, and will be taken up again in Section 0.9 on Page 58.

0.5 Instrumental Variables

The method of instrumental variables was introduced by the economist Phillip Wright in the appendix a 1928 book *The Tariff on Animal and Vegetable Oils* [20]. Phillip Wright was 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 [16].

An instrumental variable for an explanatory is a variable that is correlated with that explanatory variable, but is not correlated with any error terms or other explanatory variables, and has no direct connection to the response variable. In Econometrics, the instrumental variable usually *influences* the explanatory variable. An instrumental variable is usually not the main focus of attention; it's just a tool.

Example 0.5.1 *Credit Card Debt*

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

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

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

Definition 0.5.1 *An instrumental variable for an explanatory variable is another random variable that has non-zero covariance with the explanatory variable, and* no direct connection with any other variable in the model.

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

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

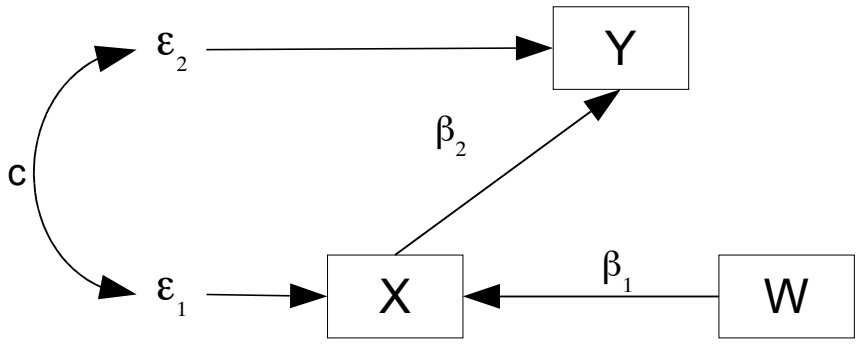
- W_i is median price of resale home in agent i's district.
- X_i is annual income of real estate agent i.
- Y_i is agent i's credit card debt.

The model equations are

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

and Figure 7 shows the path diagram. The main interest is in β_2, the link between income and credit card debt. The covariance between ϵ_1 and ϵ_2 represents all the omitted variables

Figure 7: W is median price of resale home, X is income, Y is credit card debt



that affect income and credit card debt.

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

$$\boldsymbol{\Sigma} = \begin{array}{c|ccc} & W & X & Y \\ \hline W & \sigma_w^2 & \beta_1\sigma_w^2 & \beta_1\beta_2\sigma_w^2 \\ X & \cdot & \beta_1^2\sigma_w^2 + \sigma_1^2 & \beta_2(\beta_1^2\sigma_w^2 + \sigma_1^2) + c \\ Y & \cdot & \cdot & \beta_1^2\beta_2^2\sigma_w^2 + \beta_2^2\sigma_1^2 + 2\beta_2 c + \sigma_2^2 \end{array} \tag{27}$$

The lower triangle of the covariance matrix is omitted to make it less cluttered. The notation in (27) is self-explanatory except possibly for $Var(\epsilon_{i1}) = \sigma_1^2$ and $Var(\epsilon_{i2}) = \sigma_2^2$. It is immediately apparent that the critical parameter β_2 can be recovered from $\boldsymbol{\Sigma}$ by $\beta_2 = \frac{\sigma_{13}}{\sigma_{12}}$, provided $\beta_1 \neq 0$. A nice Method of Moments estimator in terms of the sample covariances is $\widehat{\beta}_2 = \frac{\widehat{\sigma}_{13}}{\widehat{\sigma}_{12}}$.

The requirement that $\beta_1 \neq 0$ can be verified, by testing $H_0 : \sigma_{12} = 0$ with an elementary test of the correlation between housing prices and income. We expect no problem, because W is a good instrumental variable. Median resale price certainly should be related to the income of real estate agents, and furthermore the relationship is practically guaranteed to be positive. This is a feature of good a instrumental variable. Its relationship to the explanatory variable should be clear, and so obvious that it is hardly worth investigating. The usefulness of the instrumental variable is in the light it casts on relationships that are not so obvious.

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

To test the central null hypothesis $H_0 : \beta_2 = 0$, fancy software is not required. Since we have concluded with high confidence that $\beta_1 > 0$, the covariance σ_{13} equals zero if and only if $\beta_2 = 0$, and the sign of σ_{13} is the same as the sign of β_2. So it is necessary only to test the correlation between housing price and real estate agents' credit card debt. Under the normal assumption, the usual test is exact and a large sample is not required. If the normal assumption is worrisome, the non-parametric test associated with the Spearman rank correlation coefficient is a permutation test carried out on ranks, and an exact small-sample p-value is available even though some software produces a large-sample approximation by default.

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

Measurement Error All models are inexact representations of reality, but I must admit that the model in Figure 7 is seriously wrong. Our interest is in how *true* income affects *true* credit card debt. But these variables are not observed. What we have in the data file are *reported* income and *reported* credit card debt. For various reasons that the reader can easily supply, the truth and what people report about financial details are not the same thing. When we record median price of a resale home, that's unlikely to be perfectly accurate either. As we will see later in this chapter, measurement error in

the explanatory variables presents serious problems for regression analysis in general. We will also see that instrumental variables can help with measurement error as well as with omitted variables, but first it is helpful to introduce the topic of measurement error in an organized way.

0.6 The Idea of Measurement Error

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

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

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

A simple additive model for measurement error

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

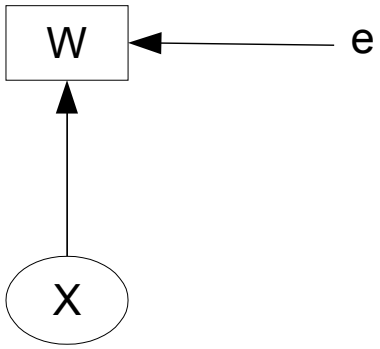
Here, we will put classification error aside because it is technically difficult, and focus on a very simple form of measurement error that applies to continuous variables. There

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,

$$W = X + e \tag{28}$$

Let's say $E(X) = \mu_x$, $E(e) = 0$, $Var(X) = \sigma_x^2$, $Var(e) = \sigma_e^2$, and $Cov(X, e) = 0$. Figure 8 is a path diagram of this model.

Figure 8: Additive Measurement Error



Because X and e are uncorrelated,

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

Variance is an index of unit-to unit variation in a measurement. The simple calculation above reveals that variation in the observable variable has two sources: variation in the actual quantity of interest, and variation in the magnitude of the random shocks that create error in measurement. To judge the quality of a measurement W, it is important to assess how much of its variance comes from variation in the true quantity of interest, and how much comes from random noise.

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

[11] Psychometric theory is the statistical theory of psychological measurement. The bible of psychometric theory is Lord and Novick's (1968) classic *Statistical theories of mental test scores* [10]. It is not too surprising that measurement error would be acknowledged and studied by psychologists. A large sector of psychological research employs "measures" of hypothetical constructs like neuroticism or intelligence (mostly paper-and-pencil tests), but no sensible person would claim that true value of such a trait is exactly the score on the test. It's true there is a famous quote "Intelligence is whatever an intelligence test measures." I have tried unsuccessfully to track down the source of this quote, and I now suspect that it is just an illustration of a philosophic viewpoint called Logical Positivism (which is how I first heard it), and not a serious statement about intelligence measurement.

[12] Reliability has a completely unrelated meaning in survival analysis and statistical quality control.

"observed score" is W. The reliability of the measurement W is

$$\begin{aligned}\rho^2 &= \left(\frac{Cov(X,W)}{SD(X)SD(W)}\right)^2 \\ &= \left(\frac{\sigma_x^2}{\sqrt{\sigma_x^2}\sqrt{\sigma_x^2+\sigma_e^2}}\right)^2 \\ &= \frac{\sigma_x^4}{\sigma_x^2(\sigma_x^2+\sigma_e^2)} \\ &= \frac{\sigma_x^2}{\sigma_x^2+\sigma_e^2}. \end{aligned} \tag{29}$$

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

A reliability of one means there is no measurement error at all, while a reliability of zero means the measurement is pure noise. In the social sciences, reliabilities above 0.9 could be called excellent, from 0.8 to 0.9 good, and from 0.7 to 0.8 acceptable. Frequently, responses to single questions have reliabilities that are much less than this. To see why reliability depends on the number of questions that measure the latent variable, see Exercise 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. On rare occasions and perhaps with great expense, it may be possible to obtain perfect or near-perfect measurements on a subset of the sample; the term *gold standard* is sometimes applied to such measurements. In that case, the reliability of the usual measurement can be estimated by a squared sample correlation between the usual measurement and the gold standard measurement. But even measurements that are called gold standard are seldom truly free of measurement error. Consequently, reliabilities that are estimated by correlating imperfect gold standards and ordinary measurements are biased downward: See Exercise 4 at the end of this section. It is clear that another approach is needed.

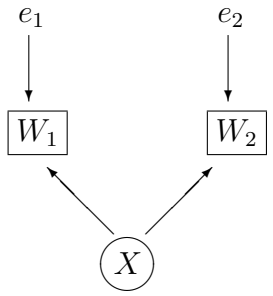
Test-retest reliability Suppose that it is possible to make the measurement of W twice, in such a way that the errors of measurement are independent on the two occasions. We have

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

where $E(X) = \mu_x$, $Var(X) = \sigma_x^2$, $E(e_1) = E(e_2) = 0$, $Var(e_1) = Var(e_2) = \sigma_e^2$, and X, e_1 and e_2 are all independent. Because $Var(e_1) = Var(e_2)$, W_1 and W_2 are called *equivalent*

[13] It's like the proportion of variance in the response variable explained by a regression, except that here the explanatory variable is the latent true score. Compare Expression (11) on Page 19.

Figure 9: Two independent measurements of a latent variable



measurements. That is, they are contaminated by error to the same degree. Figure 9 is a path diagram of this model.

It turns out that the correlation between W_1 and W_2 is exactly equal to the reliability, and this opens the door to reasonable methods of estimation. The calculation (like many in this book) is greatly simplified by using the rule for covariances of linear combinations (1) on Page 9.

$$\begin{aligned} Corr(W_1, W_2) &= \frac{Cov(W_1, W_2)}{SD(W_1)SD(W_2)} \\ &= \frac{Cov(X + e_1, X + e_2)}{\sqrt{\sigma_x^2 + \sigma_e^2}\sqrt{\sigma_x^2 + \sigma_e^2}} \\ &= \frac{Cov(X, X) + Cov(X, e_2) + Cov(e_1, X) + Cov(E_1, e_2))}{\sigma_x^2 + \sigma_e^2} \\ &= \frac{Var(X) + 0 + 0 + 0}{\sigma_x^2 + \sigma_e^2} \\ &= \frac{\sigma_x^2}{\sigma_x^2 + \sigma_e^2}, \end{aligned} \tag{30}$$

which is the reliability.

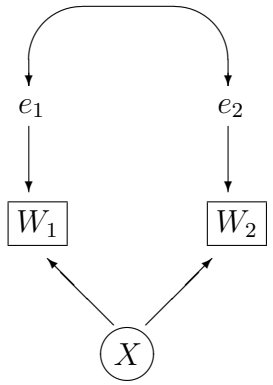
The calculation above is the basis of *test-retest reliability*[14], in which the reliability of a measurement such as an educational or psychological test is estimated by the sample correlation between two independent administrations of the test. That is, the test is given twice to the same sample of individuals, ideally with a short enough time between tests

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

so that the trait does not really change, but long enough apart so they forget how they answered the first time.

Correlated measurement error Suppose participants remembered their wrong answers or lucky guesses from the first time they took a test, and mostly gave the same answer the second time. The result would be a positive correlation between the measurement errors e_1 and e_2. Omitted variables (see Section 0.4) like level of test anxiety for educational tests or desire to make a favourable impression for attitude questionnaires can also produce a positive covariance between errors of measurement. Whatever the source, positive covariance between e_1 and e_2 is an additional source of positive covariance between W_1 and W_2 that does *not* come from the latent variable X being measured. The result is an inflated estimate of reliability and an unduly rosy picture of the quality of measurement. Figure 10 shows this situation.

Figure 10: Correlated Measurement Error



We will return more than once to the issue of correlated errors of measurement. For now, just notice how careful planning of the data collection (in this case, the time lag between the two administrations of the test) can eliminate or at least reduce the correlation between errors of measurement. In general, the best way to take care of correlated measurement error is with good research design[15].

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

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

[15]Indeed, one could argue that most principles of good research design are methods for minimizing the variance and covariance of measurement errors.

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

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

where the convergence follows from continuous mapping and the fact that sample variances and covariances are strongly consistent estimators of the corresponding population quantities; see Section A.5 in Appendix A. The conclusion is that 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.

0.7 Ignoring measurement error

Standard regression models make no provision at all for measurement error, so when such models are applied to real data, we are effectively ignoring any measurement error that may be present – pretending it's not there. This section will show that the result can be a real disaster, featuring incorrect estimates of regression parameters and Type I error probabilities approaching one as the sample size increases. Much of this material, including the history of the topic (warnings go back to at least 1936) can be found in a 2009 paper by Brunner and Austin [5].

Measurement error in the response variable

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

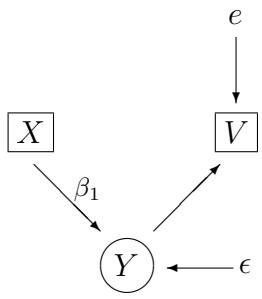
Example 0.7.1 *Measurement Error in Y*

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

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

where $Var(X_i) = \sigma_x^2$, $Var(e_i) = \sigma_e^2$, $Var(\epsilon_i) = \sigma_\epsilon^2$, and X_i, e_i, ϵ_i are all independent. Figure 11 is a path diagram of this model.

Figure 11: Measurement error in the response variable



In Example 0.7.1, the explanatory variable X_i is observable, but the response variable Y_i is latent. Instead of Y_i, we can see V_i, which is Y_i plus a piece of random noise, and also plus a constant ν that represents the difference between the expected value of the latent random variable and the expected value of its observable counterpart. This constant term could be called *measurement bias*. For example, if Y is true amount of exercise in minutes and V is reported exercise, the measurement bias ν is population mean exaggeration, in minutes.

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

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

Studying Mis-specified Models The "naive model" above is an example of a model that is *mis-specified*. That is, the model says that the data are being generated in a particular way, but this is not how the data are actually being produced. Generally speaking, correct models will usually yield better results than incorrect models, but it's not that simple. In reality, most statistical models are imperfect. The real question is how much any given imperfection really matters. As Box and Draper (1987, p. 424) put it, "Essentially all models are wrong, but some are useful." [4]

So, it is not enough to complain that a statistical model is incorrect, or unrealistic. To make the point convincingly, one must show that by being wrong in a particular way, the model can yield results that are misleading. To do this, it is necessary to have a specific *true model* in mind; typically the so-called true model is one that is obviously more believable than the model being challenged. Then, one can examine estimators or test statistics based on the mis-specified model, and see how they behave when the true model holds. We have already done this in Section 0.4 in connection with omitted variables; see Example 0.4.1 starting on Page 27.

Under the true model of Example 0.7.1 (measurement error in the response variable

only), we have $Cov(X_i, V_i) = \beta_1 \sigma_x^2$ and $Var(X_i) = \sigma_x^2$. Then,

$$\begin{aligned} \widehat{\beta}_1 &= \frac{\sum_{i=1}^n (X_i - \overline{X})(V_i - \overline{V})}{\sum_{i=1}^n (X_i - \overline{X})^2} \\ &= \frac{\widehat{\sigma}_{x,v}}{\widehat{\sigma}_x^2} \\ &\stackrel{a.s.}{\to} \frac{Cov(X_i, V_i)}{Var(X_i)} \\ &= \frac{\beta_1 \sigma_x^2}{\sigma_x^2} \\ &= \beta_1. \end{aligned} \tag{31}$$

Even when the model is mis-specified by assuming that the response variable is measured without error, the ordinary least squares estimate of the slope is consistent. There is a general lesson here about mis-specified models. Mis-specification (using the wrong model) is not always a problem; sometimes everything works out fine.

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

$$\begin{aligned} V_i &= \nu + Y_i + e_i \\ &= \nu + (\beta_0 + \beta_1 X_i + \epsilon_i) + e_i \\ &= (\nu + \beta_0) + \beta_1 X_i + (\epsilon_i + e_i) \\ &= \beta_0' + \beta_1 X_i + \epsilon_i' \end{aligned} \tag{32}$$

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

In this book and in standard statistical practice, there are many models where the response variable appears to be measured without error. But error-free measurement is a rarity at best, so these models should be viewed as re-parameterized versions of models that do acknowledge the reality of measurement error in the response variable. A critical feature of these re-parameterized models is that the measurement error is assumed independent of everything else in the model. When this fails, there is usually trouble.

Measurement error in the explanatory variables

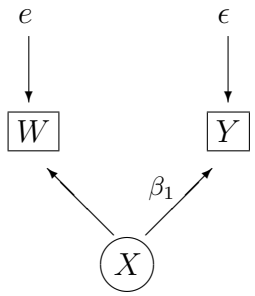
Example 0.7.2 *Measurement error in a single explanatory variable*

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

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

where $Var(X_i) = \sigma_x^2$, $Var(e_i) = \sigma_e^2$, $Var(\epsilon_i) = \sigma_\epsilon^2$, and X_i, e_i, ϵ_i are all independent. Figure 12 is a path diagram of the model.

Figure 12: Measurement error in the explanatory variable



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 0.7.2, the ordinary least squares estimate of β_1 is

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

Regardless of what model is correct, $\widehat{\sigma}_{w,y} \stackrel{a.s.}{\to} Cov(W,Y)$ and $\widehat{\sigma}_w^2 \stackrel{a.s.}{\to} Var(W)$[16], so that by the continuous mapping property of ordinary limits[17], $\widehat{\beta}_1 \stackrel{a.s.}{\to} \frac{Cov(W,Y)}{Var(W)}$.

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

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

Consequently,

$$\begin{aligned} \widehat{\beta}_1 &= \frac{\sum_{i=1}^n (W_i - \overline{W})(Y_i - \overline{Y})}{\sum_{i=1}^n (W_i - \overline{W})^2} \\ &= \frac{\widehat{\sigma}_{w,y}}{\widehat{\sigma}_w^2} \\ &\stackrel{a.s.}{\to} \frac{Cov(W,Y)}{Var(W)} \\ &= \beta_1 \left(\frac{\sigma_x^2}{\sigma_x^2 + \sigma_e^2} \right). \end{aligned} \tag{33}$$

[16]This is true because sample variances and covariances are strongly consistent estimators of the corresponding population quantities; see Section A.5 in Appendix A.

[17]Almost sure convergence acts like an ordinary limit, applying to all points in the underlying sample space, except possibly a set of probability zero. If you wanted to do this problem strictly in terms of convergence in probability, you could use the Weak Law of Large Numbers and then use Slutsky Lemma 7a of Appendix A.5.

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

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

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

Measurement error in more than one explanatory variable

In Example 0.7.2, we saw that measurement error in the explanatory variable causes the estimated regression coefficient $\widehat{\beta}_1$ to be biased toward zero as $n \to \infty$. Bias toward zero weakens the apparent relationship between X and Y; and if $\beta_1 = 0$, there is no asymptotic bias. So for the case of a single explanatory variable measured with error, the sample relationships still reflect population relationships, with the sample relationships being weaker because of inexact measurement. But this only holds for regression with a single explanatory variable. Measurement error causes a lot more trouble for multiple regression. In this example, there are two explanatory variables, both measured with error.

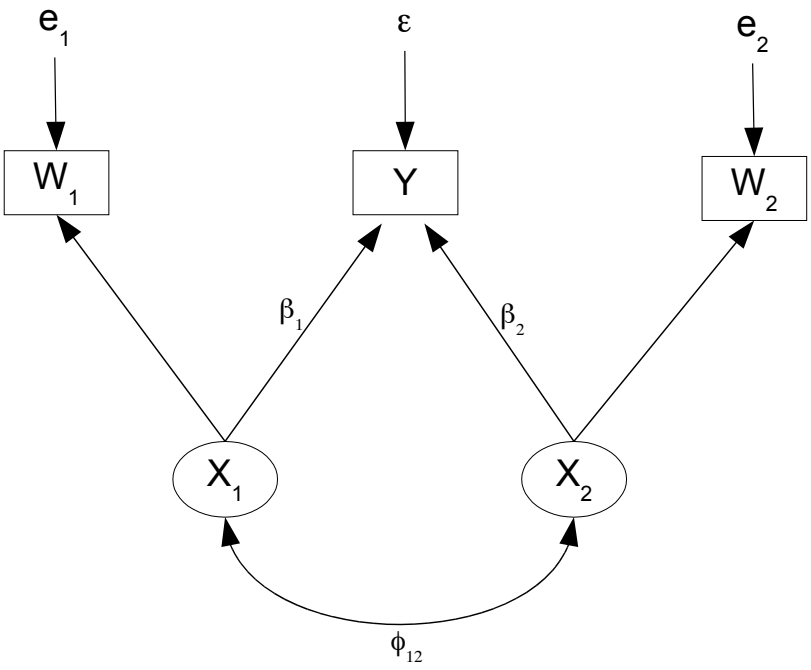
Example 0.7.3 *Measurement Error in Two Explanatory Variables*

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

Figure 13: Two explanatory variables measured with error



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

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

Figure 13 shows the path diagram.

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

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

An attractive feature of multiple regression is its ability to represent the relationship of one or more explanatory variables to the response variable, while *controlling for* other explanatory varables. In fact, this is the biggest appeal of multiple regression and similar methods for non-experimental data. In Example 0.7.3, our interest is in the relationship of X_2 to Y controlling for X_1. The main objective is to test $H_0 : \beta_2 = 0$, but we are also interested in the estimation of β_2.

The argument that follows illustrates a general way to see what happens as $n \to \infty$ for mis-specified (that is, incorrect) regression models. We have already seen special cases of it, three times. In Example 0.4.1 on omitted explanatory variables, the regression coefficient converged to the wrong target in Expression 21 on page 31. In Example 0.7.1 on measurement error in the response variable, the regression coefficient converged to the correct target in Expression 31 on page 46. In Example 0.7.2 on measurement error in a

single explanatory variable, the regression coefficient converged to the target multiplied by the reliability of the measurement, in Expression 33 on page 47.

Here is the recipe. Assume some "true" model for how the data are produced, and a mis-specified model corresponding to a natural way that people would analyze the data with a regression model. First, write the regression coefficients in terms of sample variances and covariances. The general answer is given on page 14: $\widehat{\boldsymbol{\beta}}_n = \widehat{\boldsymbol{\Sigma}}_x^{-1}\widehat{\boldsymbol{\Sigma}}_{xy}$. Then, because sample variances and covariances are consistent estimators of their population counterparts, we have the convergence $\widehat{\boldsymbol{\beta}}_n \stackrel{a.s.}{\to} \boldsymbol{\Sigma}_x^{-1}\boldsymbol{\Sigma}_{xy}$ from Page 15. This convergence follows from the formula for the least-squares estimator, and does not depend in any way on the correctness of the model. So, if you can derive $\boldsymbol{\Sigma}_x$ and $\boldsymbol{\Sigma}_{xy}$ under the true model, it is easy enough to calculate the large-sample target of the ordinary least squares estimates under the mis-specified model.

In the present application, there is just a minor notational issue. Under the naive model, the explanatory variables are called w instead of x. Adopting a notation that will be used throughout the book, denote one of the n vectors of observable data by $\mathbf{D}_i$. Here,

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

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

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

Then, we calculate the regression coefficients under the naive model.

$$\begin{aligned} \widehat{\boldsymbol{\beta}}_n &= \begin{pmatrix} \widehat{\beta}_1 \\ \widehat{\beta}_2 \end{pmatrix} \\ &= \widehat{\boldsymbol{\Sigma}}_w^{-1}\widehat{\boldsymbol{\Sigma}}_{wy} \\ &= \begin{pmatrix} \widehat{\sigma}_{1,1} & \widehat{\sigma}_{1,2} \\ \widehat{\sigma}_{1,2} & \widehat{\sigma}_{2,2} \end{pmatrix}^{-1} \begin{pmatrix} \widehat{\sigma}_{1,3} \\ \widehat{\sigma}_{2,3} \end{pmatrix} \\ &= \begin{pmatrix} \frac{\widehat{\sigma}_{22}\widehat{\sigma}_{13}-\widehat{\sigma}_{12}\widehat{\sigma}_{23}}{\widehat{\sigma}_{11}\widehat{\sigma}_{22}-\widehat{\sigma}_{12}^2} \\ \\ \frac{\widehat{\sigma}_{11}\widehat{\sigma}_{23}-\widehat{\sigma}_{12}\widehat{\sigma}_{13}}{\widehat{\sigma}_{11}\widehat{\sigma}_{22}-\widehat{\sigma}_{12}^2} \end{pmatrix}. \end{aligned} \tag{34}$$

Our primary interest is in the estimation of β_2. Because sample variances and covariances are strongly consistent estimators of the corresponding population quantities,

$$\widehat{\beta}_2 = \frac{\widehat{\sigma}_{11}\widehat{\sigma}_{23} - \widehat{\sigma}_{12}\widehat{\sigma}_{13}}{\widehat{\sigma}_{11}\widehat{\sigma}_{22} - \widehat{\sigma}_{12}^2} \stackrel{a.s.}{\to} \frac{\sigma_{11}\sigma_{23} - \sigma_{12}\sigma_{13}}{\sigma_{11}\sigma_{22} - \sigma_{12}^2}. \tag{35}$$

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

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

It will be non-zero provided at least one of

$$cov \begin{pmatrix} X_{i,1} \\ X_{i,2} \end{pmatrix} = \begin{pmatrix} \phi_{11} & \phi_{12} \\ \phi_{12} & \phi_{22} \end{pmatrix} \quad \text{and} \quad cov \begin{pmatrix} e_{i,1} \\ e_{i,2} \end{pmatrix} = \begin{pmatrix} \omega_1 & 0 \\ 0 & \omega_2 \end{pmatrix}$$

is positive definite – not a lot to ask.

The convergence of $\widehat{\beta}_2$ in Expression 35 applies regardless of what model is correct. To see what happens when the true model of Example 0.7.3 holds, we need to write the σ_{ij} quantities in terms of the parameters of the true model. A straightforward set of scalar variance-covariance calculations yields

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

Subsituting into expression 35 and simplifying[18], we obtain

$$\begin{aligned}
\widehat{\beta}_2 &= \frac{\widehat{\sigma}_{11}\widehat{\sigma}_{23} - \widehat{\sigma}_{12}\widehat{\sigma}_{13}}{\widehat{\sigma}_{11}\widehat{\sigma}_{22} - \widehat{\sigma}_{12}^2} \\
&\stackrel{a.s.}{\rightarrow} \frac{\sigma_{11}\sigma_{23} - \sigma_{12}\sigma_{13}}{\sigma_{11}\sigma_{22} - \sigma_{12}^2} \\
&= \frac{\beta_1\omega_1\phi_{12} + \beta_2(\omega_1\phi_{22} + \phi_{11}\phi_{22} - \phi_{12}^2)}{(\phi_{1,1}+\omega_1)(\phi_{2,2}+\omega_2) - \phi_{12}^2} \\
&= \beta_2 + \frac{\beta_1\omega_1\phi_{12} + \beta_2\omega_2(\phi_{11} - \omega_1)}{(\phi_{1,1}+\omega_1)(\phi_{2,2}+\omega_2) - \phi_{12}^2}
\end{aligned} \tag{36}$$

By the asymptotic normality of sample variances and covariances and the multivariate delta method (see Appendix A.5), $\widehat{\beta}_2$ has a distribution that is approximately normal for large samples, with approximate mean given by expression (36). Thus, it makes sense to call the second term in (36) 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.

[18] The simplification may be elementary, but that does not make it easy. I used Sage; see Appendix B.

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

Suppose that the null hypothesis is true, so $\beta_2 = 0$. In this case, Expression (36) becomes

$$\widehat{\beta}_2 \stackrel{a.s.}{\rightarrow} \frac{\beta_1\omega_1\phi_{12}}{(\phi_{1,1}+\omega_1)(\phi_{2,2}+\omega_2)-\phi_{12}^2}. \tag{37}$$

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

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

Brunner and Austin [5] have shown that whether H_0 is true or not, the standard error of $\widehat{\beta}_2$ goes to zero, and when the large-sample target of $\widehat{\beta}_2$ is non-zero, the p-value goes almost surely to zero. That is, the probability of making a Type I error goes to one because of measurement error in an explanatory variable — not the one being tested, but the one for which one is "controlling."

This is potentially a disaster, because the primary function of statistical hypothesis testing in the social and biological sciences is to filter out results that might be just random noise, and keep them from reaching the published research literature. Holding down the probability of a Type I error is critical. The preceding calculations show that in the very reasonable scenario where one needs to control for an explanatory variable but the measurement of that variable is imperfect (which is always the case), standard regression methods do not work as advertised. Instead, the probability of getting statistically significant results can go to one even when the null hypothesis is true and there is nothing real to discover. You should be appalled.

A large-scale simulation study All this is true as the sample size goes to infinity, but in reality no sample size can approach infinity. So it is important to see what happens for realistic sample sizes. The idea is to use computer-generated pseudo-random numbers to generate data sets in which the true parameter values are known, because actually those true parameter values are inputs to the program. Applying statistical methods to such simulated data allows one to investigate the performance of the methods empirically as well mathematically. Ideally, empirical and mathematical investigations of statistical questions are complementary, and usually reinforce one another.

Brunner and Austin [5] took this approach to the topic under discussion. They report a large simulation study in which random data sets were generated according to a factorial design with six factors. The factors were

- Sample size: $n = 50$, 100, 250, 500, 1000
- $Corr(X_1, X_2)$: $\phi_{12} = 0.00$, 0.25, 0.75, 0.80, 0.90
- Proportion of variance in Y explained by X_1: 0.25, 0.50, 0.75
- Reliability of W_1: 0.50, 0.75, 0.80, 0.90, 0.95
- Reliability of W_2: 0.50, 0.75, 0.80, 0.90, 0.95
- Distribution of the latent variables and error terms: Normal, Uniform, t, Pareto.

Thus there were $5\times5\times3\times5\times5\times4 = 7{,}500$ treatment combinations. Ten thousand random data sets were generated within each treatment combination, for a total of 75 million data sets. All the data sets were generated according to the true model of Example 0.7.3, with $\beta_2 = 0$, so that $H_0 : \beta_2 = 0$ was true in each case. For each data set, we fit the naive model (no measurement error), and tested $H_0 : \beta_2 = 0$ at $\alpha = 0.05$. The proportion of times H_0 is rejected is a Monte Carlo estimate of the Type I Error Probability.

The study yielded 7,500 estimated Type I error probabilities, and even looking at all of them is a big job. Table 1 shows a small but representative part of the results. In this table, all the variables and error terms are normally distributed, and the reliability of both explanatory variables was equal to 0.90. This means that 90% of the variance came from the real thing as opposed to random noise – a stellar value. The values of the regression coefficients were $\beta_0 = 1$, $\beta_1 = 1$ and of course $\beta_2 = 0$.

Remember that we are trying to test the effect of X_1 on Y controlling for X_2, and since we don't have X_1 and X_2, we are using W_1 and W_2 instead. In fact, because $H_0 : \beta_2 = 0$ is true, X_2 is conditionally independent of Y given $X_1 = x_1$. This means that the estimated Type I error probabilities in Table 1 should all be around 0.05 if the test is working properly.

When the correlation between X_1 and X_2 is zero (the first column of Table 1), none of the estimated Type I error probabilites is significantly different from 0.05. This is consistent with Equation (37), where $\widehat{\beta}_2$ converges to the right target when the covariance between X_1 and X_2 is zero. But as the correlation between explanatory variables increases, so does the Type I error probability – especially when the X_1 and Y is strong and the sample size is large. Observe that for the intermediate case in which 50% of variance in Y is explained by X_1 (admittedly a strong relationship, at least in the social sciences) and $n = 250$. As the correlation between X_1 and X_2 increases from zero to 0.90, the Type I error probability increases from 0.05 to about 0.60. With the strongest relationship beween X_1 and Y, and the largest sample size, the test of X_2's relationship to Y controlling for X_1 was significant 10,000 times out of 10,000. Again, this is when the null hypothesis is true, and Y is conditionally independent of X_2, given X_1.

Again, this simulation study was a 6-factor experiment with 7,500 treatment combinations. A rough way to see general trends is to look at marginal means, averaging the estimated Type I error probabilities over the other factors, for each factor in the study. Table 2 is actually six subtables, showing marginal estimated Type I error probabilities for each factor. The only one that may not be self-explanatory is "Base distribution."

Table 1: Estimated Type I Error

	Correlation between X_1 and X_2				
n	0.00	0.25	0.75	0.80	0.90
25% of variance in Y is explained by X_1					
50	0.0491†	0.0505†	0.0663	0.0740	0.0838
100	0.0541†	0.0527†	0.0896	0.0925	0.1227
250	0.0479†	0.0577†	0.1364	0.1688	0.2585
500	0.0510†	0.0588†	0.2399	0.2887	0.4587
1000	0.0489†	0.0734	0.4175	0.4960	0.7391
50% of variance in Y is explained by X_1					
50	0.0518†	0.0535†	0.0949	0.1081	0.1571
100	0.0501†	0.0541†	0.1512	0.1763	0.2710
250	0.0487†	0.0710	0.3065	0.3765	0.5994
500	0.0518†	0.0782	0.5499	0.6487	0.8740
1000	0.0500†	0.1132	0.8260	0.9120	0.9932
75% of variance in Y is explained by X_1					
50	0.0504†	0.0554†	0.1669	0.2072	0.3361
100	0.0510†	0.0599	0.3019	0.3791	0.5943
250	0.0487†	0.0890	0.6399	0.7542	0.9441
500	0.0496†	0.1296	0.9058	0.9599	0.9987
1000	0.0502†	0.2157	0.9969	0.9992	1.0000

†Not Significantly different from 0.05, Bonferroni corrected for 7,500 tests.

This is the distribution of X_1, X_2, e_1 and e_2, shifted when necessary to have expected value zero, and scaled to have variance for the particular treatment condition.

The inescapable conclusion is that ignoring measurement error in the explanatory variables can seriously inflate Type I error probabilities in multiple regression. To repeat, this is what people do all the time. The poison combination is measurement error in the variable for which you are "controlling," and correlation between latent explanatory variables. If either is zero, there is no problem. Factors affecting severity of the problem are (next slide)

- As the correlation between X_1 and X_2 increases, the problem gets worse.
- As the correlation between X_1 and Y increases, the problem gets worse.
- As the amount of measurement error in X_1 increases, the problem gets worse.
- As the amount of measurement error in X_2 increases, the problem gets *less* severe.
- As the sample size increases, the problem gets worse.
- Distribution of the variables does not matter much.

Table 2: Marginal Type I Error Probabilities

Base Distribution			
normal	Pareto	t Distr	uniform
0.38692448	0.36903077	0.38312245	0.38752571

Explained Variance		
0.25	0.50	0.75
0.27330660	0.38473364	0.48691232

Correlation between Latent Independent Variables				
0.00	0.25	0.75	0.80	0.90
0.05004853	0.16604247	0.51544093	0.55050700	0.62621533

Sample Size n				
50	100	250	500	1000
0.19081740	0.27437227	0.39457933	0.48335707	0.56512820

Reliability of W_1				
0.50	0.75	0.80	0.90	0.95
0.60637233	0.46983147	0.42065313	0.26685820	0.14453913

Reliability of W_2				
0.50	0.75	0.80	0.90	0.95
0.30807933	0.37506733	0.38752793	0.41254800	0.42503167

It is particularly noteworthy that the inflation of Type I error probability gets worse with increasing sample size. Generally in statistics, things get better as the sample size increases. This is an exception. For a large enough sample size, no amount of measurement error in the explanatory variables is safe, assuming that the latent explanatory variables are correlated.

It might be objected that null hypotheses are never exactly true in observational studies, so that estimating Type I error probability is a meaningless exercise. However, look at expression (36), the large-sample target of $\widehat{\beta}_2$ when the true value of β_2 (the parameter being tested) is not necessarily zero. Suppose that the true value of β_2 is negative, the true value of β_1 is positive, and the covariance between X_1 and X_2 is positive. This is a perfectly natural scenario. Depending on the values of the variances and covariances, it is quite possible for the second term in (36) to be a larger positive value, overwhelming β_2 and making the large-sample target of $\widehat{\beta}_2$ positive. Brunner and Austin report a smaller-scale simulation of this situation in which measurement error leads to rejection of the null hypothesis in the wrong direction nearly 100% of the time. This is a particularly nasty possibility, because findings that are opposite of the truth (especially if they are published) can only serve to muddy the waters and make scientific progress slower and more difficult.

Brunner and Austin go on to show that the inflation of Type I error probability arising from measurement error is not limited to multiple regression and measurement error of a simple additive type. It applies to other kinds of regression and other types of measurement error, including logistic regression, proportional hazards regression in survival analysis, log-linear models (for testing conditional independence in the presence of classification error, and median splits on explanatory variables, which is a kind of measurement error created by the data analyst. Even converting X_1 to ranks inflates Type I Error probability.

This is a serious problem, but only if one is interested in interpreting the results of statistical analyses to find out more about the world. If the only interest is in prediction, you just use the variables you have. You might wish your predictors were measured with less error, because that might make the predictions more accurate. But it doesn't really matter whether a given regression coefficient is positive or negative. On the other hand, if this is science then it matters.

It's worth observing that the news about true experimental studies is good. The first column of Table 1, where the covariance of explanatory variables is zero, illustrates the primary virtue of random assignment: it erases any relationship between experimental treatment and potential confounding variables. Thinking of X_2 as the treatment and X_1 as a covariate, it is apparent that in an experimental study, the Type I error probability is not inflated by measurement error in the treatment, the covariate, or both – as long as random assignment has made the latent versions of these variables independent, and the experimental procedure has been of sufficiently high quality that the corresponding measurement errors are uncorrelated.

This example also illustrates that assignment to experimental conditions need not be random to be effective. All that's needed is to somehow break up the relationship between the treatment and any possible confounding variables. In a clinical trial, for example, suppose that patients coming in to a medical clinic are assigned to experimental and control conditions alternately, and not randomly. There is no serious problem with this, because treatment condition would still be unrelated to any characteristic of the patients.

The whole issue of measurement error in the predictors is really just a sentence or two in the narrative about correlation versus causation. It goes like this. If X is related to Y, it could be that X is influencing Y, or that Y is influencing X, or that some confounding variables related to X are influencing Y. You might think that if you have an idea what those confounding variables are, you can control for them with regression methods. Unfortunately, if potential confounding variables are measured with error, the standard ways of controlling for them do not quite work (Brunner and Austin, 2009)[19].

The last two sentences are the addition to the standard narrative. It's only a couple of sentences, but it's still a big deal, because correlation-causation is a fundamental issue

[19]I could not resist citing the paper. There is no claim that Brunner and Austin discovered the problem with measurement error in the predictors. The ill effects of measurement error on estimation have been known since the 1930s, though the issue has been mostly ignored by mainstream statisticians and other users of statistical methods. What Brunner and Austin did was to review the literature and document the effect of measurement error on significance testing.

in research design. What's the solution? Surely it must be to admit that measurement error exists, and incorporate it directly into the statistical model.

0.8 Modeling measurement error

It is clear that ignoring measurement error in regression can yield conclusions that are very misleading. But as soon as we try building measurement error into the statistical model, we encounter a technical issue that will occupy a central role in this book: parameter identifiability.

A first try at including measurement error

Example 0.8.1 *Model Includes Measurement Error*

The following is basically the true model of Example 0.7.2, with everything normally distributed. Independently for $i = 1, \ldots, n$, let

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_i + \epsilon_i \\ W_i &= \nu + X_i + e_i, \end{aligned} \tag{38}$$

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 (38) 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

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

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

$$\begin{aligned} \mu_1 &= \mu_x + \nu \\ \mu_2 &= \beta_0 + \beta_1 \mu_x \\ \sigma_{1,1} &= \phi + \omega \\ \sigma_{1,2} &= \beta_1 \phi \\ \sigma_{2,2} &= \beta_1^2 \phi + \psi. \end{aligned} \tag{39}$$

It is impossible to solve these five equations for the seven model parameters[20]. That is, even with perfect knowledge of the probability distribution of the data (for the multivariate normal, that means knowing $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, period), it would be impossible to know the model parameters.

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

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

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

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

and thus the same distribution of the sample data.

No matter how large the sample size, it will be impossible to decide between $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$, because they imply exactly the same probability distribution of the observable data. The problem here is that the parameters of Model (38) are not *identifiable*. This calls for a brief discussion of identifiability, a topic of central importance in structural equation modeling.

0.9 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.

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

Definition 0.9.1 *A* Statistical Model *is a set of assertions that partly*[21] *specify the probability distribution of a set of observable data.*

Definition 0.9.2 *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 that imply exactly the same probability distribution of the observed data.

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

In Figure 14, θ_1 and θ_2 are two distinct sets of parameter values for which the distribution of the observable data is the same.

Figure 14: Two parameters values yielding the same probability distribution

Let T_n be a estimator that is consistent for both θ_1 and θ_2. What this means is that if θ_1 is the correct parameter value, eventually as n increases, the probability distribution of T_n will be concentrated in the circular neighborhood around θ_1. And if θ_1 is the correct parameter value, it the probability distribution will be concentrated around θ_2.

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

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

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

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

Definition 0.9.3 *The parameter is said to be* identifiable *at a point* $\boldsymbol{\theta}_0$ *if no other point in* Θ *yields the same probability distribution as* $\boldsymbol{\theta}_0$.

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

Definition 0.9.4 *The parameter is said to be* locally identifiable *at a point* $\boldsymbol{\theta}_0$ *if there is a neighbourhood of points surrounding* $\boldsymbol{\theta}_0$*, none of which yields the same probability distribution as* $\boldsymbol{\theta}_0$.

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

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

Definition 0.9.5 *Let* $g(\boldsymbol{\theta})$ *be a function of the parameter vector. If* $g(\boldsymbol{\theta}_0) \neq g(\boldsymbol{\theta})$ *implies* $P_{\boldsymbol{\theta}_0} \neq P_{\boldsymbol{\theta}}$ *for all* $\boldsymbol{\theta} \in \Theta$*, then the function* $g(\boldsymbol{\theta})$ *is said to be identifiable at the point* $\boldsymbol{\theta}_0$.

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

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

The idea of a *function* of the parameter vector covers a lot of territory. It includes individual parameters and sets of parameters, as well as things like products and ratios of parameters. Look at the moment structure equations (39) of Example 0.8.1. 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 of Example 0.8.1, the moment structure equations (39) consist of five equations in seven unknown parameters. It was shown by a numerical example that there were two different sets of parameter values that produced the same mean vector and covariance matrix, and hence the same distribution of the observable data. Actually, infinitely many parameter values produce the same distribution,

and it happens because there are more unknowns than equations. Theorem 2 is a strictly mathematical theorem[22] that provides the necessary details.

Theorem 2 *Let*

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

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

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

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

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

Even when the data are not assumed multivariate normal, the same process makes sense. Classical structural equation models, including models for regression with measurement error, are based on systems of simultaneous linear equations. Assuming simple random sampling from a large population, the observable data are independent and identically distributed, with a mean vector $\boldsymbol{\mu}$ and a covariance matrix $\boldsymbol{\Sigma}$ that may be written as functions of the model parameters in a straightforward way. If it is possible to solve uniquely for a given model parameter in terms of the elements of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, then that parameter is a function of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, which in turn are functions of the probability distribution of the data. A function of a function is a function, and so the parameter is a function of the probability distribution of the data. Hence, it is identifiable.

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

To summarize, we have arrived at the standard way to check parameter identifiability for any linear simultaneous equation model, not just measurement error regression. *First,*

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

calculate the expected value and covariance matrix of the observable data, as a function of the model parameters. If it is possible to solve uniquely for the model parameters in terms of the means, variances and covariances of the observable data, then the model parameters are identifiable.

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

But this would require knowing exactly what the non-normal distribution of the data might be. When it comes to analyzing actual data using linear models like the ones in this book, there are really only two alternatives. Either the distribution is assumed[23] normal, or it is acknowledged to be completely unknown. In both cases, parameters will either be identifiable from the mean and covariance matrix (usually just the covariance matrix), or they will not be identifiable at all.

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

0.10 Double measurement

Consider again the model of Example 0.8.1, a simple regression with measurement error in the single explanatory variable. This represents 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[24].

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

[24] The reason for different instruments or methods is to ensure that the errors of measurements are independent. For example, suppose a questionnaire is designed to measure racism. Respondents differ in their actual, true unobservable level of racism. They also differ in the extent to which they wish to be perceived as non-racist. If you give people two similar questionnaires in which they agree or disagree with various statements that are obviously about racism, the individuals who fake good on one questionnaire will also fake good on the other one. The result is that if e_1 and e_2 are the measurement errors in the two questionnaires, then e_1 and e_2 will surely have positive covariance. If the unknown covariance is assumed zero, the result will almost surely be incorrect estimation and inference. If the unknown covariance is another parameter in the model, it usually will create problems with identifiability. This all may seem quite technical, but there is a common-sense version. Problems with identifiability almost always correspond to shortcomings in research design. If data are collected in a way that is poorly thought out,

0.10.1 A scalar example

Example 0.10.1

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

$$\begin{aligned} W_{i,1} &= \nu_1 + X_i + e_{i,1} \\ W_{i,2} &= \nu_2 + X_i + e_{i,2} \\ Y_i &= \beta_0 + \beta_1 X_i + \epsilon_i, \end{aligned} \tag{40}$$

where

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

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

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

and variance covariance matrix

$$cov(\mathbf{D}_i) = \mathbf{\Sigma} = [\sigma_{i,j}] = \begin{pmatrix} \phi + \omega_1 & \phi & \beta_1\phi \\ & \phi + \omega_2 & \beta_1\phi \\ & & \beta_1^2\phi + \psi \end{pmatrix}. \tag{41}$$

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.

the data analysis is unlikely to yield valid conclusions. Taking two measurements that are likely to be contaminated in the same way is just not very smart.

- 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[25]. Thus, the study of parameter identifiability leads to a useful test of model fit.

- The constraint $\sigma_{1,3} = \sigma_{2,3}$ allows two solutions for β_1 in terms of the moments: $\beta_1 = \sigma_{13}/\sigma_{12}$ and $\beta_1 = \sigma_{23}/\sigma_{12}$. Does this mean the solution for β_1 is not "unique?" No; everything is okay. Because $\sigma_{1,3} = \sigma_{2,3}$, the two solutions are actually the same. If a parameter can be recovered from the moments in any way at all, it is identifiable.

- For the other model parameters appearing in the covariance matrix, the additional measurement of the explanatory variable also appears to have done the trick. It is easy to solve for ϕ, ω_1, ω_2 and ψ in terms of $\sigma_{i,j}$ values. Thus, these parameters are identifiable.

- On the other hand, the additional measurement did not help with the means and intercepts *at all*. Even assuming β_1 known because it can be recovered from $\boldsymbol{\Sigma}$, 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 work up to a point, but the lack of unique values for μ_x, ν_1, ν_2 and β_0 would cause numerical problems. A good solution is to *re-parameterize* the model, absorbing $\mu_x + \nu_1$ into a parameter called μ_1, $\mu_x + \nu_2$ into a parameter called μ_2, and $\beta_0 + \beta_1\mu_x$ into a parameter called μ_3. The parameters in $\boldsymbol{\mu} = (\mu_1, \mu_2, \mu_3)^\top$ lack meaning and interest[26], 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.4, 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.20) from Appendix A.

$$L(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = |\boldsymbol{\Sigma}|^{-n/2}(2\pi)^{-np/2} \exp -\frac{n}{2}\left\{tr(\widehat{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}^{-1}) + (\overline{\mathbf{D}} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\overline{\mathbf{D}} - \boldsymbol{\mu})\right\} \tag{42}$$

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 $\boldsymbol{\Sigma}$.

[25] 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?"

[26] 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.

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

$$L(\boldsymbol{\theta}) = |\boldsymbol{\Sigma}(\boldsymbol{\theta})|^{-n/2}(2\pi)^{-np/2} \exp -\frac{n}{2}\left\{tr(\widehat{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1})\right\}, \tag{43}$$

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 scalar double measurement Example 0.10.1, $\boldsymbol{\theta} = (\phi, \omega_1, \omega_2, \beta_1, \psi)^\top$, and $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ is given by Expression (41). Maximum likelihood estimation is numerical, and the full range of large-sample likelihood methods described in Section A.6.3 of Appendix A is available.

Testing goodness of model fit

When there are more covariance structure equations than unknown parameters and the parameters are identifiable, the parameters are said to be *over-identified*. In this case, the model implies functional connections between some variances and covariances. In the small example we are considering, it is clear from Expression (41) on page 63 that $\sigma_{13} = \sigma_{23}$, because they both equal $\beta_1\phi$. This is a testable null hypothesis, and if it is rejected, the model is called into question.

The traditional way to do the test[27] is to compare the fit of the model to the fit of a completely unrestricted multivariate normal using the test statistic

$$G^2 = -2\ln\frac{L\left(\overline{\mathbf{D}}, \boldsymbol{\Sigma}(\widehat{\boldsymbol{\theta}})\right)}{L(\overline{\mathbf{D}}, \widehat{\boldsymbol{\Sigma}})} = n\left(tr\left(\widehat{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}(\widehat{\boldsymbol{\theta}})^{-1}\right) - \ln\left|\widehat{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}(\widehat{\boldsymbol{\theta}})^{-1}\right| - p\right), \tag{44}$$

where $\widehat{\boldsymbol{\Sigma}}$ is the ordinary sample variance-covariance matrix with n in the denominator, and $L(\cdot,\cdot)$ is the multivariate normal likelihood (42) on page 64. The degrees of freedom equals the number of covariance structure equations minus the number of parameters. The idea is that if there are k parameters and m unique variances and covariances, the model imposes $m - k$ equality constraints on the variances and covariances[28]. Those are the constraints being tested, even when we don't know exactly what they are. The goodness of fit test is examined more closely in Chapter 5.

The matrix $\boldsymbol{\Sigma}(\widehat{\boldsymbol{\theta}})$ is called the *reproduced covariance matrix*. It is the covariance matrix of the observable data, written as a function of the model parameters and evaluated at the MLE. For the present example,

$$\boldsymbol{\Sigma}(\widehat{\boldsymbol{\theta}}) = \begin{pmatrix} \widehat{\phi} + \widehat{\omega}_1 & \widehat{\phi} & \widehat{\beta}_1\widehat{\phi} \\ & \widehat{\phi} + \widehat{\omega}_2 & \widehat{\beta}_1\widehat{\phi} \\ & & \widehat{\beta}_1^2\widehat{\phi} + \widehat{\psi} \end{pmatrix}$$

[27] The test is documented on page 447 of Jöreskog's classic (1978) article [8] in *Psychometrika*, but I believe it had been in Jöreskog and Sörbom's LISREL software for years before that.

[28] Here's why. In most cases, it is possible to choose just k of the m variances and covariances, and establish identifiability by solving k equations in k unknowns. In this case, there are $m - k$ unused, redundant equations. Each sets a variance or covariance equal to some function of the model parameters of the model parameters. Substituting the solutions for the parameters in terms of σ_{ij} back into the unused equations will yield $m - k$ equality constraints on the variances and covariances.

The reproduced covariance matrix obeys all model-induced constraints, while $\widehat{\boldsymbol{\Sigma}}$ does not. However, they should be close if the model is right. In the limiting case where $\widehat{\boldsymbol{\Sigma}} = \boldsymbol{\Sigma}(\widehat{\boldsymbol{\theta}})$, the G^2 statistic in (44) equals zero.

When the parameter vector is identifiable and there are more unique variances and covariances than parameters, we call the parameter vector *over-identifiable*. An alternative terminology is to say that the "model is over-identified." The equality restrictions on $\boldsymbol{\Sigma}$ imposed by the model are called *over-identifying restrictions*. The likelihood ratio test for goodness of fit is testing the null hypothesis that the over-identifying restrictions are true.

Suppose that the entire parameter vector is identifiable, and $m = k$. That is, the number of parameters is equal to the number of unique variances and covariances. In this case, identifiability is established by solving k equations in k unknowns. The function from parameters to the variances and covariances is one-to-one (injective), and the model imposes no constraints on the variances and covariances. In this case the parameter vector is said to be *just identifiable*. Alternatively, the model is often said to be "just identified," or *saturated*. In this case, $\widehat{\boldsymbol{\Sigma}} = \boldsymbol{\Sigma}(\widehat{\boldsymbol{\theta}})$ by the invariance principle, and the likelihood ratio test statistics for goodness of fit automatically equals zero. The degrees of freedom $m - k = 0$ also. These values are usually displayed by software, which could be confusing unless you know why. It means the model is not testable. It is incapable of being challenged by any data set, at least using this technology.

0.10.2 Computation with `lavaan`

A variety of commercial software is available for fitting structural equation models, including LISREL, EQS, Amos and Mplus. I myself have used mostly SAS `proc calis` until recently. In keeping with the open-source philosophy of this text, we will use the free, open-source R package `lavaan`; the name is short for LAtent VAriable ANalysis. The software is described very well by Rosseel [13] in his 2012 article in the *Journal of Statistical Software*. The capabilities of `lavaan` have grown since the article was published. A nice tutorial is available at `http://lavaan.ugent.be/tutorial`.

This first illustration of `lavaan` will use a data set simulated from the model of Example 0.10.1, the same example we have been studying. It may be a toy example, but it's an educational toy. Readers familiar with `lavaan` will notice that I am using synax that favours explicitness over brevity, for now. R input and output will be interspersed with explanation.

When I begin an R session, I like to clear the deck with `rm(list=ls())`, removing any existing R objects that may be in the workspace. The statement `options(scipen=999)` suppresses scientific notation. This is just a matter of taste.

The `lavaan` package may be installed with the `install.packages` command. You only need to do this once, which is why it's commented out below. `library(lavaan)` is necessary to load the package, every time.

```
> rm(list=ls()); options(scipen=999)
> # install.packages("lavaan", dependencies = TRUE)
```

```
> library(lavaan)
This is lavaan 0.5-23.1097
lavaan is BETA software! Please report any bugs.
```

Next, we read the data, look at the first few lines, and obtain a summary and correlation matrix. Notice that that the data file has only observable variables (obviously), and that their means are certainly not zero. In practice, we would examine the data much more carefully. This vital step in data analysis will not be mentioned again.

```
> babydouble = read.table("http://www.utstat.toronto.edu/~brunner/openSEM
 /data/Babydouble.data.txt")
> head(babydouble)
     W1    W2     Y
1  9.94 12.24 15.23
2 12.42 11.32 14.55
3 10.43 10.40 12.40
4  9.07  9.85 17.09
5 11.04 11.98 16.83
6 10.40 10.85 15.04

> summary(babydouble)
       W1               W2              Y
 Min.   : 6.190   Min.   : 6.76   Min.   : 3.98
 1st Qu.: 8.932   1st Qu.: 9.11   1st Qu.:10.97
 Median : 9.720   Median :10.05   Median :13.22
 Mean   : 9.809   Mean   :10.06   Mean   :13.10
 3rd Qu.:10.655   3rd Qu.:10.99   3rd Qu.:15.46
 Max.   :12.830   Max.   :13.57   Max.   :21.62

> cor(babydouble)
          W1        W2         Y
W1 1.0000000 0.5748331 0.1714324
W2 0.5748331 1.0000000 0.1791539
Y  0.1714324 0.1791539 1.0000000
```

Notice that the sample correlations of W_1 with Y and W_2 with Y are very close. This is consistent with the model-induced constraint $\sigma_{13} = \sigma_{23}$, especially if $\omega_1 = \omega_2$.

Next comes specification of the model to be fit. Again, this is the model of Example 0.10.1 on page 63. The entire model specification is in a *model string*, assigned to the string variable `dmodel1`. If the model is big and you are using it repeatedly, you can compose the model string in a separate file and bring it in with `readlines`.

```
> dmodel1 =  'Y ~ beta1*F         # Latent variable model
+             F =~ 1*W1 + 1*W2    # Measurement model
+             # Variances (covariances would go here too)
```

```
+             F~~phi*F      # Var(F) = phi
+             Y~~psi*Y      # Var(epsilon) = psi
+             W1~~omega1*W1 # Var(e1) = omega1
+             W2~~omega2*W2 # Var(e2) = omega2
+             '
```

It's best to discuss the model string line by line.

`Y ~ beta1*F`: This is reminiscent of R's `lm` syntax. The translation is $Y = \beta_1 X + \epsilon$. Notice that there is no β_0. Though you can specify intercepts and expected values in `lavaan` if you wish, by default they are invisible. Thus the whole process of re-parameterization and swallowing all the non-identifiable expected values and intercepts into $\boldsymbol{\mu}$ (see page 64) is implicit.

As in the syntax of the EQS software, names of latent variables must begin with the letter F or f, for "factor." Anticipating Chapter 2 on factor analysis, a factor is another term for latent variable. Since X is a latent variable, it is re-named F. The error term ϵ is invisible, but it's there — again as in the `lm` syntax.

`F =~ 1*W1 + 1*W2`: This looks like F is being produced by W_1 and W_2, when actually it's the other way around. However, if you read $\sim$ and $=\sim$ as two different flavours of "is modelled as," it makes more sense. The statment stands for two model equations:

$$\begin{aligned} W_1 &= 1 * X + e_1 \\ W_2 &= 1 * X + e_2 \end{aligned}$$

These two statements constitute the *measurement model* for this simple example. The observable variables W_1 and W_2 are called *indicators* of X. An indicator of a latent variable is an observable variable that arises from only that latent variable plus an error term. In `lavaan`, a latent variable must have indicators. Otherwise, it is assumed observable even if its name begins with the letter F and it's not in the input data set. The explicit "1*" syntax is necessary if you want the coefficients to equal one. Otherwise, `lavaan` will assume you want coefficients that are free parameters in the model, but you don't feel like naming them. It will try to be helpful, with results that are unfortunate in this case[29].

`F~~ phi*F`: As the comment statement says, this means $Var(X) = \phi$. The double tilde is a way of naming variances, or setting variances equal to numeric constants, if that's what you want to do. Notice that the symbol F appears on both sides. If you had two different variable names, the statement would specify a covariance. Since a variance may be viewed as the covariance of a random variable with itself, this is good notation. Also be aware that if a covariance is not specified, it equals zero.

`Y~~ psi*Y`: In contrast to the preceding statement, this one is *not* saying that $Var(Y) = \psi$. It is saying $Var(\epsilon) = \psi$. Here's the rule. If a variable appears on the left side of any model equation, then the $\sim\sim$ notation specifies the variance or covariance of the error

[29] `lavaan`'s "helpful" behaviour really is helpful for many users under many circumstances. It is based on rules for parameter identifiability that will be developed later in this text.

term in the equation. If the variable appears only on the right side (possibly in more than one equation), the $\sim\sim$ notation specifies the variance or covariance of the variable itself. In this way, though error terms are never named in `lavaan`, you can name their variances, and you can name their covariances with other variables and error terms.

`W1~~omega1*W1`: $Var(e_1) = \omega_1$

`W2~~omega2*W2`: $Var(e_2) = \omega_2$

A covariance between the measurement errors e_1 and e_2 would be specified with something like `W1~~omega12*W2`. A covariance of c between X and ϵ would be specified with `F~~c*Y`.

Next, we fit the model and look at a summary. We use the `lavaan` function[30] (same name as the `lavaan` package).

```
> dfit1 = lavaan(dmodel1, data=babydouble)
> summary(dfit1)
lavaan (0.5-23.1097) converged normally after  23 iterations

  Number of observations                           150

  Estimator                                         ML
  Minimum Function Test Statistic                0.007
  Degrees of freedom                                 1
  P-value (Chi-square)                           0.933

Parameter Estimates:

  Information                                 Expected
  Standard Errors                             Standard

Latent Variables:
                   Estimate  Std.Err  z-value  P(>|z|)
  F =~
    W1                1.000
    W2                1.000

Regressions:
                   Estimate  Std.Err  z-value  P(>|z|)
  Y ~
    F       (bet1)    0.707    0.290    2.442    0.015
```

[30] Model fitting can also be accomplished with the `sem` and `cfa` functions. With this "user friendly" approach, the model specification in the model string is less elaborate, and the software makes choices about the model for you. These choices are intended to be helpful, and may or may not be what you want.

```
Variances:
                   Estimate  Std.Err  z-value  P(>|z|)
    F        (phi)    1.104    0.181    6.104    0.000
   .Y        (psi)    9.775    1.153    8.481    0.000
   .W1      (omg1)    0.834    0.158    5.265    0.000
   .W2      (omg2)    0.800    0.156    5.123    0.000
```

We first learn that the numerical parameter estimation converged in 23 iterations, $n = 150$, and estimation was by maximum likelihood – the default. The "`Minimum Function Test Statistic`" is the G^2 statistic given in expression (44) on page 65: the likelihood ratio test for goodness of model fit.

Then there is a section entitled `Latent Variables`, saying that F is manafested by the indicators W_1 and W_2. The "estimates" are the fixed numerical constants of 1.000, specified in the model string. More generally, this section would include all the latent variables in a model. If coefficients (factor loadings) linking the latent variables to their indicators were not pre-specified, their estimates would appear here, together with tests of difference from zero.

The next section of the summary is `Regressions`. These correspond to all the model equations using the $\sim$ rather than the $\sim=$ notation, whether the variables involved are latent or observed. Here, we have maximum likelihood estimates, standard errors, Z-tests for whether the parameter equals zero, and two-sided p-values. The standard errors are what you would expect. They are square roots of the diagonal elements of the inverse of the Hessian of the minus log likelihood. If this does not make sense, see the maximum likelihood review in Appendix A. Also, observe that in the `summary` display, the parameter names are abbreviated to four characters.

The last section of the summary is `Variances`. Covariances would go here too, if any had been specified in the model. We have maximum likelihood estimates of the variance parameters, standard errors, and two-sided Z-tests for whether the parameter equals zero. When the variance in question is the variance of an error term rather than of the variable itself, the variable name is preceded by a dot, as in `.Y`, `.W1` and `.W2`.

Testing whether variances equal zero It might seem strange to test whether variances equal zero, when they are automatically assumed greater than zero according to the model. It's not as silly as you might think. Look at Equation (41) on page 63, which gives the covariance matrix of the observable variables for this model, in terms of the model parameters. The covariance $\sigma_{1,2}$ equals ϕ, which is a variance. That means that the covariance between W_1 and W_2 must be greater than zero if the model is correct; this would not necessarily be true for an arbitrary covariance matrix.

The other variance parameters, because they are identifiable, can also be written as functions of the variances and covariances $\sigma_{i,j}$. This means that they also correspond to functions of the variances and covariances — functions that must be greater than zero if the model is correct. In this way, we see that the model also imposes *inequality constraints*

on the covariance matrix $\mathbf{\Sigma}$. The most obvious of these constraints[31] can be tested by looking at the estimates of the variance parameters in the model. If the variance estimates are less than zero, particularly if they are *significantly* less than zero, the model is thrown into question.

The conclusion is that testing whether variances equal zero is another way to test model fit. A good practice is to check the equality constraint first with the likelihood ratio test for goodness of fit, and then worry about inequality constraints provided that the first test is non-significant. It is quite common for inequality violations to disappear once the equality violations have been fixed.

The R object created by the `lavaan` function contains a large amount of additional information. The `parameterEstimates` function returns a data frame that gives more detail about the parameter estimates, including confidence intervals.

```
> parameterEstimates(dfit1)
  lhs op rhs  label   est    se     z pvalue ci.lower ci.upper
1   Y  ~   F  beta1 0.707 0.290 2.442  0.015    0.140    1.275
2   F =~  W1        1.000 0.000    NA     NA    1.000    1.000
3   F =~  W2        1.000 0.000    NA     NA    1.000    1.000
4   F ~~   F    phi 1.104 0.181 6.104  0.000    0.750    1.459
5   Y ~~   Y    psi 9.775 1.153 8.481  0.000    7.516   12.034
6  W1 ~~  W1 omega1 0.834 0.158 5.265  0.000    0.524    1.145
7  W2 ~~  W2 omega2 0.800 0.156 5.123  0.000    0.494    1.105
```

The `dfit1` function yields details abut the model fitting, including the automatic starting values used by `lavaan`.

```
> parTable(dfit1)
  id lhs op rhs user block group free ustart exo  label plabel start   est    se
1  1   Y  ~   F    1     1     1    1     NA   0  beta1   .p1. 0.000 0.707 0.290
2  2   F =~  W1    1     1     1    0      1   0          .p2. 1.000 1.000 0.000
3  3   F =~  W2    1     1     1    0      1   0          .p3. 1.000 1.000 0.000
4  4   F ~~   F    1     1     1    2     NA   0    phi   .p4. 0.050 1.104 0.181
5  5   Y ~~   Y    1     1     1    3     NA   0    psi   .p5. 5.164 9.775 1.153
6  6  W1 ~~  W1    1     1     1    4     NA   0 omega1   .p6. 0.968 0.834 0.158
7  7  W2 ~~  W2    1     1     1    5     NA   0 omega2   .p7. 0.953 0.800 0.156
```

A vector containing the parameter estimates may be obtained with the `coef` function. This is very useful when the parameter estimates are to be used in further calculations.

```
> coef(dfit1) # A vector of MLEs
 beta1    phi    psi omega1 omega2
 0.707  1.104  9.775  0.834  0.800
```

[31] It can be challenging to obtain all the inequality constraints, particularly in a minimal form. See Chapter 5

The `fitted` function returns a list, whose first element is the reproduced covariance matrix $\boldsymbol{\Sigma}(\widehat{\boldsymbol{\theta}})$. If means are specified it also returns what might be called the "reproduced mean vector" $\boldsymbol{\mu}(\widehat{\boldsymbol{\theta}})$.

```
> fitted(dfit1) # Sigma(thetahat)
$cov
   W1     W2     Y
W1  1.939
W2  1.104  1.904
Y   0.781  0.781 10.327

$mean
W1 W2  Y
 0  0  0
```

As usual with R, the `vcov` function returns the estimated asymptotic covariance matrix, the inverse of the observed Fisher information (Hessian).

```
> vcov(dfit1)
       beta1  phi    psi    omega1 omega2
beta1   0.084
phi    -0.007  0.033
psi    -0.035  0.002  1.328
omega1  0.003 -0.004 -0.002  0.025
omega2  0.003 -0.005 -0.002 -0.007  0.024
```

Even though the upper triangular entries are not shown, that's just a display method. The whole symmetric matrix is available for furter calculation.

The `logLik` function returns the log likelihood evaluated at the MLE.

```
> logLik(dfit1)
'log Lik.' -878.512 (df=5)
```

It would be possible to use `logLike` to compute likelihood ratio tests, but the `anova` function is more convenient. One can fit a restricted model by specifying the constraints in the `lavaan` statement.

```
> # Fit a restricted model (restricted by H0)
> dfit1r = lavaan(dmodel1, data=babydouble, constraints = 'omega1==omega2')
> anova(dfit1r,dfit1)
Chi Square Difference Test

       Df  AIC    BIC  Chisq Chisq diff Df diff Pr(>Chisq)
dfit1   1 1767 1782.1 0.0071
dfit1r  2 1765 1777.1 0.0262   0.019189       1     0.8898
```

To test a null hypothesis with multiple constraints, put the constraints on separate lines. This is the code for testing $H_0 : \omega_1 = \omega_2, \phi = 1$.

```
> # Put multiple constraints on separate lines.
> dfit1r2 = lavaan(dmodel1, data=babydouble, constraints = 'omega1==omega2
+                                                           phi==1')
> anova(dfit1r2,dfit1)
```

Illustrating a Wald test[32] of $H_9 : \omega_1 = \omega_2$, we first define the publicly available `Wtest` function, and then enter the $\mathbf{L}$ matrix and do the calculation.

```
> # For Wald tests: Wtest = function(L,Tn,Vn,h=0) # H0: L theta = h
> source("http://www.utstat.utoronto.ca/~brunner/Rfunctions/Wtest.txt")
> LL = cbind(0,0,0,1,-1); LL
     [,1] [,2] [,3] [,4] [,5]
[1,]    0    0    0    1   -1

> Wtest(LL,coef(dfit1),vcov(dfit1))
         W          df    p-value
0.01918586 1.00000000 0.88983498
```

It is only a little surprising that the Wald and likelihood ratio test statistics are so close. The two tests are asymptotically equivalent under the null hypothesis, meaning that the difference between the two test statistic values goes to zero in probability when H_0 is true. In this case, the null hypothesis is exactly true (these are simulated data), and the sample size of $n = 150$ is fairly large.

The `lavaan` software makes it remarkably convenient to estimate non-linear functions of the parameters, along with standard errors calculated using the multivariate delta method (see the end of Section A.5 in Appendix A). This is accomplished with the `:=` operator, as shown below. In this example, two functions of the parameter vector are specified. The first function is $\omega_1 - \omega_2$; because this function is linear, the Z-test for whether it equals zero is equivalent to the Wald test of $H_0 : \omega_1 = \omega_2$ directly above. The second function is the reliability of $W_!$. Using Equation (29) on page 41, this is $\frac{\phi}{\phi+\omega_1}$.

```
> # Non-linear functions of the parameters with :=
> dmodel1b = 'Y ~ beta1*F         # Latent variable model
+             F =~ 1*W1 + 1*W2    # Measurement model
+             # Variances (covariances would go here too)
+             F~~phi*F      # Var(F) = phi
+             Y~~psi*Y      # Var(epsilon) = psi
+             W1~~omega1*W1 # Var(e1) = omega1
+             W2~~omega2*W2 # Var(e2) = omega2
+             diff := omega1-omega2
```

[32]The Wald test of the linear null hypothesis $\mathbf{L}\boldsymbol{\theta} = \mathbf{h}$ is given in Appendix A, Equation (A.33) on page 235

```
+             rel1 := omega1/(omega1+phi)
+             '
> dfit1b = lavaan(dmodel1b, data=babydouble)
> parameterEstimates(dfit1b)
   lhs op                  rhs  label   est    se     z pvalue ci.lower ci.upper
1    Y  ~                    F  beta1 0.707 0.290 2.442  0.015    0.140    1.275
2    F =~                   W1        1.000 0.000    NA     NA    1.000    1.000
3    F =~                   W2        1.000 0.000    NA     NA    1.000    1.000
4    F ~~                    F    phi 1.104 0.181 6.104  0.000    0.750    1.459
5    Y ~~                    Y    psi 9.775 1.153 8.481  0.000    7.516   12.034
6   W1 ~~                   W1 omega1 0.834 0.158 5.265  0.000    0.524    1.145
7   W2 ~~                   W2 omega2 0.800 0.156 5.123  0.000    0.494    1.105
8 diff :=       omega1-omega2   diff 0.035 0.252 0.139  0.890   -0.458    0.528
9 rel1 := omega1/(omega1+phi)   rel1 0.430 0.066 6.540  0.000    0.301    0.559
```

The Z statistic of 0.139 for the null hypothesis $\omega_1 - \omega_2 = 0$ matches the Wald test of the same null hypothesis, with $W = Z^2$.

```
> sqrt(0.01918586)
[1] 0.138513
```

Trying to fit models with non-identifiable parameters This sub-section may be skipped without loss of continuity. It contains more details about how `lavaan` works, and also material on the connection of identifiability to maximum likelihood estimation. The account of how double measurement can help with identifiability is continued on page 82.

Trying to estimate the parameters of a structural equation model without first checking identifiability is like jumping out of an airplane without checking that your backpack contains a parachute and not just a sleeping bag. You shouldn't do it. Unfortunately, people do it all the time. Sometimes it's because they have little or no idea what parameter identifiability is. Sometimes it's because the model is a little non-standard, and checking identifiability is too much work[33] Sometimes, it's because of coding errors. Typos in the model string can easily specify a model that's non-identifiable, because a mis-spelled parameter name is assumed to represent a different parameter. Anyway, it's interesting to see how `lavaan` deals with models you *know* are not identified. The main lesson is that sometimes it complains, and sometimes it just returns a meaningless answer with no obvious indication that anything is wrong.

Example 0.10.2

In this first example, `lavaan` complains loudly. The model is obtained by taking `dmodel1` (that's the model of Example 0.10.1 on page 63) and adding unknown coefficients λ_1 and λ_2 linking F to W_1 and W_2 respectively. The result is that there are now two more

[33] In later chapters, we will use Sage to ease the burden of symbolic calculation. See Appendix B.

parameters, for a total of seven. There are still only six variances and covariances, so the model fails the Parameter Count Rule (Rule 1 on page 61), and we know the parameters can be identifiable on at most a set of volume zero in the parameter space.

```
> dmodel2  =  'Y ~ beta1*F          # Latent variable model
+             F =~ lambda1*W1 + lambda2*W2    # Measurement model
+             # Variances (covariances would go here too)
+             F~~phi*F      # Var(F) = phi
+             Y~~psi*Y      # Var(epsilon) = psi
+             W1~~omega1*W1 # Var(e1) = omega1
+             W2~~omega2*W2 # Var(e2) = omega2
+             '
```

When we try to fit the model, it is clear that something is wrong.

```
> dfit2 = lavaan(dmodel2, data=babydouble)
Warning message:
In lav_model_vcov(lavmodel = lavmodel, lavsamplestats = lavsamplestats,  :
  lavaan WARNING: could not compute standard errors!
  lavaan NOTE: this may be a symptom that the model is not identified.
```

In this case, `lavaan` correctly guessed that the parameters were not identifiable. Here's what happened.

When `lavaan` does maximum likelihood estimation, it is minimizing a function proportional to the minus log likelihood plus a constant[34]. If the parameter vector is massively non-identifiable as in the present case, the typical parameter vector belongs to an infinite, connected set whose members all yield exactly the same covariance matrix and hence the same value of the function being minimized. The graph of the function does not look like a high-dimensional bowl. Instead, it resembles a high-dimensional river valley. The non-unique minimum is on the flat surface of the water at the bottom of the valley. The numerical search starts somewhere up in the hills, and then trickles downhill, usually until it comes to the river. Then it stops. The stopping place (the MLE) depends entirely on where the search began.

The surface is not strictly concave up at the stopping point, so the Hessian matrix (see Expression A.26 in Appendix A) is not positive definite. However, the valley function is convex, so that the Hessian has to be non-negative definite. Consequently all its eigenvalues are greater than or equal to zero. They can't all be positive, or the Hessian would be positive definite. This means there must be at least one zero eigenvalue. Hence, the determinant of the Hessian is zero and its inverse does not exist.

[34]The constant is $L(\overline{\mathbf{D}}, \widehat{\boldsymbol{\Sigma}})$, the multivariate normal likelihood evaluated at the unrestricted MLE of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. The function is also divided by n, which can help with numerical accuracy. When the search finds a minimum, multiplication by $2n$ yields the test statistic given in Equation (44).

The standard errors of the MLEs are the square roots of the diagonal elements of the estimated asymptotic variance-covariance matrix. This matrix is obtained by inverting the Hessian of the minus log likelihood; see Expression (A.31) in Appendix A. Since the inverse does not exist, the standard errors can't be computed, and `lavaan` issues a warning about it. This whole scenario is so common that `lavaan` also speculates – correctly in this case – that the problem arises from lack of parameter identifiability.

This is not an error; it's just a warning. So a model fit object is created.

```
> summary(dfit2)
lavaan (0.5-23.1097) converged normally after  25 iterations

  Number of observations                           150

  Estimator                                         ML
  Minimum Function Test Statistic                   NA
  Degrees of freedom                                -1

Parameter Estimates:

  Information                                 Expected
  Standard Errors                             Standard

Latent Variables:
                   Estimate  Std.Err  z-value  P(>|z|)
  F =~
    W1      (lmb1)    1.022       NA
    W2      (lmb2)    1.060       NA

Regressions:
                   Estimate  Std.Err  z-value  P(>|z|)
  Y ~
    F       (bet1)    0.736       NA

Variances:
                   Estimate  Std.Err  z-value  P(>|z|)
    F        (phi)    1.019       NA
   .Y        (psi)    9.776       NA
   .W1      (omg1)    0.871       NA
   .W2      (omg2)    0.761       NA
```

After "normal" convergence (hummm), the `Minimum Function Test Statistic` is `NA`, or missing even though it could be computed. The degrees of freedom are -1, impossible for a chi-squared statistic. The degrees of freedom are calculated as number of unique variances and covariances minus number of parameters. When it's negative, this is a sure

sign the model has failed the parameter count rule, and the parameter vector can't be identifiable. The software could check this and inform the user, but as of this writing it does not. Parameter estimates (corresponding to the point where the search stopped) are given, but standard errors are `NA` and there are no significance tests.

Example 0.10.3

In this next example, we modify the model of Example 0.10.1 again, keeping the unknown factor loadings λ_1 and λ_2 that connect the latent explanatory variable F to its indicators W_1 and W_2, but making the two measurement error variances equal: $\omega_1 = \omega_2 = \omega$. Everything else remains the same. The model has six unknown parameters and six unique variances and covariances, so it passes the test of the parameter count rule. This means identifiability is possible, but not guaranteed.

```
> # dmodel3 passes the parameter count rule, but its parameters are not identifiable.
> dmodel3  = 'Y ~ beta1*F                    # Latent variable model
+             F =~ lambda1*W1 + lambda2*W2    # Measurement model
+             F~~phi*F      # Var(F) = phi
+             Y~~psi*Y      # Var(epsilon) = psi
+             W1~~omega*W1 # Var(e1) = omega
+             W2~~omega*W2 # Var(e2) = omega
+             '
> dfit3 = lavaan(dmodel3, data=babydouble)
>
```

`lavaan` fits the model and silently returns the R prompt. Looking at `summary`,

```
> summary(dfit3)
lavaan (0.5-23.1097) converged normally after  19 iterations

  Number of observations                           150

  Estimator                                         ML
  Minimum Function Test Statistic                0.014
  Degrees of freedom                                 0
  Minimum Function Value               0.0000466299101

Parameter Estimates:

  Information                                 Expected
  Standard Errors                             Standard

Latent Variables:
                   Estimate  Std.Err  z-value  P(>|z|)
  F =~
```

```
    W1      (lmb1)    1.048    0.089   11.797    0.000
    W2      (lmb2)    1.034    0.089   11.658    0.000

Regressions:
                   Estimate  Std.Err  z-value  P(>|z|)
  Y ~
    F       (bet1)    0.736    0.275    2.671    0.008

Variances:
                   Estimate  Std.Err  z-value  P(>|z|)
    F        (phi)    1.019    0.087   11.713    0.000
   .Y        (psi)    9.776    1.153    8.481    0.000
   .W1      (omeg)    0.817    0.094    8.660    0.000
   .W2      (omeg)    0.817    0.094    8.660    0.000
```

Everything seems to be fine, but it's not fine! The parameters of this model are not identifiable, and as in the previos example (Example 0.10.2), the MLE is not unique. At first glance, it's not obvious why.

The matrix equation (45) gives the covariance matrix of $(W_{i,1}, W_{i,2}, Y_i)^\top$, expressing the six covariance structure equations in six unknowns, in a compact form.

$$\begin{pmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} \\ & \sigma_{22} & \sigma_{23} \\ & & \sigma_{33} \end{pmatrix} = \begin{pmatrix} \lambda_1^2\phi + \omega & \lambda_1\lambda_2\phi & \lambda_1\beta_1\phi \\ & \lambda_2^2\phi + \omega & \lambda_2\beta_1\phi \\ & & \beta_1^2\phi + \psi \end{pmatrix}. \tag{45}$$

First, it is clear that if just one of $\lambda_1 = 0$, $\lambda_2 = 0$ or $\beta_1 = 0$, the zero value would be detectable from the covariance matrix, making that parameter identifiable. However, the remaining four equations in five unknowns would fail the parameter count rule, so that the other parameters would not be identifiable. If two or three of λ_1, λ_2 and β_1 were equal to zero, it would be impossible to tell which ones they were. Solving the remaining three equations in six unknowns is a hopeless task, and the entire parameter vector would be non-identifiable.

All these identifiability problems are local, and would have no effect on numerical maximum likelihood unless the true parameter values in question were zero. So consider points in the parameter space where λ_1, λ_2 and β_1 are all non-zero. In this case, ω and ψ are identifiable, because

$$\omega = \sigma_{11} - \frac{\sigma_{12}\sigma_{13}}{\sigma_{23}} \quad \text{and} \quad \psi = \sigma_{33} - \frac{\sigma_{13}\sigma_{23}}{\sigma_{12}}.$$

In fact, ω is over-identified, and this imposes the testable constraint $\sigma_{11} = \sigma_{22}$ on the covariance matrix. As for the other parameters, let $\boldsymbol{\theta}_1$ be an arbitrary point in the parameter space. Letting $c \neq 0$, consider the two parameter vectors

$\boldsymbol{\theta}_1$	λ_1	λ_2	β_1	ϕ	ω	ψ
$\boldsymbol{\theta}_c$	$c\lambda_1$	$c\lambda_2$	$c\beta_1$	$\frac{\phi}{c^2}$	ω	ψ

(46)

It is clear that $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_c$ both yield the same covariance matrix (45), and hence the same value of the likelihood function. In fact, every point in the parameter space belongs to an infinite family $\{\boldsymbol{\theta}_c : c \neq 0\}$ whose members all have the same the same likelihood.

This means that if a numerical search locates a minimum, that point is just one of an infinite number of points in the parameter space where that same minimum value is attained. Furthermore the set is connected, and we are back to the river valley picture of Example 0.10.2. Also as in Example 0.10.2, the Hessian matrix has at least one zero eigenvalue, its determinant equals zero, and its inverse does not exist. Therefore, the Hessian of the minus log likelihood cannot be inverted to give an estimated asymptotic covariance matrix.

`lavaan` inverted the matrix to get standard errors anyway, but that must be due to numerical inaccuracy[35]. Floating point operations are not quite the same as exact mathematics, and one suspects that numerically, the determinant of the Hessian was not quite zero, and `lavaan` got the standard errors by (approximately) inverting a matrix that was almost the Hessian, but not quite. If the almost-Hessian was nearly singular, its inverse should be nearly singular too. Checking the determinant,

```
> V = vcov(dfit3) ; det(V)
[1] 0.00000000000000000000000000000000000000000001801361
```

This makes sense. Also notice the usefulness (in this case) of `options(scipen=999)`, which suppresses scientific notation. Trying to get beck to the Hessian,

```
> H = solve(V)
Error in solve.default(V) :
  system is computationally singular: reciprocal condition number = 1.95535e-19
```

This suggests that however `lavaan` inverted the almost-Hessian to obtain estimated asymptotic variances and covariances, it was not using R's `solve` function.

A good way to confirm the account of what's happening here is to choose a different set of starting values. Then, the numerical search should trickle downhill into the valley until it reaches a different point on the likelihood river. The estimated parameters should be very different (except for ψ and ω), but the value of the likelihood function should be the same. In the first test, I will try to start the search exactly in the river, at a point fairly distant from the first MLE. If the map provided by expression (46) is correct, this should work.

To specify starting value of a regression coefficient in `lavaan`, one replaces the coefficient with `start(`*number*`)`, where *number* is a numeric starting value. A generic example is `Y~start(4.2)*X`. This is excellent when you are letting `lavaan` name parameters automatically, but what if you want to also name the regression coefficient? Somewhat oddly, you specify the connection between X and Y twice, and `lavaan` picks up the information in two passes through the syntax. The generic example would look

[35] This is not a criticism of `lavaan`. It's a fact of life in scientific computing.

like this: `Y~beta*X + start(4.2)*X`. A similar syntax works for variances, like this: `Y~~sigmasq*Y + start(1.0)*Y`.

Since the estimated β_1 for model `dmodel3` was positive, we will make it negative this time. As far as I can tell, the starting values have to be literal numbers, and not R variables.

```
> c = -2
> thetac = coef(dfit3); thetac
  beta1 lambda1 lambda2     phi     psi   omega   omega
  0.736   1.048   1.034   1.019   9.776   0.817   0.817

> thetac[1] = c*thetac[1]; thetac[2] = c*thetac[2]; thetac[3] = c*thetac[3]
> thetac[4] = thetac[4]/c^2
> cat(thetac)
-1.471502 -2.095291 -2.068575 0.2548175 9.775661 0.816833 0.816833
```

The `cat` function was used to get more decimal places in the output, because I needed to copy and paste the numbers into the model string. To start right in the river, we need as much accuracy as possible.

```
> dmodel3b = 'Y ~ beta1*F + start(-1.471502)*F
+             F =~ lambda1*W1 + start(-2.095291)*W1 +
+                  lambda2*W2 + start(-2.068575)*W2
+             # Variances (covariances would go here too)
+             F~~phi*F + start(0.2548175)*F      # Var(F) = phi
+             Y~~psi*Y + start(9.775661)*Y       # Var(epsilon) = psi
+             W1~~omega*W1 + start(0.816833)*W1  # Var(e1) = omega
+             W2~~omega*W2 + start(0.816833)*W2  # Var(e2) = omega
+             '
> dfit3b = lavaan(dmodel3b, data=babydouble)
> show(dfit3b)
lavaan (0.5-23.1097) converged normally after   2 iterations

  Number of observations                           150

  Estimator                                         ML
  Minimum Function Test Statistic                0.014
  Degrees of freedom                                 0
  Minimum Function Value               0.0000466299101
```

This time the search found a minimum in two iterations rather than 19. Binding the starting and ending values into a matrix for easy inspection, we see that they are identical, at least to R's accuracy of display. This means that essentially, we started the numerical search at one of the infinitely many MLEs — as planned.

```
> rbind(thetac,coef(dfit3b))

           beta1   lambda1   lambda2       phi      psi    omega    omega
thetac -1.471502 -2.095291 -2.068575 0.2548175 9.775661 0.816833 0.816833
       -1.471502 -2.095291 -2.068575 0.2548175 9.775661 0.816833 0.816833
```

As expected, the parameter estimates are quite different from the first set we located, except for the estimates of the identifiable parameters ψ and ω.

```
> rbind(coef(dfit3),coef(dfit3b))      |
          beta1   lambda1   lambda2       phi      psi    omega    omega
[1,]  0.7357509  1.047646  1.034288 1.0192699 9.775661 0.816833 0.816833
[2,] -1.4715020 -2.095291 -2.068575 0.2548175 9.775661 0.816833 0.816833
```

Though the locations of the MLEs are different, the log likelihood at those points is the same. Again, the theoretical analysis is confirmed.

```
> c( logLik(dfit3), logLik(dfit3b) )
[1] -878.5155 -878.5155
```

In one last variation, the search starts fairly close to the river[36] but not exactly on target, and finds its way to yet another MLE. Here, starting values are provided for λ_1, λ_2, β_1 and ϕ. `lavaan` provides starting values for ψ and ω.

```
> dmodel3c = 'Y ~ beta1*F + start(6)*F
+             F =~ lambda1*W1 + start(8)*W1 +
+                  lambda2*W2  + start(8)*W2
+             # Variances (covariances would go here too)
+             F~~phi*F + start(1/64)*F      # Var(F) = phi
+             Y~~psi*Y       # Var(epsilon) = psi
+             W1~~omega*W1   # Var(e1) = omega
+             W2~~omega*W2   # Var(e2) = omega
+             '
> dfit3c = lavaan(dmodel3c, data=babydouble)
> c( logLik(dfit3), logLik(dfit3b), logLik(dfit3b) )
[1] -878.5155 -878.5155 -878.5155


> rbind( coef(dfit3), coef(dfit3b), coef(dfit3c) )
          beta1   lambda1   lambda2       phi      psi    omega    omega
[1,]  0.7357509  1.047646  1.034288 1.0192699 9.775661 0.816833 0.816833
[2,] -1.4715020 -2.095291 -2.068575 0.2548175 9.775661 0.816833 0.816833
[3,]  5.7803725  8.230750  8.125805 0.0165135 9.775661 0.816833 0.816833
```

[36] To find a point that is fairly close, observe from (46) that the product $\lambda_1\lambda_2\phi$ must be constant for all points on the river. The constant is pretty close to 1, and β_1 should be around 3/4 of λ_1. So $\beta_1 = 6$, $\lambda_1 = \lambda_2 = 8$ and $\phi = 1/64$ should do it.

So the search located another point with the same maximum log likelihood, fairly far from the other two. For the parameters that are not identifiable, the answer depends on the starting value.

When the parameters of a model are all identifiable, the minus log likelihood should have a unique global minimum, and `lavaan`'s default starting values should be adequate most of the time. However even when the parameters are identifiable, local maxima and minima are possible. If you suspect the search may have located a local minimum (perhaps because some of the MLEs are extremely large), you may need to specify your own starting values. Try several sets. The `parTable` function can be used to verify that the starting values were the ones you intended. In the display below, `ustart` are the starting values given by the user, some of which are `NA` because they were not specified. The `start` column are the starting values used by the software, and the `est` column (estimates) is where the search ended — at the parameter estimates.

```
> parTable(dfit3c)
  id  lhs op  rhs user block group free ustart exo   label plabel start   est    se
1  1    Y  ~    F    1     1     1    1  6.000   0   beta1   .p1. 6.000 5.780 1.895
2  2    F =~   W1    1     1     1    2  8.000   0 lambda1   .p2. 8.000 8.231 0.822
3  3    F =~   W2    1     1     1    3  8.000   0 lambda2   .p3. 8.000 8.126 0.819
4  4    F ~~    F    1     1     1    4  0.016   0     phi   .p4. 0.016 0.017 0.004
5  5    Y ~~    Y    1     1     1    5     NA   0     psi   .p5. 5.164 9.776 1.153
6  6   W1 ~~   W1    1     1     1    6     NA   0   omega   .p6. 0.968 0.817 0.094
7  7   W2 ~~   W2    1     1     1    7     NA   0   omega   .p7. 0.953 0.817 0.094
8  8 .p6. == .p7.    2     0     0    0     NA   0                0.000 0.000 0.000
```

0.10.3 The Double Measurement Design in Matrix Form

Consider the general case of regression with measurement error in both the explanatory variables and the response variables. Independently for $i = 1, \ldots, n$, let

$$\begin{aligned} \mathbf{W}_{i,1} &= \boldsymbol{\nu_1} + \mathbf{X}_i + \mathbf{e}_{i,1} \\ \mathbf{V}_{i,1} &= \boldsymbol{\nu_2} + \mathbf{Y}_i + \mathbf{e}_{i,2} \\ \mathbf{W}_{i,2} &= \boldsymbol{\nu_3} + \mathbf{X}_i + \mathbf{e}_{i,3} \\ \mathbf{V}_{i,2} &= \boldsymbol{\nu_4} + \mathbf{Y}_i + \mathbf{e}_{i,4}, \\ \mathbf{Y}_i &= \boldsymbol{\alpha} + \boldsymbol{\beta}\mathbf{X}_i + \boldsymbol{\epsilon}_i \end{aligned} \tag{47}$$

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 $\mathbf{\Phi}$, 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 $\mathbf{\Psi}$, 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 and a set of constant terms that could be called *measurement bias*[37].

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

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

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

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

$\boldsymbol{\alpha}$, $\boldsymbol{\nu}_1$, $\boldsymbol{\nu}_2$, $\boldsymbol{\nu}_3$ and $\boldsymbol{\nu}_4$ are vectors of constants.

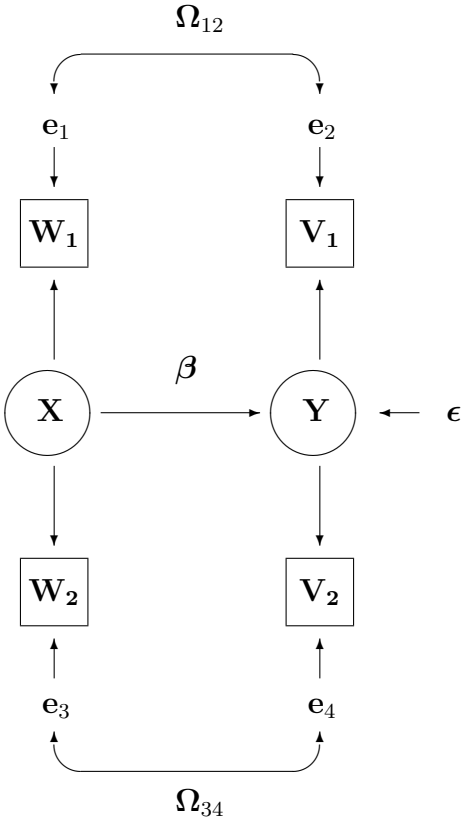
$E(\mathbf{X}_i) = \boldsymbol{\mu}_x$.

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 general, correlation within measurement methods is almost unavoidable. The ability of the double measurement model to admit the existence of correlated measurement error and still be identifiable is a great virtue.

In symbolic terms, $\mathbf{e}_{i,1}$ is error in measuring the explanatory variables by method one, and $\mathbf{e}_{i,2}$ is error in measuring the response variables by method one. $cov(\mathbf{e}_{i,1}) = \mathbf{\Omega}_{11}$ need not be diagonal, so method one's errors of measurement for the explanatory variables may be correlated with one another. Similarly, $cov(\mathbf{e}_{i,2}) = \mathbf{\Omega}_{22}$ need not be diagonal, so method one's errors of measurement for the response variables may be correlated with one

[37] 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.

Figure 15: The Double Measurement Model



another. And, errors of measurement using the same method may be correlated between the explanatory and response variables. For method one, this is represented by the matrix $cov(\mathbf{e}_{i,1}, \mathbf{e}_{i,2}) = \mathbf{\Omega}_{12}$. 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 have to be equal. In particular, the corresponding diagonal elements may be unequal. This means that measurements of a variable by two different methods do not need to be equally precise.

The model is depicted in Figure 15. 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.

Parameter identifiability As usual in structural equation models, the moments (specifically, the expected values and variance-covariance matrix) of the observable data are functions of the model parameters. If the model parameters are also functions of the moments, then they are identifiable[38]. For the double measurement model, the parameters appearing in the covariance matrix of the observable variables are identifiable, but the parameters appearing only in the mean vector are not. Accordingly, we split the job into two parts, starting with the covariance matrix. The first part is typical of easier proofs for structural equation models. The goal is to solve for the model parameters in terms of elements of the variance-covariance matrix of the observable data. This shows the parameters are functions of the distribution, so that no two distinct parameter values could yield the same distribution of the observed data.

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

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

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

Now we express all the $\boldsymbol{\Sigma}_{ij}$ sub-matrices in terms of the parameter matrices of Model (47) by straightforward variance-covariance calculations. Students may be reminded that things go smoothly if one substitutes for everything in terms of explanatory variables and error terms before actually starting to calculate covariances. For example,

$$\begin{aligned}
\boldsymbol{\Sigma}_{12} &= cov(\mathbf{W}_{i,1}, \mathbf{V}_{i,1}) \\
&= cov\left(\boldsymbol{\nu_1} + \mathbf{X}_i + \mathbf{e}_{i,1},\, \boldsymbol{\nu_2} + \mathbf{Y}_i + \mathbf{e}_{i,2}\right) \\
&= cov\left(\boldsymbol{\nu_1} + \mathbf{X}_i + \mathbf{e}_{i,1},\, \boldsymbol{\nu_2} + \boldsymbol{\alpha} + \boldsymbol{\beta}\mathbf{X}_i + \boldsymbol{\epsilon}_i + \mathbf{e}_{i,2}\right) \\
&= cov\left(\mathbf{X}_i + \mathbf{e}_{i,1},\, \boldsymbol{\beta}\mathbf{X}_i + \boldsymbol{\epsilon}_i + \mathbf{e}_{i,2}\right) \\
&= cov(\mathbf{X}_i, \boldsymbol{\beta}\mathbf{X}_i) + cov(\mathbf{X}_i, \boldsymbol{\epsilon}_i) + cov(\mathbf{X}_i, \mathbf{e}_{i,2}) + cov(\mathbf{e}_{i,1}, \boldsymbol{\beta}\mathbf{X}_i) + cov(\mathbf{e}_{i,1}, \boldsymbol{\epsilon}_i) + cov(\mathbf{e}_{i,1}, \mathbf{e}_{i,2}) \\
&= cov(\mathbf{X}_i, \mathbf{X}_i)\boldsymbol{\beta}^\top + 0 + 0 + 0 + 0 + \boldsymbol{\Omega}_{12} \\
&= \boldsymbol{\Phi}\boldsymbol{\beta}^\top + \boldsymbol{\Omega}_{12}.
\end{aligned}$$

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

[38] Meaning identifiable from the moments. For multivariate normal models and also in general practice, a parameter is identifiable from the mean vector and covariance matrix, or not at all.

$$\begin{aligned}
\boldsymbol{\Sigma} &= \left(\begin{array}{c|c|c|c} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} & \boldsymbol{\Sigma}_{13} & \boldsymbol{\Sigma}_{14} \\ \hline & \boldsymbol{\Sigma}_{22} & \boldsymbol{\Sigma}_{23} & \boldsymbol{\Sigma}_{24} \\ \hline & & \boldsymbol{\Sigma}_{33} & \boldsymbol{\Sigma}_{34} \\ \hline & & & \boldsymbol{\Sigma}_{44} \end{array}\right) \\
&= \left(\begin{array}{c|c|c|c} \boldsymbol{\Phi}+\boldsymbol{\Omega}_{11} & \boldsymbol{\Phi}\boldsymbol{\beta}^\top+\boldsymbol{\Omega}_{12} & \boldsymbol{\Phi} & \boldsymbol{\Phi}\boldsymbol{\beta}^\top \\ \hline & \boldsymbol{\beta}\boldsymbol{\Phi}\boldsymbol{\beta}^\top+\boldsymbol{\Psi}+\boldsymbol{\Omega}_{22} & \boldsymbol{\beta}\boldsymbol{\Phi} & \boldsymbol{\beta}\boldsymbol{\Phi}\boldsymbol{\beta}^\top+\boldsymbol{\Psi} \\ \hline & & \boldsymbol{\Phi}+\boldsymbol{\Omega}_{33} & \boldsymbol{\Phi}\boldsymbol{\beta}^\top+\boldsymbol{\Omega}_{34} \\ \hline & & & \boldsymbol{\beta}\boldsymbol{\Phi}\boldsymbol{\beta}^\top+\boldsymbol{\Psi}+\boldsymbol{\Omega}_{44} \end{array}\right)
\end{aligned} \tag{48}$$

The equality (48) corresponds to a system of ten matrix equations in nine matrix unknowns. The unknowns are the parameter matrices of Model (47): $\boldsymbol{\Phi}$, $\boldsymbol{\beta}$, $\boldsymbol{\Psi}$, $\boldsymbol{\Omega}_{11}$, $\boldsymbol{\Omega}_{22}$, $\boldsymbol{\Omega}_{33}$, $\boldsymbol{\Omega}_{44}$, $\boldsymbol{\Omega}_{12}$, 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.

$$\begin{aligned}
\boldsymbol{\Phi} &= \boldsymbol{\Sigma}_{13} \\
\boldsymbol{\beta} &= \boldsymbol{\Sigma}_{23}\boldsymbol{\Phi}^{-1} = \boldsymbol{\Sigma}_{14}^\top\boldsymbol{\Phi}^{-1} \\
\boldsymbol{\Psi} &= \boldsymbol{\Sigma}_{24} - \boldsymbol{\beta}\boldsymbol{\Phi}\boldsymbol{\beta}^\top \\
\boldsymbol{\Omega}_{11} &= \boldsymbol{\Sigma}_{11} - \boldsymbol{\Phi} \\
\boldsymbol{\Omega}_{22} &= \boldsymbol{\Sigma}_{22} - \boldsymbol{\beta}\boldsymbol{\Phi}\boldsymbol{\beta}^\top - \boldsymbol{\Psi} \\
\boldsymbol{\Omega}_{33} &= \boldsymbol{\Sigma}_{33} - \boldsymbol{\Phi} \\
\boldsymbol{\Omega}_{44} &= \boldsymbol{\Sigma}_{44} - \boldsymbol{\beta}\boldsymbol{\Phi}\boldsymbol{\beta}^\top - \boldsymbol{\Psi} \\
\boldsymbol{\Omega}_{12} &= \boldsymbol{\Sigma}_{12} - \boldsymbol{\Phi}\boldsymbol{\beta}^\top \\
\boldsymbol{\Omega}_{34} &= \boldsymbol{\Sigma}_{34} - \boldsymbol{\Phi}\boldsymbol{\beta}^\top
\end{aligned} \tag{49}$$

This shows that the parameters appearing in the covariance matrix $\boldsymbol{\Sigma}$ are identifiable. This includes the critical parameter matrix $\boldsymbol{\beta}$, which determines the connection between explanatory variables and response variables.

Intercepts

Now Model (47) Let $\boldsymbol{\mu} = E(\mathbf{D}_i)$. This vector of expected values may be written as a partitioned vector, as follows.

$$\boldsymbol{\mu} = \left(\begin{array}{c} \boldsymbol{\mu}_1 \\ \hline \boldsymbol{\mu}_2 \\ \hline \boldsymbol{\mu}_3 \\ \hline \boldsymbol{\mu}_4 \end{array}\right) = \left(\begin{array}{c} E(\mathbf{W}_{i,1}) \\ \hline E(\mathbf{V}_{i,1}) \\ \hline E(\mathbf{W}_{i,2}) \\ \hline E(\mathbf{V}_{i,2}) \end{array}\right) = \left(\begin{array}{c} \boldsymbol{\nu}_1 + \boldsymbol{\mu}_x \\ \hline \boldsymbol{\nu}_2 + \boldsymbol{\alpha} + \boldsymbol{\beta}\boldsymbol{\mu}_x \\ \hline \boldsymbol{\nu}_3 + \boldsymbol{\mu}_x \\ \hline \boldsymbol{\nu}_4 + \boldsymbol{\alpha} + \boldsymbol{\beta}\boldsymbol{\mu}_x \end{array}\right). \tag{50}$$

The parameters that appear in $\boldsymbol{\mu}$ but not $\boldsymbol{\Sigma}$ are contained in $\boldsymbol{\nu}_1$, $\boldsymbol{\nu}_2$, $\boldsymbol{\nu}_3$, $\boldsymbol{\nu}_4$, $\boldsymbol{\mu}_x$ and $\boldsymbol{\alpha}$. To identify these parameters, one would need to solve the equations in (50) uniquely for these six parameter vectors. Even with $\boldsymbol{\beta}$ considered known and fixed because it is identified in (49), this is impossible in most of the parameter space, because (50) 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 (50). This can be viewed as a re-parameterization of the model. In practice, the functions $\boldsymbol{\mu_1} \ldots, \boldsymbol{\mu_4}$ of the parameters are usually not of much interest. They are estimated by the corresponding sample means, conveniently forgotten, and almost never mentioned.

To summarize, the parameters appearing in the covariance matrix are identifiable. This includes $\boldsymbol{\beta}$, the quantity of primary interest. Means and intercepts are not identifiable, but they are absorbed in a re-parameterization and set aside. It's no great loss. In practice, if data are collected following the double measurement recipe, then the data analysis may proceed with no worries about parameter identifiability.

For the double measurement model, there are more covariance structure equations than unknowns. Thus the model is over-identified, and testable. Notice in the covariance structure equations (48), that $\boldsymbol{\Sigma}_{14} = \boldsymbol{\Sigma}_{23}^\top$. As in the scalar Example 0.10.1 (see page 63), 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.

Estimation and testing

Normal model As in Example 0.10.1, the (collapsed) expected values are estimated by the corresponding vector of sample means, and then set aside. Under a multivariate normal model, these terms literally disappear from the likelihood function (42) on page 64. The resulting likelihood is (43) on page 65. The full range of large-sample likelihood methods is available. Maximum likelihood estimates are asymptotically normal, and asymptotic standard errors are convenient by-products of the numerical minimization as described in Section A.6.3 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 (47) 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 (49) 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 (49), one is tempted to just put hats on everything to obtain Method-of-Moments estimators. However, we can do a little better. Note that while $\boldsymbol{\Phi} = \boldsymbol{\Sigma}_{12}$ is a symmetric matrix in the population and $\widehat{\boldsymbol{\Sigma}}_{12}$ *converges* to a symmetric matrix, $\widehat{\boldsymbol{\Sigma}}_{12}$ will be non-symmetric for any finite sample size (with probability one if the distributions involved are continuous). A better estimator is obtained by averaging pairs of off-diagonal elements:

$$\widehat{\boldsymbol{\Phi}}_M = \frac{1}{2}(\widehat{\boldsymbol{\Sigma}}_{13} + \widehat{\boldsymbol{\Sigma}}_{13}^\top), \tag{51}$$

where the subscript M indicates a Method-of-Moments estimator. Using the second line of (49), a reasonable though non-standard estimator of $\boldsymbol{\beta}$ is

$$\widehat{\boldsymbol{\beta}}_M = \frac{1}{2}\left(\widehat{\boldsymbol{\Sigma}}_{14}^\top + \widehat{\boldsymbol{\Sigma}}_{23}\right)\widehat{\boldsymbol{\Phi}}_M^{-1} \tag{52}$$

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[39] basis for large-sample interval estimation and testing.

However, there is no need to bother. Research on the robustness of the normal model for structural equation models (Amemiya, Fuller and Pantula, 1987; Anderson and Rubin, 1956; Anderson and Amemiya, 1988; Anderson, 1989; Anderson and Amemiya, 1990; Browne, 1988; Browne and Shapiro, 1988; Satorra and Bentler, 1990) shows that procedures for (such as likelihood ratio and Wald tests) based on a multivariate normal model are asymptotically valid even when the normal assumption is false. And Satorra and Bentler (1990) describe Monte Carlo work suggesting that normal-theory methods generally perform better than at least one method (Browne, 1984) that is specifically designed

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

to be distribution-free. Since the methods suggested by the estimator (52) are similar to Browne's weighted least squares approach, they are also unlikely to be superior to the standard normal-theory tools.

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

The BMI Health Study

Body mass index. (BMI) is defined as weight in kilograms divided by height in meters squared. It represents weight relative to height, and is a measure of how thick, or hefty a person is. People with a BMI less than 18 are described as underweight, those over 25 are described as overweight, and those over 30 are described as obese. However, many professional athletes have BMI numbers in the overweight range.

High BMI tends to be associated with poor health, and with indicators such as high blood pressure and high cholesterol. However, people with high BMI tend to be older and fatter. Perhaps age and physical condition are responsible for the association of BMI to health. The natural idea is to look at the connection of BMI to health indicators, controlling for age and some indicator of physical condition like percent body fat. The problem is that percent body fat (and to a lesser extent, age) are measured with error. As discussed in Section 0.7, standard ways of controlling for them with ordinary regression are highly suspect. The solution is double measurement regression.

Example 0.10.4 *The BMI health study*[40]

In this study, there are five latent variables. Each one was were measured twice, by different personnel at different locations and mostly by different methods. The variables are age, BMI, percent body fat, cholesterol level, and diastolic blood pressure.

- In measurement set one, age was self report. In measurement set two, age was based on a passport or birth certificate.
- In measurement set one, the height and weight measurements maknig up BMI were conducted in a doctor's office, following no special procedures. In measurement set two, they were conducted by a lab technician. Patients had tp remove their shoes, and wore a hospital gown.
- In measurement set one, estimated percent body fat was based on measurements with tape and calipers, conducted in the doctor's office. In measurement set two, percent body fat was estimated by submerging the participant in a water tank (hydrostatic weighing).

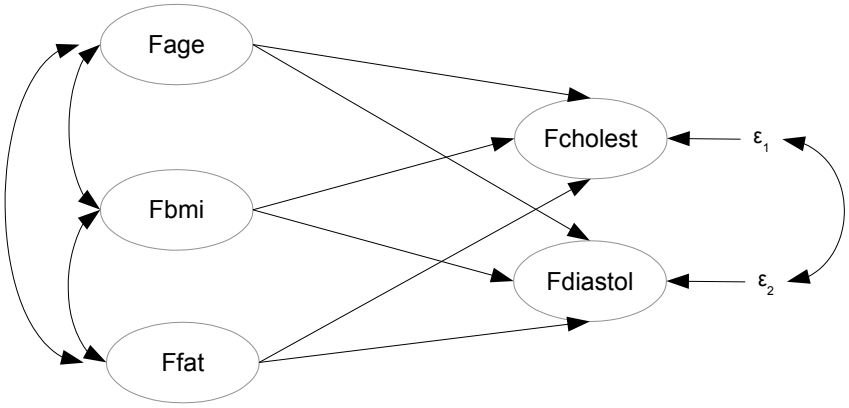
[40]This study is fictitious, and the data come from a combination of random number generation and manual editing. As far as I know, nothing like this has actually been done. I believe it should be.

- In measurement set one, serum (blood) cholesterol level was measured in lab 1. In measurement set two, it was measured in lab 2. There is no known difference between the labs in quality.
- In measurement set one, diastolic blood pressure was measured in the doctor's office usng a standard manual blood pressure cuff. In measurement set two, bolld pressure was measured in the lab by a digital device, and was mostly automatic.

Measurement set two was of generally higher quality than measurement set one. Correlation of measurement errors is possible within sets, but unlikely between sets.

Figure 16 shows a regression model for the latent variables. Because there are two response variables, it is multivariate regression. The names of the latent variables begin with the letter "F," for factor. This is a naming convention of the `lavaan` software. First,

Figure 16: Latent variable model for the BMI health study



we read the data and take a look. The variables are self-explanatory. There are 500 cases.

```
 > bmidata = read.table("http://www.utstat.toronto.edu/~brunner/openSEM/data/bmi.data.txt")
> head(bmidata)
  age1 bmi1 fat1 cholest1 diastol1 age2 bmi2 fat2 cholest2 diastol2
1   63 24.5 16.5    195.4       38   60 23.9 20.1    203.5       66
2   42 13.0  1.9    184.3       86   44 14.8  2.6    197.3       78
3   32 22.5 14.6    354.1      104   33 21.7 20.4    374.3       73
4   59 25.5 19.0    214.6       93   58 28.5 20.0    203.7      106
5   45 26.5 17.8    324.8       97   43 25.0 12.3    329.7       92
6   31 19.4 17.1    280.7       92   42 19.9 19.9    276.7       87
```

The standard, naive approach to analyzing these data is to ignore the possibility of measurement error, and use ordinary linear regression. One could either use just the better set of measurements (set 2), or average them. Averaging is a little better, because it improves reliability.

```
> age = (age1+age2)/2; bmi = (bmi1+bmi2)/2; fat = (fat1+fat2)/2
> cholest = (cholest1+cholest2)/2; diastol = (diastol1+diastol2)/2
```

There are two response variables (cholesterol level and diastolic blood pressure), so we fit a conventional multivariate linear model, and look at the multivariate test of BMI controlling for age and percent body fat. The full model has age, percent body fat and BMI, while the restricted model has just age and percent body fat.

```
> fullmod = lm( cbind(cholest,diastol) ~ age + fat + bmi)
> restrictedmod = update(fullmod, . ~ . - bmi) # Remove var(s) being tested
> anova(fullmod,restrictedmod) # Gives multivariate test.
Analysis of Variance Table

Model 1: cbind(cholest, diastol) ~ age + fat + bmi
Model 2: cbind(cholest, diastol) ~ age + fat
  Res.Df Df Gen.var.  Pillai approx F num Df den Df    Pr(>F)
1    496      591.89
2    497  1   599.36 0.02869   7.3106      2    495 0.0007431 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

The conclusion is that controlling for age and percent body fat, BMI is related to cholesterol, or diastolic blood pressure, or both. The `summary` function gives two sets of univariate output. Primary interest is in the t-tests for `bmi`.

```
> summary(fullmod) # Two sets of univariate output
Response cholest :

Call:
lm(formula = cholest ~ age + fat + bmi)

Residuals:
     Min       1Q   Median       3Q      Max
-148.550  -34.243    2.626   33.661  165.582

Coefficients:
            Estimate Std. Error t value             Pr(>|t|)
(Intercept) 220.0610    21.0109  10.474 < 0.0000000000000002 ***
age          -0.2714     0.2002  -1.356              0.17578
fat           2.2334     0.5792   3.856              0.00013 ***
bmi           0.5164     1.0154   0.509              0.61128
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 52.43 on 496 degrees of freedom
```

```
Multiple R-squared:  0.09701,Adjusted R-squared:  0.09155
F-statistic: 17.76 on 3 and 496 DF,  p-value: 0.00000000005762


Response diastol :

Call:
lm(formula = diastol ~ age + fat + bmi)

Residuals:
    Min      1Q  Median      3Q     Max
-44.841  -7.140  -0.408   7.612  41.377

Coefficients:
            Estimate Std. Error t value             Pr(>|t|)
(Intercept) 49.69194    4.52512  10.981 < 0.0000000000000002 ***
age          0.12648    0.04311   2.934              0.003504 **
fat          0.64056    0.12474   5.135          0.000000406 ***
bmi          0.82627    0.21869   3.778              0.000177 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 11.29 on 496 degrees of freedom
Multiple R-squared:  0.3333,Adjusted R-squared:  0.3293
F-statistic: 82.67 on 3 and 496 DF,  p-value: < 0.00000000000000022
```

For cholesterol, we have $t = 0.509$ and $p = 0.61128$. The conclusion is that controlling for age and percent body fat, there is no evidence of a connection between body mass index and serum cholesterol level.

For diastolic blood pressure, the test of BMI controlling for age and percent body fat is $t = 3.778$ and $p = 0.000177$. This time the conclusion is that even controlling for age and percent body fat, higher BMI is associated with higher average diastolic blood pressure – a bad sign for health. However, this "even controlling for" conclusion is exactly the kind of mistake that is often caused by ignoring measurement error; see Section 0.7. So, we specify a proper double measurement regression model. Note the covariances between measurement errors within sets of measurements, but not between.

```
bmimodel1 =
    ############################################################
    # Latent variable model
    # ---------------------
   'Fcholest ~ beta11*Fage + beta12*Fbmi + beta13*Ffat
    Fdiastol ~ beta21*Fage + beta22*Fbmi + beta23*Ffat
    #
    # Measurement model
    # -----------------
```

```
    Fage =~ 1*age1 + 1*age2
    Fbmi =~ 1*bmi1 + 1*bmi2
    Ffat =~ 1*fat1 +1*fat2
    Fcholest =~ 1*cholest1 + 1*cholest2
    Fdiastol =~ 1*diastol1 + 1*diastol2
    #
    # Variances and covariances
    # -------------------------
    # Of latent explanatory variables
    Fage ~~ phi11*Fage; Fage ~~ phi12*Fbmi; Fage ~~ phi13*Ffat
                        Fbmi ~~ phi22*Fbmi; Fbmi ~~ phi23*Ffat
                                           Ffat ~~ phi33*Ffat
    # Of error terms in latent the regression (epsilon_ij)
    Fcholest ~~ psi11*Fcholest; Fcholest ~~ psi12*Fdiastol
                                Fdiastol ~~ psi22*Fdiastol
    # Of measurement errors (e_ijk) for measurement set 1
    age1 ~~ w111*age1; age1 ~~ w112*bmi1; age1 ~~ w113*fat1;
    age1 ~~ w114*cholest1; age1 ~~ w115*diastol1
            bmi1 ~~ w122*bmi1; bmi1 ~~ w123*fat1; bmi1 ~~ w124*cholest1; bmi1 ~~ w125*diastol1
                    fat1 ~~ w133*fat1; fat1 ~~ w134*cholest1; fat1 ~~ w135*diastol1
                            cholest1 ~~ w144*cholest1; cholest1 ~~ w145*diastol1
                                                       diastol1 ~~ w155*diastol1
    # Of measurement errors (e_ijk) for measurement set 2
    age2 ~~ w211*age2; age2 ~~ w212*bmi2; age2 ~~ w213*fat2;
    age2 ~~ w214*cholest2; age2 ~~ w215*diastol2
            bmi2 ~~ w222*bmi2; bmi2 ~~ w223*fat2; bmi2 ~~ w224*cholest2; bmi2 ~~ w225*diastol2
                    fat2 ~~ w233*fat2; fat2 ~~ w234*cholest2; fat2 ~~ w235*diastol2
                            cholest2 ~~ w244*cholest2; cholest2 ~~ w245*diastol2
                                                       diastol2 ~~ w255*diastol2
  ' ################## End of bmimodel1 ##################
```

But then when we try to fit this perfectly nice model, there is trouble.

```
> # install.packages("lavaan", dependencies = TRUE) # Only need to do this once
> library(lavaan)
> fit1 = lavaan(bmimodel1, data=bmidata)
Warning message:
In lavaan(bmimodel1, data = bmidata) :
  lavaan WARNING: model has NOT converged!
```

Taking a look at just the fit of the model,

```
> show(fit1)
** WARNING ** lavaan (0.5-23.1097) did NOT converge after 5685 iterations
** WARNING ** Estimates below are most likely unreliable

  Number of observations                           500

  Estimator                                         ML
```

```
  Minimum Function Test Statistic                NA
  Degrees of freedom                             NA
  P-value                                        NA
```

What happened here is that the numerical search for the MLE stopped after 5,685 steps, without finding a local minimum.

The output of `summary` is quite voluminous. There are 55 parameters, and everything you do will generate a lot of output. `parTable` is the most informative, because it shows the starting values for the numerical search as well as the parameter estimates. Actually, they should not really be called estimates; they are just the place where the search stopped when it ran out of iterations.

```
> parTable(fit1)
   id      lhs op      rhs user block group free ustart exo  label plabel    start        est se
1   1 Fcholest  ~     Fage    1     1     1    1     NA   0 beta11   .p1.    0.000     87.986 NA
2   2 Fcholest  ~     Fbmi    1     1     1    2     NA   0 beta12   .p2.    0.000   1307.372 NA
3   3 Fcholest  ~     Ffat    1     1     1    3     NA   0 beta13   .p3.    0.000   -698.539 NA
4   4 Fdiastol  ~     Fage    1     1     1    4     NA   0 beta21   .p4.    0.000   -696.326 NA
5   5 Fdiastol  ~     Fbmi    1     1     1    5     NA   0 beta22   .p5.    0.000 -10306.408 NA
6   6 Fdiastol  ~     Ffat    1     1     1    6     NA   0 beta23   .p6.    0.000   5531.568 NA
7   7     Fage =~     age1    1     1     1    0      1   0          .p7.    1.000      1.000  0
8   8     Fage =~     age2    1     1     1    0      1   0          .p8.    1.000      1.000  0
9   9     Fbmi =~     bmi1    1     1     1    0      1   0          .p9.    1.000      1.000  0
10 10     Fbmi =~     bmi2    1     1     1    0      1   0         .p10.    1.000      1.000  0
11 11     Ffat =~     fat1    1     1     1    0      1   0         .p11.    1.000      1.000  0
12 12     Ffat =~     fat2    1     1     1    0      1   0         .p12.    1.000      1.000  0
13 13 Fcholest =~ cholest1    1     1     1    0      1   0         .p13.    1.000      1.000  0
14 14 Fcholest =~ cholest2    1     1     1    0      1   0         .p14.    1.000      1.000  0
15 15 Fdiastol =~ diastol1    1     1     1    0      1   0         .p15.    1.000      1.000  0
16 16 Fdiastol =~ diastol2    1     1     1    0      1   0         .p16.    1.000      1.000  0
17 17     Fage ~~     Fage    1     1     1    7     NA   0  phi11  .p17.    0.050    146.894 NA
18 18     Fage ~~     Fbmi    1     1     1    8     NA   0  phi12  .p18.    0.000      3.027 NA
19 19     Fage ~~     Ffat    1     1     1    9     NA   0  phi13  .p19.    0.000     24.137 NA
20 20     Fbmi ~~     Fbmi    1     1     1   10     NA   0  phi22  .p20.    0.050     11.534 NA
21 21     Fbmi ~~     Ffat    1     1     1   11     NA   0  phi23  .p21.    0.000     21.875 NA
22 22     Ffat ~~     Ffat    1     1     1   12     NA   0  phi33  .p22.    0.050     43.806 NA
23 23 Fcholest ~~ Fcholest    1     1     1   13     NA   0  psi11  .p23.    0.050   1439.256 NA
24 24 Fcholest ~~ Fdiastol    1     1     1   14     NA   0  psi12  .p24.    0.000   8682.195 NA
25 25 Fdiastol ~~ Fdiastol    1     1     1   15     NA   0  psi22  .p25.    0.050 -68805.034 NA
26 26     age1 ~~     age1    1     1     1   16     NA   0   w111  .p26.   83.763     18.784 NA
27 27     age1 ~~     bmi1    1     1     1   17     NA   0   w112  .p27.    0.000      4.303 NA
28 28     age1 ~~     fat1    1     1     1   18     NA   0   w113  .p28.    0.000      0.968 NA
29 29     age1 ~~ cholest1    1     1     1   19     NA   0   w114  .p29.    0.000      3.101 NA
30 30     age1 ~~ diastol1    1     1     1   20     NA   0   w115  .p30.    0.000     12.518 NA
31 31     bmi1 ~~     bmi1    1     1     1   21     NA   0   w122  .p31.   10.925      9.257 NA
32 32     bmi1 ~~     fat1    1     1     1   22     NA   0   w123  .p32.    0.000      7.088 NA
33 33     bmi1 ~~ cholest1    1     1     1   23     NA   0   w124  .p33.    0.000     -0.995 NA
34 34     bmi1 ~~ diastol1    1     1     1   24     NA   0   w125  .p34.    0.000     12.457 NA
35 35     fat1 ~~     fat1    1     1     1   25     NA   0   w133  .p35.   30.023     17.494 NA
36 36     fat1 ~~ cholest1    1     1     1   26     NA   0   w134  .p36.    0.000     10.042 NA
37 37     fat1 ~~ diastol1    1     1     1   27     NA   0   w135  .p37.    0.000     -6.847 NA
38 38 cholest1 ~~ cholest1    1     1     1   28     NA   0   w144  .p38. 1548.559    199.263 NA
```

```
39 39 cholest1 ~~ diastol1    1     1     1   29     NA   0   w145  .p39.    0.000     -1.087 NA
40 40 diastol1 ~~ diastol1    1     1     1   30     NA   0   w155  .p40.  162.507    203.809 NA
41 41     age2 ~~     age2    1     1     1   31     NA   0   w211  .p41.   76.888      8.187 NA
42 42     age2 ~~     bmi2    1     1     1   32     NA   0   w212  .p42.    0.000      1.026 NA
43 43     age2 ~~     fat2    1     1     1   33     NA   0   w213  .p43.    0.000     -3.523 NA
44 44     age2 ~~ cholest2    1     1     1   34     NA   0   w214  .p44.    0.000     -4.153 NA
45 45     age2 ~~ diastol2    1     1     1   35     NA   0   w215  .p45.    0.000      5.971 NA
46 46     bmi2 ~~     bmi2    1     1     1   36     NA   0   w222  .p46.    7.156      3.043 NA
47 47     bmi2 ~~     fat2    1     1     1   37     NA   0   w223  .p47.    0.000     -2.598 NA
48 48     bmi2 ~~ cholest2    1     1     1   38     NA   0   w224  .p48.    0.000     -4.928 NA
49 49     bmi2 ~~ diastol2    1     1     1   39     NA   0   w225  .p49.    0.000      7.475 NA
50 50     fat2 ~~     fat2    1     1     1   40     NA   0   w233  .p50.   27.528      9.624 NA
51 51     fat2 ~~ cholest2    1     1     1   41     NA   0   w234  .p51.    0.000    -10.696 NA
52 52     fat2 ~~ diastol2    1     1     1   42     NA   0   w235  .p52.    0.000     -8.263 NA
53 53 cholest2 ~~ cholest2    1     1     1   43     NA   0   w244  .p53. 1608.286    347.391 NA
54 54 cholest2 ~~ diastol2    1     1     1   44     NA   0   w245  .p54.    0.000    -14.307 NA
55 55 diastol2 ~~ diastol2    1     1     1   45     NA   0   w255  .p55.   88.049     59.783 NA
```

Looking at the `est` (estimate) column, one can see the pathology. Some of the values are extremely large, positive and negative. The $\widehat{\beta}_{ij}$ at the beginning of the display are completely unbelievable given the scales of the variables involved. This kind of thing is almost a sure sign that the numerical search has wandered off and gotten lost somewhere far from the actual MLE.

The minus log likelihood functions for structural equation models are characterized by hills and valleys. There can be lots of local maxima and minima. While there will be a deep hole somehere for a sufficiently large sample is the model is correct, the only guarantee of finding it is to start the search close to the hole, where the surface is already sloping down in the right direction. Otherwise, what happens will depend on the detailed topography of the minus log likelihood, and finding the correct MLE is a matter of luck.

What seems to have happened in the present case is that `lavaan`'s default starting values, which often work quite well, were fairly far from the global minimum. The search proceeded downhill but perhaps only slightly downhill after a while[41], off into the distance in an almost featureless plain. It was never going to arrive anywhere, and it's fortunate that there was a limit on the number of iterations.

It took me a while to notice this, but actually the search wandered far outside the parameter space. Look at the `est` value for `psi22`, parameter 25. This is ψ_{22}, the variance of $\epsilon_{i,2}$. The value is -68805.034, an enormous negative variance. There was no warning.

I tried setting boundaries to prevent variances from becoming negative, hoping the search would bounce off the barrier into a better region of the parameter space. I added the following to the model string `bmimodel1`,

```
# Bounds (Variances are positive)
# ------
phi11 > 0; phi22 > 0 ; phi33 > 0
psi11 > 0; psi22 > 0
```

[41] The `verbose = TRUE` option on the `lavaan` statement generated 5,685 lines of output, not shown here. The decrease in the minus log likelihood was more and more gradual.

```
w111 > 0; w122 > 0; w133 > 0; w144 > 0; w155 > 0;
w211 > 0; w222 > 0; w233 > 0; w244 > 0; w255 > 0
```

and then re-ran `lavaan`. The search converged "normally" after 1,245 iterations, but it stuck to the barrier rather than bouncing off, yielding $\widehat{\psi}_{22} = 0$. The estimates for β_{ij} were almost as huge as before, and entirely unrealistic.

The only cure for this disease is better starting values. Commercial software for structural equation modeling uses a deep and sophisticated bag of tricks to pick starting values, and SAS `proc calis` has no trouble fitting this double measurement model. However, as of this writing, `lavaan`'s automatic starting values work okay only most of the time[42].

Here is a way to obtain good starting values for any structural equation model, provided the parameters are identifiable. Recall how the proof of identifiability goes. For any model, the covariance matrix is a function of the model parameters: $\boldsymbol{\Sigma} = g(\boldsymbol{\theta})$. This equality represents the *covariance structure equations*. The parameters that appear in $\boldsymbol{\Sigma}$ are identifiable if the covariance structure equations can be solved to yield $\boldsymbol{\theta} = g^{-1}(\boldsymbol{\Sigma})$. Provided the solution is available explicitly[43], a method of moments estimator is $\widehat{\boldsymbol{\theta}}_M = g^{-1}(\widehat{\boldsymbol{\Sigma}})$, where $\widehat{\boldsymbol{\Sigma}}$ denotes the sample variance-covariance matrix. Typically, the function g^{-1} is continuous in most of the parameter space. In this case, the method of moments estimator is guaranteed to be consistent by the Law of Large Numbers and continuous mapping. Since the MLE is also consistent, it will be close to $\widehat{\boldsymbol{\theta}}_M$ for large samples, and $\widehat{\boldsymbol{\theta}}_M$ should provide an excellent set of starting values.

For double measurement regression, the solution (49) represents $\boldsymbol{\theta} = g^{-1}(\boldsymbol{\Sigma})$. One may start with Expression (51) for $\widehat{\boldsymbol{\Phi}}_M$ and Expression (52) for $\widehat{\boldsymbol{\beta}}_M$ (see page 88), and then use (49) for the rest of the parameters. This is done in the R work below.

```
> # Obtain the MOM estimates to use as starting values.
> head(bmidata)
  age1 bmi1 fat1 cholest1 diastol1 age2 bmi2 fat2 cholest2 diastol2
1   63 24.5 16.5    195.4       38   60 23.9 20.1    203.5       66
2   42 13.0  1.9    184.3       86   44 14.8  2.6    197.3       78
3   32 22.5 14.6    354.1      104   33 21.7 20.4    374.3       73
4   59 25.5 19.0    214.6       93   58 28.5 20.0    203.7      106
5   45 26.5 17.8    324.8       97   43 25.0 12.3    329.7       92
6   31 19.4 17.1    280.7       92   42 19.9 19.9    276.7       87

> W1 = as.matrix(bmidata[,1:3])  # age1 bmi1 fat1
> V1 = as.matrix(bmidata[,4:5])  # cholest1 diastol1
> W2 = as.matrix(bmidata[,6:8])  # age2 bmi2 fat2
> V2 = as.matrix(bmidata[,9:10]) # cholest2 diastol2
```

[42] I'm not complaining. I am deeply grateful for `lavaan`, and if I want better starting values I should develop the software myself. To me, this is not the most interesting project in the world, so it is on the back burner.

[43] For some models, an explicit solution is hard to obtain, even if you can prove it exists. That's the main obstacle to automating this process.

```
> var(W1,W2) # Matrix of sample covariances
           age2      bmi2     fat2
age1 148.220782  3.621581 25.29808
bmi1   5.035726 13.194016 21.42201
fat1  23.542289 20.613490 45.13296

> # Using S as short for Sigmahat, and not worrying about n vs. n-1,
> S11 = var(W1); S12 = var(W1,V1); S13 = var(W1,W2); S14 = var(W1,V2)
>                S22 = var(V1);    S23 = var(V1,W2); S24 = var(V1,V2)
>                                  S33 = var(W2);    S34 = var(W2,V2)
>                                                    S44 = var(V2)
> # The matrices below should all have "hat" in the name, because they are estimates
> Phi = (S13+t(S13))/2
> rownames(Phi) = colnames(Phi) = c('Fage','Fbmi','Ffat'); Phi
           Fage      Fbmi     Ffat
Fage 148.220782  4.328654 24.42019
Fbmi   4.328654 13.194016 21.01775
Ffat  24.420185 21.017749 45.13296

> Beta = 0.5*(t(S14)+S23) %*% solve(Phi)
> rownames(Beta) = c('Fcholest','Fdiastol')
> colnames(Beta) = c('Fage','Fbmi','Ffat'); Beta
              Fage       Fbmi     Ffat
Fcholest -0.3851327 -0.1885072 2.968322
Fdiastol  0.0224190 -0.3556138 1.407425

> Psi = S24 - Beta %*% Phi %*% t(Beta)
> rownames(Psi) = colnames(Psi) = c('Fcholest','Fdiastol') # epsilon1, epsilon2
> Psi
           Fcholest  Fdiastol
Fcholest 2548.17303 -44.56069
Fdiastol  -28.70087  57.64153

> # Oops, it should be symmetric.
> Psi = ( Psi+t(Psi) )/2; Psi
           Fcholest  Fdiastol
Fcholest 2548.17303 -36.63078
Fdiastol  -36.63078  57.64153

> Omega11 = S11 - Phi; Omega11
          age1     bmi1      fat1
age1 19.640040 4.610807  1.634183
bmi1  4.610807 8.699533  8.754484
fat1  1.634183 8.754484 15.033932
```

```
> Omega12 = S12 - ( S14+t(S23) )/2; Omega12
      cholest1  diastol1
age1  4.499017 12.164192
bmi1 -1.517733 10.671443
fat1  3.888565 -2.196681

> Omega22 = S22-S24 # A little rough but consistent
> Omega22 = (Omega22 + t(Omega22)  )/2
> Omega22
          cholest1  diastol1
cholest1 213.76117  11.24971
diastol1  11.24971 196.44520

> Omega33 = S33 - Phi; Omega33
          age2      bmi2      fat2
age2  5.862661 -1.219843 -2.155736
bmi2 -1.219843  1.146991 -1.714769
fat2 -2.155736 -1.714769 10.033984

> Omega34 = S34 - ( S14+t(S23) )/2; Omega34
      cholest2   diastol2
age2 -2.978041  0.7795992
bmi2 -1.206256  2.1081739
fat2 -6.422983 -4.9125882

> Omega44 = S44 - S24 ; Omega44 = ( Omega44 + t(Omega44) )/2
> Omega44
          cholest2  diastol2
cholest2 333.45335 -21.65923
diastol2 -21.65923  47.23065

> round(Beta,3)
           Page   Fbmi  Ffat
Fcholest -0.385 -0.189 2.968
Fdiastol  0.022 -0.356 1.407
```

Please look at the last set of numbers. It is worth noting how far these method-of-moments estimates are from the stopping place of the first numerical search. Repeating earlier material for comparison, ...

```
   id      lhs op      rhs user block group free ustart exo  label plabel    start        est se
1   1 Fcholest  ~     Page    1     1     1    1     NA   0 beta11   .p1.    0.000     87.986 NA
2   2 Fcholest  ~     Fbmi    1     1     1    2     NA   0 beta12   .p2.    0.000   1307.372 NA
3   3 Fcholest  ~     Ffat    1     1     1    3     NA   0 beta13   .p3.    0.000   -698.539 NA
4   4 Fdiastol  ~     Page    1     1     1    4     NA   0 beta21   .p4.    0.000   -696.326 NA
```

```
5   5 Fdiastol  ~     Fbmi    1     1     1    5     NA   0 beta22   .p5.    0.000 -10306.408 NA
6   6 Fdiastol  ~     Ffat    1     1     1    6     NA   0 beta23   .p6.    0.000   5531.568 NA
```

While the method-of-moments estimates are promising as starting values, there is no doubt that entering them all manually is a major pain. I was motivated and I was confident it would work, so I did it. The model string is given below. As in Example 0.10.3, variables appear twice, once to specify the parameter name and a second time to specify the starting value.

```
> bmimodel2 =
+     #
+     # Latent variable model
+     # ---------------------
+         'Fcholest ~ beta11*Fage        + beta12*Fbmi        + beta13*Ffat +
+                     start(-0.385)*Fage + start(-0.189)*Fbmi + start(2.968)*Ffat
+          Fdiastol ~ beta21*Fage        + beta22*Fbmi        + beta23*Ffat +
+                     start(0.022)*Fage  + start(-0.356)*Fbmi + start(1.407)*Ffat
+     #
+     # Measurement model
+     # -----------------
+         Fage =~ 1*age1 + 1*age2
+         Fbmi =~ 1*bmi1 + 1*bmi2
+         Ffat =~ 1*fat1 +1*fat2
+         Fcholest =~ 1*cholest1 + 1*cholest2
+         Fdiastol =~ 1*diastol1 + 1*diastol2
+     #
+     # Variances and covariances
+     # -------------------------
+     # Of latent explanatory variables
+         Fage ~~ phi11*Fage + start(148.220782)*Fage
+         Fage ~~ phi12*Fbmi + start(4.328654)*Fbmi
+         Fage ~~ phi13*Ffat + start(24.42019)*Ffat
+         Fbmi ~~ phi22*Fbmi + start(13.194016)*Fbmi
+         Fbmi ~~ phi23*Ffat + start(21.01775)*Ffat
+         Ffat ~~ phi33*Ffat + start(45.13296)*Ffat
+     # Of error terms in latent the regression (epsilon_ij)
+         Fcholest ~~ psi11*Fcholest + start(2548.17303)*Fcholest
+         Fcholest ~~ psi12*Fdiastol + start(-36.63078)*Fdiastol
+         Fdiastol ~~ psi22*Fdiastol + start(57.64153)*Fdiastol
+     # Of measurement errors (e_ijk) for measurement set 1
+         age1 ~~ w111*age1 + start(19.640040)*age1
+         age1 ~~ w112*bmi1 + start(4.610807)*bmi1
+         age1 ~~ w113*fat1 + start(1.634183)*fat1
+         age1 ~~ w114*cholest1 + start(4.499017)*cholest1
+         age1 ~~ w115*diastol1 + start(12.164192)*diastol1
```

```
+         bmi1 ~~ w122*bmi1 + start(8.699533)*bmi1
+         bmi1 ~~ w123*fat1 + start(8.754484)*fat1
+         bmi1 ~~ w124*cholest1 + start(-1.517733)*cholest1
+         bmi1 ~~ w125*diastol1 + start(10.671443)*diastol1
+         fat1 ~~ w133*fat1 + start(15.033932)*fat1
+         fat1 ~~ w134*cholest1 + start(3.888565)*cholest1
+         fat1 ~~ w135*diastol1 + start(-2.196681)*diastol1
+         cholest1 ~~ w144*cholest1 + start(213.76117)*cholest1
+         cholest1 ~~ w145*diastol1 + start(11.24971)*diastol1
+         diastol1 ~~ w155*diastol1 + start(196.44520)*diastol1
+     # Of measurement errors (e_ijk) for measurement set 2
+         age2 ~~ w211*age2 + start(5.862661)*age2
+         age2 ~~ w212*bmi2 + start(-1.219843)*bmi2
+         age2 ~~ w213*fat2 + start(-2.155736)*fat2
+         age2 ~~ w214*cholest2 + start(-2.978041)*cholest2
+         age2 ~~ w215*diastol2 + start(0.7795992)*diastol2
+         bmi2 ~~ w222*bmi2 + start(1.146991)*bmi2
+         bmi2 ~~ w223*fat2 + start(-1.714769)*fat2
+         bmi2 ~~ w224*cholest2 + start(-1.206256)*cholest2
+         bmi2 ~~ w225*diastol2 + start(2.1081739)*diastol2
+         fat2 ~~ w233*fat2 + start(10.033984)*fat2
+         fat2 ~~ w234*cholest2 + start(-6.422983)*cholest2
+         fat2 ~~ w235*diastol2 + start(-4.9125882)*diastol2
+         cholest2 ~~ w244*cholest2 + start(333.45335)*cholest2
+         cholest2 ~~ w245*diastol2 + start(-21.65923)*diastol2
+         diastol2 ~~ w255*diastol2 + start(47.23065)*diastol2
+     # Bounds (Variances are positive)
+     # ------
+         phi11 > 0; phi22 > 0 ; phi33 > 0
+         psi11 > 0; psi22 > 0
+         w111 > 0; w122 > 0; w133 > 0; w144 > 0; w155 > 0;
+         w211 > 0; w222 > 0; w233 > 0; w244 > 0; w255 > 0
+     ' ################## End of bmimodel2 ##################
> fit2 = lavaan(bmimodel2, data=bmidata)
> summary(fit2)
lavaan (0.5-23.1097) converged normally after 327 iterations

  Number of observations                           500

  Estimator                                         ML
  Minimum Function Test Statistic                4.654
  Degrees of freedom                                10
  P-value (Chi-square)                           0.913
```

```
Parameter Estimates:

  Information                                 Expected
  Standard Errors                             Standard

Latent Variables:
                   Estimate  Std.Err  z-value  P(>|z|)
  Fage =~
    age1              1.000
    age2              1.000
  Fbmi =~
    bmi1              1.000
    bmi2              1.000
  Ffat =~
    fat1              1.000
    fat2              1.000
  Fcholest =~
    cholest1          1.000
    cholest2          1.000
  Fdiastol =~
    diastol1          1.000
    diastol2          1.000

Regressions:
                   Estimate  Std.Err  z-value  P(>|z|)
  Fcholest ~
    Fage    (bt11)   -0.320    0.228   -1.404    0.160
    Fbmi    (bt12)    0.393    1.708    0.230    0.818
    Ffat    (bt13)    2.774    0.980    2.829    0.005
  Fdiastol ~
    Fage    (bt21)    0.020    0.050    0.407    0.684
    Fbmi    (bt22)   -0.480    0.419   -1.145    0.252
    Ffat    (bt23)    1.480    0.235    6.312    0.000

Covariances:
                   Estimate  Std.Err  z-value  P(>|z|)
  Fage ~~
    Fbmi    (ph12)    4.161    2.141    1.944    0.052
    Ffat    (ph13)   23.321    3.986    5.851    0.000
  Fbmi ~~
    Ffat    (ph23)   20.976    1.584   13.244    0.000
 .Fcholest ~~
   .Fdiastl (ps12)  -45.870   24.969   -1.837    0.066
 .age1 ~~
```

```
   .bmi1    (w112)    3.998    0.945    4.231    0.000
   .fat1    (w113)    2.389    1.505    1.587    0.112
   .cholst1 (w114)    2.705    9.091    0.297    0.766
   .diastl1 (w115)   10.562    3.824    2.762    0.006
 .bmi1 ~~
   .fat1    (w123)    8.968    0.956    9.382    0.000
   .cholst1 (w124)   -0.888    4.178   -0.212    0.832
   .diastl1 (w125)   10.060    2.274    4.424    0.000
 .fat1 ~~
   .cholst1 (w134)    7.916    6.741    1.174    0.240
   .diastl1 (w135)   -2.928    3.409   -0.859    0.390
 .cholest1 ~~
   .diastl1 (w145)   -0.107   16.907   -0.006    0.995
 .age2 ~~
   .bmi2    (w212)   -0.661    0.735   -0.899    0.369
   .fat2    (w213)   -2.703    1.369   -1.974    0.048
   .cholst2 (w214)   -1.964    8.962   -0.219    0.827
   .diastl2 (w215)    2.274    2.710    0.839    0.401
 .bmi2 ~~
   .fat2    (w223)   -1.849    0.705   -2.624    0.009
   .cholst2 (w224)   -2.650    3.476   -0.762    0.446
   .diastl2 (w225)    2.652    1.487    1.784    0.074
 .fat2 ~~
   .cholst2 (w234)  -11.370    6.546   -1.737    0.082
   .diastl2 (w235)   -4.839    2.536   -1.908    0.056
 .cholest2 ~~
   .diastl2 (w245)   -8.964   12.605   -0.711    0.477

Variances:
                   Estimate  Std.Err  z-value  P(>|z|)
    Fage    (ph11)  147.330    9.699   15.190    0.000
    Fbmi    (ph22)   13.341    0.986   13.528    0.000
    Ffat    (ph33)   44.485    3.101   14.345    0.000
   .Fcholst (ps11) 2534.507  171.258   14.799    0.000
   .Fdiastl (ps22)   56.169    9.221    6.092    0.000
   .age1    (w111)   18.584    2.914    6.378    0.000
   .bmi1    (w122)    8.665    0.708   12.239    0.000
   .fat1    (w133)   16.124    1.659    9.717    0.000
   .cholst1 (w144)  200.103   57.422    3.485    0.000
   .diastl1 (w155)  195.040   14.323   13.617    0.000
   .age2    (w211)    6.861    2.701    2.540    0.011
   .bmi2    (w222)    1.089    0.491    2.220    0.026
   .fat2    (w233)    9.332    1.539    6.065    0.000
   .cholst2 (w244)  344.454   60.290    5.713    0.000
```

```
   .diastl2 (w255)    48.350    8.246    5.864    0.000

Constraints:
                                                |Slack|
    phi11 - 0                                   147.330
    phi22 - 0                                    13.341
    phi33 - 0                                    44.485
    psi11 - 0                                  2534.507
    psi22 - 0                                    56.169
    w111 - 0                                     18.584
    w122 - 0                                      8.665
    w133 - 0                                     16.124
    w144 - 0                                    200.103
    w155 - 0                                    195.040
    w211 - 0                                      6.861
    w222 - 0                                      1.089
    w233 - 0                                      9.332
    w244 - 0                                    344.454
    w255 - 0                                     48.350
```

With these starting values, the maximum likelihood search converged after 327 iterations. The likelihood ratio chi-squared test of model fit indicated no problems: $G^2 = 4.654, df = 10, p = 0.913$. Primary interest is in the relationship of latent (true) BMI to latent cholesterol level and latent blood pressure, controlling for latet aga and latent percent body fat. When measurmtn error was taken into account using double measurement, neither relationship was statistically significant at the 0.05 level. For cholesterol, $Z = 0.230$ and $p = 0.818$. For diastolic blood pressure, $Z = -1.145$ and $p = 0.252$. This is in contrast to the conclusion from naive ordinary least squarees regression, which was that controlling for age and percent body fat, higher BMI was associated with higher average diastolic blood pressure. Brunner and Austin (1992; also see Section 0.7) have shown how this kind of "even controlling for" conclusion is the kind of error that tends to creep in with ordinary regression, when the explanatory variables are measured with error. Double measurement regression has more credibility.

Plenty more tests based on this model are possible and worthwhile, but BMI controlling for age and percent body fat is the main issue. Just as a demonstration, I will illustrate one more test, a likelihood rati test of BMI controlling for age and percent body fat, for cholesterol and diastolic blood pressure simultaneously. The null hypothesis is $H_0 : \beta_{21} = \beta_{22} = 0$. We begin by fitting a restricted model[44].

```
> nobmi = lavaan(bmimodel2, data=bmidata,
+             constraints = 'beta12 == 0
+                            beta22 == 0')
```

[44]It is a relief that the non-zero starting values for β_{21} and β_{22} in `bmimodel2` do not conflict with the constraint that sets them equal to zero.

```
>
> anova(nobmi,fit2)
Chi Square Difference Test

      Df   AIC   BIC  Chisq Chisq diff Df diff Pr(>Chisq)
fit2  10 35758 35947 4.6537
nobmi 12 35755 35936 6.1457      1.492       2     0.4743
```

Again, the conclusion is that allowing for age and percent body fat, there is no evidence of a connection between BMI and either health indicator.

0.11 Extra Response Variables

Sometimes, double measurement is not a practical alternative. Perhaps 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.

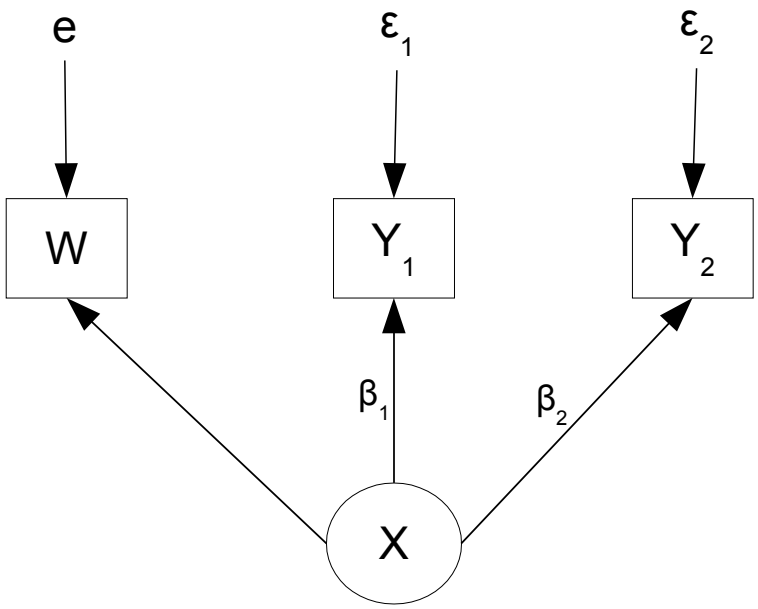
In such situations, all one can do is to use what information happens to be available. While most research studies will not contain multiple measurements of the explanatory variables, they often will have quite a few possible response variables. These variables might already be part of the data set, or possibly the researchers could go back and collect them without an unbearable amount of effort. It helps if these extra response variables are from a different domain than the response variable of interest, so one can make a case that the extra variables and the response variables of interest are not affected by common omitted variables. In the path diagrams, this is represented by the absence of curved, double-headed arrows connecting error terms. It is a critical part of the recipe.

One explanatory variable

In a simple measurement error regression model like the one in Example 0.8.1, suppose that we have access to data for a second response variable that depends on the latent explanatory variable X_i. Our main interest is still in the response variable Y_i. The other response variable may or may not be interesting in its own right; it is included as a way of getting around the identifiability problem.

Example 0.11.1 *One Extra Response Variable*

Figure 17: Y_2 is an extra response variable



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

$$\begin{aligned} W_i &= \nu + X_i + e_i \\ Y_{i,1} &= \alpha_1 + \beta_1 X_i + \epsilon_{i,1} \\ Y_{i,2} &= \alpha_2 + \beta_2 X_i + \epsilon_{i,2} \end{aligned} \tag{53}$$

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. Figure 17 shows a path diagram of this model.

It is usually helpful to check the Parameter Count Rule (Rule 1 on page 61) before doing detailed calculations. For this model, there are ten parameters: $\boldsymbol{\theta} = (\nu, \alpha_1, \alpha_2, \beta_1, \beta_2, \mu_x, \phi, \omega, \psi_1, \psi_2)$. Writing the vector of observable data for case i as $\mathbf{D}_i = (W_i, Y_{i,1}, Y_{i,2})^\top$, we see that $\boldsymbol{\mu} = E(\mathbf{D}_i)$ has three elements and $\boldsymbol{\Sigma} = cov(\mathbf{D}_i)$ has $3(3+1)/2 = 6$ unique elements. Thus identifiability of the entire parameter vector is ruled out in most of the parameter space. 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 are pessimistic about identifying expected values and intercepts. So consider first the covariance matrix. Elements of $\boldsymbol{\Sigma} = cov(\mathbf{D}_i)$ 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)

$$\begin{aligned} Cov(W_i, Y_{i,i}) &= Cov(X_i + e_i,\ \beta_1 X_i + \epsilon_{i,1}) \\ &= \beta_1 Cov(X_i, X_i) + Cov(X_i, \epsilon_{i,1}) + \beta_1 Cov(e_i, X_i) + Cov(e_i, \epsilon_{i,1}) \\ &= \beta_1 Var(X) + 0 + 0 + 0 \\ &= \beta_1 \phi \end{aligned}$$

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 in the parameter space.

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 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 extra variable is related to the response variable. Otherwise, why include it[45]?

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:

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

[45]Moreover, one can rule out $\beta_1 = 0$ by a routine test of the correlation between W and Y_2. This kind of test is very helpful (assuming the data are in hand), because for successful inference it's not necessary for the entire parameter to be identifiable everywhere in the parameter space. It's only necessary for the interesting part of the parameter vector to be identifiable in the region of the parameter space where the true parameter is located.

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

$$\begin{aligned} \mu_1 &= \nu + \mu_x \\ \mu_2 &= \alpha_1 + \beta_1 \mu_x \\ \mu_3 &= \alpha_2 + \beta_2 \mu_x \end{aligned} \tag{54}$$

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 (54). 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

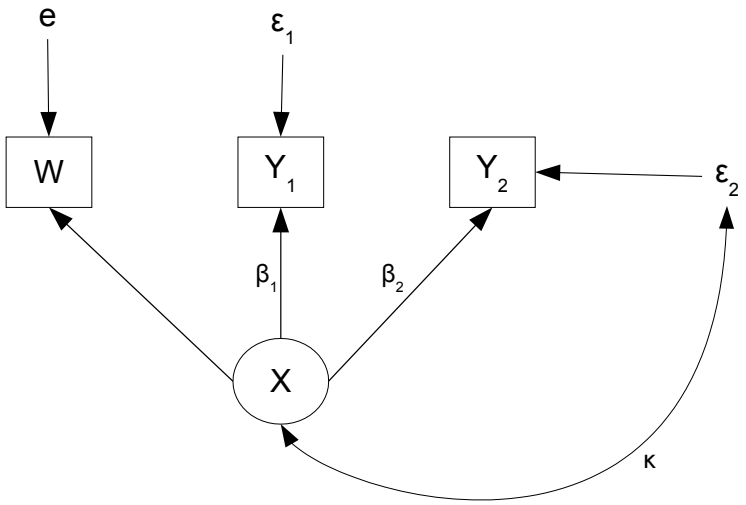
$$\boldsymbol{\theta} = (\mu_1, \mu_2, \mu_3, \beta_1, \beta_2, \phi, \omega, \psi_1, \psi_2).$$

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 (32) on page 46, 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. 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* of the parameters that are identifiable, and at the same time allow valid inference about the original parameters that are of primary interest.

Example 0.11.1 is very interesting for another reason. The purpose of all this is to test $H_0 : \beta_1 = 0$, but even if an assumption of normality is justified, the usual normal theory tests will break down. Though β_1 is identifiable when the null hypothesis is true, the entire parameter vector is not. It will be impossible to fit the restricted model needed for a likelihood ratio test, because infinitely many sets $(\beta_2, \phi, \psi_2, \omega)$ yield the same covariance matrix when $\beta_1 = 0$. The Wald test will suffer too. The Hessian (see Section A.6.3 in Appendix A) will be nearly singular at the unrestricted MLE, resulting in huge standard errors. Again because of the nearly singular Hessian, numerical problems in locating the unrestricted MLE are likely. These problem will be worse for larger sample sizes, as the point where the likelihood function happens to be highest approaches the level region where the parameter is not identifiable at all.

Figure 18: Error term correlated with the explanatory variable



There is a general lesson here, and a way out in this particular case. The general lesson is to re-verify parameter identifiability when the null hypothesis is true, bearing in mind that likelihood methods depend on identifiability of the entire parameter vector[46]. It is better to anticipate trouble and avoid it than to be confused by it once it happens.

As for the way out of this haunted house, note that if $\beta_2 \neq 0$, the null hypothesis $\beta_1 = 0$ is true if and only if $\sigma_{12} = \sigma_{23} = 0$. This null hypothesis can be tested using a generic multivariate normal model. The likelihood ratio test, like the Wald test, will have two degrees of freedom. If the normal assumption is a source of discomfort, testing a couple of Spearman rank correlations with a Bonferroni correction is an alternative. More generally, we will see shortly that having more than one extra response variable can yield identifiability whether or not $H_0 : \beta_1 = 0$ is true. This is a better solution if it's possible, because it makes the analysis more routine.

Example 0.11.2 *Correlation between explanatory variables and error terms*

Recalling Section 0.4 on omitted variables in regression, it is remarkable that while the explanatory variable X_i must not be correlated with the error term $\epsilon_{i,1}$, the error term $\epsilon_{i,2}$ (corresponding to the extra variable $Y_{i,2}$) is allowed to be correlated with X_i, perhaps reflecting the operation of omitted explanatory variables that affect $Y_{i,2}$ and have non-zero covariance with X_i. Figure 18 shows a path diagram of this model.

Suppose $Cov(X_i, \epsilon_{i,2}) = \kappa$, which might be non-zero. This means that seven unknown parameters appear in the six covariance structure equations, and the Parameter Count Rule warns us that it will be impossible to identify them all. Proceeding anyway, the

[46] This is true of any estimation method based on a numerical search.

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 extra variable so that $\beta_2 \neq 0$,

$$\frac{\sigma_{23}}{\sigma_{13}} = \frac{\beta_1(\beta_2\phi+\kappa)}{\beta_2\phi+\kappa} = \beta_1. \tag{55}$$

In fact, if $\kappa \neq 0$, we don't even need $\beta_2 \neq 0$ to identify β_1. That is, the extra variable does not need be influenced by the explanatory variable. It need only be influenced by some unknown variable or variables that are *correlated* with the explanatory variable. Far from being a problem in this case, the omitted variables made it easier to get at β_1.

As in Example 0.11.1, testing $H_0 : \beta_1 = 0$ is non-standard because while β_1 is identifiable, the entre parameter vector is not. We can deal with this kind of complication if we need to, but everything is much easier with more than one extra variable.

Example 0.11.3 *More Than One Extra 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 extra variables. Now the model is, independently for $i = 1, \ldots, n$,

$$\begin{aligned} W_i &= \nu + X_i + e_i \\ Y_{i,1} &= \alpha_1 + \beta_1 X_i + \epsilon_{i,1} \\ Y_{i,2} &= \alpha_2 + \beta_2 X_i + \epsilon_{i,2} \\ Y_{i,3} &= \alpha_3 + \beta_3 X_i + \epsilon_{i,3}, \end{aligned} \tag{56}$$

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 $\mathbf{D}_i = (W_i, Y_{i,1}, Y_{i,2}, Y_{i,3})^\top$,

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

and

$$\boldsymbol{\Sigma} = \begin{pmatrix} \phi+\omega & \beta_1\phi & \beta_2\phi & \beta_3\phi \\ & \beta_1^2\phi+\psi_1 & \beta_1\beta_2\phi & \beta_1\beta_3\phi \\ & & \beta_2^2\phi+\psi_2 & \beta_2\beta_3\phi \\ & & & \beta_3^2\phi+\psi_3 \end{pmatrix}. \tag{57}$$

As before, it is impossible to identify the intercepts and expected values, so we re-parameterize, absorbing them into a vector of expected values which we estimate with the corresponding vector of sample means and then never mention again.

To establish identifiability of the parameters that appear in the covariance matrix, the task is to solve the following ten equations for the eight unknown parameters ϕ, ω, β_1, β_2, β_3, ψ_1, ψ_2, and ψ_3:

$$\begin{aligned}
\sigma_{11} &= \phi + \omega \\
\sigma_{12} &= \beta_1\phi \\
\sigma_{13} &= \beta_2\phi \\
\sigma_{14} &= \beta_3\phi \\
\sigma_{22} &= \beta_1^2\phi + \psi_1 \\
\sigma_{23} &= \beta_1\beta_2\phi \\
\sigma_{24} &= \beta_1\beta_3\phi \\
\sigma_{33} &= \beta_2^2\phi + \psi_2 \\
\sigma_{34} &= \beta_2\beta_3\phi \\
\sigma_{44} &= \beta_3^2\phi + \psi_3
\end{aligned} \tag{58}$$

Assuming the extra variables are well-chosen so that both β_2 and β_3 are both non-zero,

$$\frac{\sigma_{13}\sigma_{14}}{\sigma_{34}} = \frac{\beta_2\beta_3\phi^2}{\beta_2\beta_3\phi} = \phi. \tag{59}$$

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

$$\begin{aligned}
\phi &= \frac{\sigma_{13}\sigma_{14}}{\sigma_{34}} \\
\omega &= \sigma_{11} - \frac{\sigma_{13}\sigma_{14}}{\sigma_{34}} \\
\beta_1 &= \frac{\sigma_{12}\sigma_{34}}{\sigma_{13}\sigma_{14}} \\
\beta_2 &= \frac{\sigma_{34}}{\sigma_{14}} \\
\beta_3 &= \frac{\sigma_{34}}{\sigma_{13}} \\
\psi_1 &= \sigma_{22} - \frac{\sigma_{12}^2\sigma_{34}}{\sigma_{13}\sigma_{14}} \\
\psi_2 &= \sigma_{33} - \frac{\sigma_{13}\sigma_{34}}{\sigma_{14}} \\
\psi_3 &= \sigma_{44} - \frac{\sigma_{14}\sigma_{34}}{\sigma_{13}}
\end{aligned} \tag{60}$$

This proves identifiability at all points in the parameter space where $\beta_2 \neq 0$ and $\beta_3 \neq 0$. The extra variables Y_2 and Y_3 have been chosen so as to guarantee this, and in any case the assumption is testable.

The solution (60) 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 (57), saying "Denote the i, j element of $\boldsymbol{\Sigma}$ by σ_{ij}," skipping the system of equations (58) because they are present in (57), and showing the solution for ϕ in (59), *including* the stipulation that β_2 and β_3 are both non-zero. Then, instead of the explicit solution (60), 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 (like method-of-moments estimation).

With two (or more) extra 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 of Example 0.11.3 imposes equality constraints on the covariance matrix of the observable data. In the solution given by (60), 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 (58) 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

$$\sigma_{12}\sigma_{34} = \sigma_{14}\sigma_{23} = \sigma_{13}\sigma_{24}. \tag{61}$$

Since all three products equal $\beta_1\beta_2\beta_3\phi^2$, it is clear that the model implies the equality constraints (61) 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[47] 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 (61). 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 usually imply *inequality* constraints on the variances and covariances, whether the parameters are identifiable or not. For example, in (60), $\phi = \frac{\sigma_{13}\sigma_{14}}{\sigma_{34}}$. Because ϕ is a variance, we have the inequality restriction $\frac{\sigma_{13}\sigma_{14}}{\sigma_{34}} > 0$, something that is not automatically true 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.

Multiple explanatory variables

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

Example 0.11.4 *Two explanatory variables and two extra response variables*

Here is an example with two explanatory variables and a single extra response variable for each one. Independently for $i = 1, \ldots, n$,

$$\begin{array}{rcl} W_{i,1} &=& \nu_1 + X_{i,1} + e_{i,1} \\ Y_{i,1} &=& \alpha_1 + \beta_1 X_{i,1} + \epsilon_{i,1} \\ Y_{i,2} &=& \alpha_2 + \beta_2 X_{i,1} + \epsilon_{i,2} \\ W_{i,2} &=& \nu_2 + X_{i,2} + e_{i,2} \\ Y_{i,3} &=& \alpha_3 + \beta_3 X_{i,2} + \epsilon_{i,3} \\ Y_{i,4} &=& \alpha_4 + \beta_4 X_{i,2} + \epsilon_{i,4} \end{array} \tag{62}$$

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

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

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

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

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 $\mathbf{D}_i = (W_{i,1}, Y_{i,1}, Y_{i,2}, W_{i,2}, Y_{i,3}, Y_{i,4})^\top$, the covariance matrix $\boldsymbol{\Sigma} = cov(\mathbf{D}_i)$ 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[48] 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 extra 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.

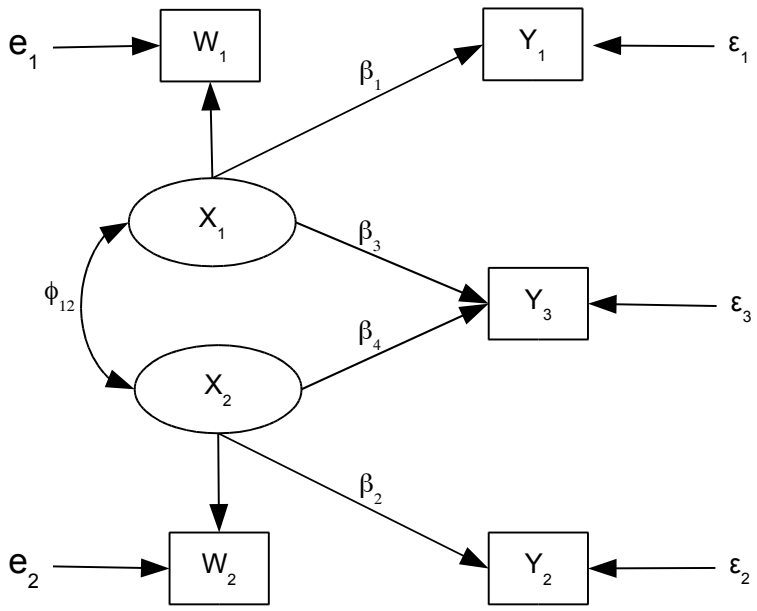
Example 0.11.5 *Two explanatory variables, one response variable of primary interest, and one extra response variable for each explanatory variable.*

In this example, each explanatory variable has its own extra variable, but they share a response variable of primary interest. This is more interesting, because now one can speak of one explanatory variable *controlling* for the other, as in ordinary regression. Figure 19 shows the path diagram.

The formal statement of this model dispenses with intercepts and expected values. They are really present, but because they are not identifiable separately, they are not

[48] 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.

Figure 19: Two explanatory variables with one extra response variable each, plus a single response variable of interest



even mentioned. This is common in structural equation modeling. Independently for $i = 1, \ldots, n$, let

$$\begin{aligned} W_{i,1} &= X_{i,1} + e_{i,1} \\ W_{i,2} &= X_{i,2} + e_{i,2} \\ Y_{i,1} &= \beta_1 X_{i,1} + \epsilon_{i,1} \\ Y_{i,2} &= \beta_2 X_{i,2} + \epsilon_{i,2} \\ Y_{i,3} &= \beta_3 X_{i,1} + \beta_4 X_{i,2} + \epsilon_{i,3} \end{aligned}$$

where

- The $X_{i,j}$ variables are latent, while the $W_{i,j}$ and $Y_{i,j}$ variables are observable.
- $e_{i,1} \sim N(0, \omega_1)$ and $e_{i,2} \sim N(0, \omega_2)$.
- $\epsilon_{i,j} \sim N(0, \psi_j)$ for $j = 1, 2, 3$.
- $e_{i,j}$ and $\epsilon_{i,j}$ are independent of each other and of $X_{i,j}$.
- $X_{i,j}$ have covariance matrix

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

Denote the vector of observable data by $\mathbf{D}_i = (W_{i,1}, Y_{i,1}, W_{i,2}, Y_{i,2}, Y_{i,3})^\top$, with $cov(\mathbf{D}_i) = \mathbf{\Sigma} = [\sigma_{ij}]$.

Among other things, this example illustrates how the search for identifiability can be supported by exploratory data analysis. Hypotheses about *single* covariances, like $H_0 : \sigma_{ij} = 0$ can be tested by looking at tests of the corresponding correlations. These tests, including non-parametric tests based on the Spearman rank correlation, are easily obtained using the `cor.test` function.

The parameter vector[49] for this problem is $\boldsymbol{\theta} = (\phi_{11}, \phi_{12}, \phi_{22}, \omega_1, \omega_2, \beta_1, \beta_2, \beta_3, \beta_4, \psi_1, \psi_2, \psi_3)^\top$. There are 12 parameters and 5 observable variables, so that the covariance matrix has $5(5+1)/2 = 15$ unique variances and covariances. Thus there are 15 covariance structure equations in 12 unknowns, and the parameter count rule tells us that identifiability in most of the parameter space is possible but not guaranteed.

The matrix equation 63 shows the covariance structure equations in a compact form.

$$\begin{pmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} & \sigma_{14} & \sigma_{15} \\ & \sigma_{22} & \sigma_{23} & \sigma_{24} & \sigma_{25} \\ & & \sigma_{33} & \sigma_{34} & \sigma_{35} \\ & & & \sigma_{44} & \sigma_{45} \\ & & & & \sigma_{55} \end{pmatrix} = \tag{63}$$

$$\begin{pmatrix} \omega_1 + \phi_{11} & \beta_1\phi_{11} & \phi_{12} & \beta_2\phi_{12} & \beta_3\phi_{11} + \beta_4\phi_{12} \\ & \beta_1^2\phi_{11} + \psi_1 & \beta_1\phi_{12} & \beta_1\beta_2\phi_{12} & \beta_1(\beta_3\phi_{11} + \beta_4\phi_{12}) \\ & & \omega_2 + \phi_{22} & \beta_2\phi_{22} & \beta_3\phi_{12} + \beta_4\phi_{22} \\ & & & \beta_2^2\phi_{22} + \psi_2 & \beta_2(\beta_3\phi_{12} + \beta_4\phi_{22}) \\ & & & & (\beta_3\phi_{11} + \beta_4\phi_{12})\beta_3 + (\beta_3\phi_{12} + \beta_4\phi_{22})\beta_4 + \psi_3 \end{pmatrix}$$

In our study of identifiability for this example, we will confine our attention to that part of the parameter space where $\beta_1 \neq 0$ and $\beta_2 \neq 0$. After all, the variables Y_1 and Y_2 were introduced only to help with identifiability, and they are useless unless they are related to the explanatory variables. The issue may be resolved empirically by testing $H_0 : \sigma_{23} = 0$ and $H_0 : \sigma_{14} = 0$ with `cor.test`. One should proceed to model fitting only if both null hypotheses are comfortably rejected. In any case, the rest of this discussion assumes that β_1 and β_2 are both non-zero.

The parameter ϕ_{12} is identifiable, since $\phi_{12} = \sigma_{13}$. Consider two cases. The first case is $\phi_{12} \neq 0$. In this region of the parameter space, β_1 is identified from $\beta_1 = \sigma_{23}/\phi_{12}$, and β_2 is identified from $\beta_2 = \sigma_{14}/\phi_{12}$. Then, $\phi_{11} = \sigma_{12}/\beta_1$ and $\phi_{22} = \sigma_{34}/\beta_2$.

With ϕ_{11}, ϕ_{12} and ϕ_{22} identified, they may be treated as known. Then, β_3 and β_4 are identified from σ_{14} and σ_{34} by solving two linear equations in two unknowns. Writing the equations in matrix form,

$$\begin{pmatrix} \phi_{11} & \phi_{12} \\ \phi_{12} & \phi_{22} \end{pmatrix} \begin{pmatrix} \beta_3 \\ \beta_4 \end{pmatrix} = \begin{pmatrix} \sigma_{14} \\ \sigma_{34} \end{pmatrix}.$$

The solution exists if and only if the covariance matrix of the latent explanatory variables has an inverse, which is not much to ask. At this point, all parameters have been identified

[49] That is, the vector of parameters appearing in $\mathbf{\Sigma} = cov(\mathbf{D}_i)$.

except the variances of the e_{ij} and ϵ_{ij}. Accordingly, ω_1, ψ_1, ω_2, ψ_2 and ψ_3 are obtained from the diagonal elements of $\boldsymbol{\Sigma}$, by subtraction. The conclusion is that all parameters are identifiable provided $\phi_{12} \neq 0$. In most observational studies, explanatory variables will be correlated. That means the parameters of this model are identifiable for most applications.

Now consider the case where $\phi_{12} = 0$; that is, the latent explanatory variables are uncorrelated. This might apply in a designed experiment with random assignment. The covariance structure equations are now

$$\begin{pmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} & \sigma_{14} & \sigma_{15} \\ & \sigma_{22} & \sigma_{23} & \sigma_{24} & \sigma_{25} \\ & & \sigma_{33} & \sigma_{34} & \sigma_{35} \\ & & & \sigma_{44} & \sigma_{45} \\ & & & & \sigma_{55} \end{pmatrix} = \tag{64}$$

$$\begin{pmatrix} \omega_1 + \phi_{11} & \beta_1\phi_{11} & 0 & 0 & \beta_3\phi_{11} \\ & \beta_1^2\phi_{11} + \psi_1 & 0 & 0 & \beta_1\beta_3\phi_{11} \\ & & \omega_2 + \phi_{22} & \beta_2\phi_{22} & \beta_4\phi_{22} \\ & & & \beta_2^2\phi_{22} + \psi_2 & \beta_2\beta_4\phi_{22} \\ & & & & \beta_3^2\phi_{11} + \beta_4^2\phi_{22} + \psi_3 \end{pmatrix}.$$

The parameter ϕ_{12} is still identifiable from σ_{13}, but three equations are lost since $\phi_{12} = 0$ also implies $\sigma_{14} = \sigma_{23} = \sigma_{24} = 0$. Thus there are eleven equations in the eleven remaining unknown parameters. The condition of the the parameter count rule is satisfied, and identifiability of the entire parameter vector is still possible.

Using (64), $\beta_3 = \sigma_{25}/\sigma_{12}$ and $\beta_4 = \sigma_{45}/\sigma_{34}$. If β_3 and β_4 are non-zero, solution for the rest of the parameters is routine. But if $\beta_3 = 0$, then β_1 is no longer identifiable. Similarly, if $\beta_4 = 0$, then β_2 is no longer identifiable. Since the whole point of this model is likely to test something like $H_0 : \beta_3 = 0$, it's important to examine the situation where this null hypothesis is true.

Suppose one could be sure that $Cov(X_i, X_2) = \phi_{12} = 0$. Consider two ways of testing $H_0 : \beta_3 = 0$. The first and most obvious way is to just compare the likelihood ratio test statistic to a chi-squared critical value with one degree of freedom. The second way is to look at equality (64) and observe that $\beta_3 = 0$ implies both $\sigma_{15} = 0$ and $\sigma_{25} = 0$. This hypothesis imposes the same constraints on the covariance matrix and could also be tested with a likelihood ratio test, but it would be a two degree of freedom test. Which is correct?

This is a situation where one must break down and actually think about the conditions under which the likelihood ratio statistic has the distribution it is supposed to. The critical condition is Wilks' (1936) regularity condition zero, which just happens to be identifiability. Since the parameter vector is not identifiable under the null hypothesis, Wilks' theorem does not apply, and we are in a situation that Davison (?, pages) would call non-regular. Two degrees of freedom under H_0 are much more believable than one. Furthermore, if the distribution of the likelihood ratio statistic under the null hypothesis is really $\chi^2(2)$, then its expected value is two rather than one, and using the critical value

with one *df* will result in rejection at a rate higher than α. This is undesirable, to say the least. The moral of this story is that the study of identifiability should specifically consider those parts of the parameter space where important null hypotheses are true.

Also, be aware that the 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.

0.12 Instrumental Variables Again

In Section 0.5, the method of instrumental variables was introduced as a solution to the problems that arise when explanatory variables that are missing from the model cause non-zero covariances between the error term and variables that are in the model. We will now see that instrumental variables can help with measurement error too.

0.13 Exercises for Chapter 0

- Exercises 0.2: Conditional and unconditional regression

  1. Everybody knows that $Var(Y_i) = \sigma^2$ for a regression model, but that's really a conditional variance. Independently for $i = 1, \ldots, n$, let

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

  where $\epsilon_1, \ldots \epsilon_n$ are independent random variables with expected value zero and common variance σ^2, $E(X_{i,1}) = \mu_1$, $Var(X_{i,1}) = \sigma_1^2$, $E(X_{i,2}) = \mu_2$, $Var(X_{i,2}) = \sigma_2^2$, and $Cov(X_{i,1}, X_{i,2}) = \sigma_{12}$. Calculate $Var(Y_i)$; show your work.

  2. Suppose that the model (3) has an intercept. How many integral signs are there in the second line of (6)? The answer is a function of n and p.

  3. The usual univariate multiple regression model with independent normal errors is

  $$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

  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 $\sigma^2\mathbf{I}_n$, with $\sigma^2 > 0$ an unknown constant. But of course in practice, the explanatory variables are random, not fixed. Clearly, if the model holds *conditionally* upon the values of the explanatory variables, then all the usual results hold, again conditionally upon the particular values of the explanatory variables. The probabilities (for example, p-values) are conditional probabilities, and the F statistic does not have an F distribution, but a conditional F distribution, given $\mathbf{X} = \mathbf{x}$.

     (a) Show that the least-squares estimator $\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ is conditionally unbiased.

     (b) Show that $\widehat{\boldsymbol{\beta}}$ is also unbiased unconditionally.

     (c) A similar calculation applies to the significance level of a hypothesis test. Let F be the test statistic (say for an extra-sum-of-squares F-test), and f_c be the critical value. If the null hypothesis is true, then the test is size α, conditionally upon the explanatory variable values. That is, $P(F > f_c|\mathbf{X} = \mathbf{x}) = \alpha$. Find the *unconditional* probability of a Type I error. Assume that the explanatory variables are discrete, so you can write a multiple sum.

- Exercises **??**: The Centering Rule

  Maybe refer to some exercises from the Appendix.

- Exercises 0.4: Omitted variables

1. In the following regression model, the independent variables X_1 and X_2 are random variables. The true model is

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

independently for $i = 1, \ldots, n$, where $\epsilon_i \sim N(0, \sigma^2)$.
The mean and covariance matrix of the independent variables are given by

$$E\begin{pmatrix} X_{i,1} \\ X_{i,2} \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} \quad \text{and} \quad Var\begin{pmatrix} X_{i,1} \\ X_{i,2} \end{pmatrix} = \begin{pmatrix} \phi_{11} & \phi_{12} \\ \phi_{12} & \phi_{22} \end{pmatrix}$$

Unfortunately $X_{i,2}$, which has an impact on Y_i and is correlated with $X_{i,1}$, is not part of the data set. Since $X_{i,2}$ is not observed, it is absorbed by the intercept and error term, as follows.

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

The primes just denote a new β_0 and a new ϵ_i. It was necessary to add and subtract $\beta_2\mu_2$ in order to obtain $E(\epsilon_i') = 0$. And of course there could be more than one omitted variable. They would all get swallowed by the intercept and error term, the garbage bins of regression analysis.

   (a) What is $Cov(X_{i,1}, \epsilon_i')$?
   (b) Calculate the variance-covariance matrix of $(X_{i,1}, Y_i)$ under the true model.
   (c) Suppose we want to estimate β_1. The usual least squares estimator is

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^n (X_{i,1} - \overline{X}_1)(Y_i - \overline{Y})}{\sum_{i=1}^n (X_{i,1} - \overline{X}_1)^2}.$$

   You may just use this formula; you don't have to derive it. Is $\widehat{\beta}_1$ a consistent estimator of β_1 for all points in the parameter space if the true model holds? Answer Yes or no and show your work. Remember, X_2 is not available, so you are doing a regression with one independent variable. You may use the consistency of the sample variance and covariance without proof.
   (d) Are there *any* points in the parameter space for which $\widehat{\beta}_1$ is a consistent estimator when the true model holds?

2. Ordinary least squares is often applied to data sets where the independent variables are best modeled as random variables. In what way does the usual conditional linear regression model imply that (random) independent variables have zero covariance with the error term? Hint: Assume $\mathbf{X}_i$ as well as ϵ_i continuous. What is the conditional distribution of ϵ_i given $\mathbf{X}_i = \mathbf{x}_i) = 0$?

3. Show that $E(\epsilon_i|X_i = x_i) = 0$ for all x_i implies $Cov(X_i, \epsilon_i) = 0$, so that a standard regression model without the normality assumption still implies zero covariance (though not necessarily independence) between the error term and explanatory variables.

- Exercises 0.6: Measurement error

1. Calculate expression (29) for the reliability, showing the details that were skipped. The point of this question (besides exercising your variance-covariance muscles and keeping you busy so you don't have a personal life) is to see whether you feel comfortable assuming $\mu = 0$ even though it may not be.
2. In a study of diet and health, suppose we want to know how much snack food each person eats, and we "measure" it by asking a question on a questionnaire. Surely there will be measurement error, and suppose it is of a simple additive nature. But we are pretty sure people under-report how much snack food they eat, so a model like $W = X + e$ with $E(e) = 0$ is hard to defend. Instead, let

$$W = \nu + X + e,$$

where $E(X) = \mu$, $E(e) = 0$, $Var(X) = \sigma_x^2$, $Var(e) = \sigma_e^2$, and $Cov(X, e) = 0$ The unknown constant ν could be called *measurement bias*. Calculate the reliability of W for this model. Is it the same as (29), or does $\nu \neq 0$ make a difference?

3. Continuing Exercise 2, suppose that two measurements of W are available.

$$\begin{aligned} W_1 &= \nu_1 + X + e_1 \\ W_2 &= \nu_2 + X + e_2, \end{aligned}$$

where $E(X) = \mu$, $Var(X) = \sigma_T^2$, $E(e_1) = E(e_2) = 0$, $Var(e_1) = Var(e_2) = \sigma_e^2$, and X, e_1 and e_2 are all independent. Calculate $Corr(W_1, W_2)$. Does this correlation still equal the reliability?

4. Let X be a latent variable, $W = X + e_1$ be the usual measurement of X with error, and $G = X + e_2$ be a measurement of X that is deemed "gold standard," but of course it's not completely free of measurement error. It's better than W in the sense that $0 < Var(e_2) < Var(e_1)$, but that's all you can really say. This is a realistic scenario, because nothing is perfect. Accordingly, let

$$\begin{aligned} W &= X + e_1 \\ G &= X + e_2, \end{aligned}$$

where $E(X) = \mu$, $Var(X) = \sigma_x^2$, $E(e_1) = E(e_2) = 0$, $Var(e_1) = \sigma_1^2$, $Var(e_2) = \sigma_2^2$ and that X, e_1 and e_2 are all independent of one another. Prove that the

squared correlation between W and G is strictly less than the reliability of W. Show your work.

The idea here is that the squared *population* correlation[50] between an ordinary measurement and an imperfect gold standard measurement is strictly less than the actual reliability of the ordinary measurement. If we were to estimate such a squared correlation by the corresponding squared *sample* correlation, all we would be doing is estimating a quantity that is not the reliability. On the other hand, we would be estimating a lower bound for the reliability — and this could be reassuring if it is a high number.

5. In this continuation of Exercise 4, show what happens when you calculate the squared *sample* correlation between a usual measurement and an imperfect gold standard, and let $n \to \infty$. It's just what you would think.

6. Suppose we have two equivalent measurements with uncorrelated measurement error:

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

where $E(X) = \mu$, $Var(X) = \sigma_x^2$, $E(e_1) = E(e_2) = 0$, $Var(e_1) = Var(e_2) = \sigma_e^2$, and X, e_1 and e_2 are all independent. What if we were to measure the true score X by adding the two imperfect measurements together? Would the result be more reliable?

   (a) Let $S = W_1 + W_2$. Calculate the reliability of S. Is there any harm in assuming $\mu = 0$?

   (b) Suppose you take k independent measurements (in psychometric theory, these would be called equivalent test items). What is the reliability of $S = \sum_{i=1}^{k} W_i$? Show your work.

   (c) What happens as the number of measurements $k \to \infty$?

   This exercise establishes the well-known principle that longer tests tend to be more reliable. The measurement of practically anything can be improved by measuring it independently several times and then averaging the results — assuming this is possible.

7. Suppose we have two equivalent measurements with *correlated* measurement error:

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

[50] When we do Greek-letter calculations, we are figuring out what is happening in the population from which a data set might be a random sample.

where $E(X) = \mu$, $Var(X) = \sigma_x^2$, $E(e_1) = E(e_2) = 0$, $Var(e_1) = Var(e_2) = \sigma_e^2$, and e_1 and e_2 are all independent of X but $Cov(e_1, e_2) = \kappa$. Calculate $Corr(W_1, W_2)$; show your work. What is the relationship of your answer to the reliability if $\kappa > 0$ (which is typical of correlated measurement error)? The point of this question is that correlated measurement errors are more the rule than the exception in practice, and it's poison.

- Exercises 0.7: Ignoring measurement error

1. The following is perhaps the simplest example of what happens to regression when there is measurement error in the explanatory variable. Independently for $i = 1, \ldots, n$, let

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

where $E(X_i) = E(\epsilon_i) = 0$, $Var(X_i) = \sigma_x^2$, $Var(\epsilon_i) = \sigma_\epsilon^2$, $Var(e_i) = \sigma_e^2$, and X_i, ϵ_i and e_i are all independent. Notice that W_i is just X_i plus a piece of random noise. This is a simple additive model of measuremnt error.

Unfortunately, we cannot observe the X_i values. All we can see are the pairs (X_i, W_i) for $i = 1, \ldots, n$. So we do what everybody does, and fit the *naive* (mis-specified, wrong) model

$$Y_i = W_i\beta + \epsilon_i$$

and estimate β with the usual formula for regression through the origin. Where does $\widehat{\beta}_n$ go as $n \to \infty$? Show your work.

2. Recall the simulation study of inflated Type I error when independent variables are measured with error but one ignores it and uses ordinary regression anyway. We needed to produce correlated (latent, that is unobservable) independent variables from different distributions. Here's how we did it.

   (a) It is easy to simulate a collection of independent random variables from any distribution, and then standardize them to have expected value zero and variance one. Let $E(X) = \mu$ and $Var(X) = \sigma^2$. Now define $Z = \frac{X-\mu}{\sigma}$. Find

      i. $E(Z)$
      ii. $Var(Z)$

   (b) Okay, now let R_1, R_2 and R_3 be independent random variables from any distribution you like, but standardized to have expected value zero and variance one. Now let

$$\begin{aligned} W_1 &= \sqrt{1-\phi}\, R_1 + \sqrt{\phi}\, R_3 \text{ and} \\ W_2 &= \sqrt{1-\phi}\, R_2 + \sqrt{\phi}\, R_3. \end{aligned}$$

   Find

i. $Cov(W_1, W_2)$
ii. $Corr(W_1, W_2)$

(c) This one is more efficient. Let R_1 and R_2 be independent random variables with expected value zero and variance one. Now let

$$\begin{aligned} W_1 &= \sqrt{\frac{1+\phi}{2}}\, R_1 + \sqrt{\frac{1-\phi}{2}}\, R_2 \\ W_2 &= \sqrt{\frac{1+\phi}{2}}\, R_1 - \sqrt{\frac{1-\phi}{2}}\, R_2 \end{aligned}$$

Find
i. $Cov(W_1, W_2)$
ii. $Corr(W_1, W_2)$

(d) Briefly state how you know the following. No proof is required.
- If the R variables are normal and $\phi = 0$, both methods yield X_1 and X_2 independent.
- But if the Rs are non-normal, then $\phi = 0$ only implies independence for the first method.

- Exercises 0.8: Modeling measurement error

1. Let $X_1, \ldots, X_n$ be a random sample from a normal distribution with mean θ_1 and variance $\theta_2 + \theta_3$, where $-\infty < \theta_1 < \infty$, $\theta_2 > 0$ and $\theta_3 > 0$. Are the prameters of this model identifiable? Answer Yes or No and prove your answer. This is fast.
2. Let $X_1, \ldots, X_n$ be a random sample from a normal distribution with mean θ and variance θ^2, where $-\infty < \theta < \infty$. Is θ identifiable? Answer Yes or No and justify your answer. This is even faster than the last one.
3. For this problem you may want to read about the *invariance principle* of maximum likelihood estimation in Appendix A. Consider the simple regression model

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

where β is an unknown constant, $X_i \sim N(0, \phi)$, $\epsilon_i \sim N(0, \psi)$ and the random variables X_i and ϵ_i are independent. X_i and Y_i are observable variables.

(a) What is the parameter vector $\boldsymbol{\theta}$ for this model? It has three elements.
(b) What is the distribution of the data vector $(X_i, Y_i)^\top$? Of course the expected value is zero; obtain the covariance matrix in terms of $\boldsymbol{\theta}$ values. Show your work.
(c) Now solve three equations in three unknowns to express the three elements of $\boldsymbol{\theta}$ in terms of $\sigma_{i,j}$ values.
(d) Are the parameters of this model identifiable? Answer Yes or No and state how you know.

(e) For a sample of size n, give the MLE $\widehat{\mathbf{\Sigma}}$. Your answer is a matrix containing three scalar formulas (or four formulas, if you write down the same thing for $\widehat{\sigma}_{1,2}$ and $\widehat{\sigma}_{2,1}$). Write your answer in terms of X_i and Y_i quantities. You are *not* being asked to derive anything. Just translate the matrix MLE into scalar form.

(f) Use the invariance principle to obtain the formula for $\widehat{\beta}$ and simplify. Show your work.

(g) Give the formula for $\widehat{\phi}$. Use the invariance principle.

(h) Obtain the formula for $\widehat{\psi}$ and simplify. Use the invariance principle. Show your work.

4. Consider the regression model

$$\begin{aligned} Y_{i,1} &= \beta_1 X_i + \epsilon_{i,1} \\ Y_{i,2} &= \beta_2 X_i + \epsilon_{i,2}, \end{aligned}$$

where $X_i \sim N(0, \phi)$, and X_i is independent of $\epsilon_{i,1}$ and $\epsilon_{i,2}$. The error terms $\epsilon_{i,1}$ and $\epsilon_{i,2}$ are bivariate normal, with mean zero and covariance matrix

$$\mathbf{\Psi} = \begin{pmatrix} \psi_{1,1} & \psi_{1,2} \\ \psi_{1,2} & \psi_{2,2} \end{pmatrix}.$$

The variables X_i, $Y_{i,1}$ and $Y_{i,2}$ are observable; there is no measurement error.

(a) What is the parameter vector $\boldsymbol{\theta}$ for this model? It has six elements.

(b) Calculate the covariance matrix of the observable variables; show your work.

(c) Are the parameters of this model identifiable? Answer Yes or No and justify your answer.

5. Here is a multivariate regression model with no intercept and no measurement error. Independently for $i = 1, \ldots, n$,

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

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.

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

$\boldsymbol{\beta}$ 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 $q \times 1$ 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 $\mathbf{X}_i$.

Are the parameters of this model identifiable? Answer Yes or No and show your work.

6. Consider the following simple regression through the origin with measurement error in both the explanatory and response variables. Independently for $i = 1, \ldots, n$,

$$\begin{aligned} Y_i &= \beta X_i + \epsilon_i \\ W_{i,1} &= X_i + e_{i,1} \\ W_{i,2} &= X_i + e_{i,2} \\ V_i &= Y_i + e_{i,3} \end{aligned}$$

where X_i and Y_i are latent variables, ϵ_i, $e_{i,1}$, $e_{i,2}$, $e_{i,3}$ and X_i and are independent normal random variables with expected value zero, $Var(X_i) = \phi$, $Var(\epsilon_i) = \psi$, and $Var(e_{i,1}) = Var(e_{i,2}) = Var(e_{i,3}) = \omega$. The regression coefficient β is a fixed constant. The observable variables are $W_{i,1}, W_{i,1}$ and V_i.

   (a) Calculate the variance-covariance matrix of the observable variables. Show your work.

   (b) Write down the moment structure equations.

   (c) Are the parameters of this model identifiable? Answer Yes or No and prove your answer.

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

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

where $E(X_i) = \mu \neq 0$, $E(\epsilon_i) = E(e_i) = 0$, $Var(X_i) = \phi$, $Var(\epsilon_i) = \psi$, $Var(e_i) = \omega$, and X_i, e_i andϵ_i are all independent. The variables X_i is latent, while W_i and Y_i are observable.

   (a) Does this model pass the test of the parameter count rule? Answer Yes or No and give the numbers.

   (b) Is the parameter vector identifiable? Answer Yes or No and prove your answer. If the answer is No, give a simple example of two different sets of parameter values that yield the same (bivariate normal) distribution of the observable data.

   (c) Let

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^n W_i Y_i}{\sum_{i=1}^n W_i^2}.$$

   Is $\widehat{\beta}_1$ a consistent estimator of β? Answer Yes or No and prove your answer.

(d) Let

$$\widehat{\beta}_2 = \frac{\sum_{i=1}^n Y_i}{\sum_{i=1}^n W_i}.$$

- Is $\widehat{\beta}_2$ a consistent estimator of β? Answer Yes or No and justify your answer.
- We know from Theorem 1 that consistent estimation is impossible when the parameter is not identifiable. Does this example contradict Theorem 1?

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

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

where

- X_i is a normally distributed *latent* variable with mean zero 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, ϵ_i, $e_{i,1}$ and $e_{i,2}$ are all independent of one another.

(a) What is the parameter vector $\boldsymbol{\theta}$ for this model?

(b) Does this problem pass the test of the Parameter Count Rule? Answer Yes or No and give the numbers.

(c) Calculate the variance-covariance matrix of the observable variables. Show your work.

(d) Is the parameter vector identifiable? Answer Yes or No and prove your answer.

(e) Propose a consistent estimator of the parameter β, and show it is consistent.

- Exercises 0.9
-
-
-
-