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

0.1 Regression: Conditional or Unconditional?

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

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

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

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

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

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

Now please take a step back and think about this model, rather than just accepting it without question. In particular, think about why the x variables should be constants. It's true that if they are constants then all the calculations are easier, but in the typical application of regression to observational[1] data, it makes more sense to view the independent variables as random variables rather than constants. Why? Because if you took repeated samples from the same population, the values of the independent 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 independent 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 independent 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 independent variable values. No! This sample was taken from a population, and the conclusions should apply to that population, not to subsets of the population with these particular values of the independent variables.

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

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

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

Using the double expectation formula $E\{Y\} = E\{E\{Y|X\}\}$,

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

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 independent variables are random.*

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

[1] *Observational* data are just observed, rather than being controlled by the investigator. For example, the number of minutes outside per day 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.

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

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

Size α Tests Suppose Model (1) is conditionally correct, and we plan to use an F test. Conditionally upon the x values, the F statistic has an F distribution when the null hypothesis is true, but unconditionally it does not. Rather, its distribution is a *mixture* of Fs, 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 independent variable values $\mathbf{x}$. In that case,

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

So, the so-called F-test has the correct Type I error rate when the independent 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 independent variables are random and we assume they are fixed, the resulting estimators and tests might be of generally low quality, even though the estimators are unbiased and the tests have the right Type I error rate. Now we will see that given a fairly reasonable set of assumptions, this objection has no merit.

Denoting the independent variable values by $\mathbf{X}$ and the dependent 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 independent variables does not depend upon the values of $\boldsymbol{\beta}$ or σ^2 in any way. This is surely not too hard to believe.

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

Because likelihoods are just joint densities or probability mass functions viewed as functions of the parameter, the notation of Appendix A.4.4 may be stretched just a little bit to write the likelihood function for the unconditional model (with $\mathbf{X}$ random) in terms of conditional densities as

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

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

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

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

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

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

It turns out that there is a very serious problem with applying standard regression methods to observational data[2], but it's not because the independent variables are random. It's because they are random and measured with error.

Exercises 0.1

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

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

[2] Please notice how radical this claim is. Regression methods are applied to observational data *all the time*, and we teach students how to do it in almost every Statistics class where regression is mentioned. I am saying that this standard practice a very bad idea. It's not wrong theoretically; the theory is great. But the applications are almost guaranteed to be misleading. See Section 0 for details.

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 $Cov(X_{i,1}, X_{i,2}) = \kappa$. Calculate $Var(Y_i)$; show your work.

2. Suppose that the model (1) has an intercept. How many integral signs are there in the second line of (2)? 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 independent variables are random, not fixed. Clearly, if the model holds *conditionally* upon the values of the independent variables, then all the usual results hold, again conditionally upon the particular values of the independent 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 independent 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 independent variables are discrete, so you can write a multiple sum.

0.2 Measurement error

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

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

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

0.2.1 A simple additive model for measurement error

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

Here, we will put classification error aside because it is technically very 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,

$$W = X + e \tag{4}$$

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

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

So, it is impossible to tell how much of the variance in the observable variable W comes from variation in the true quantity of interest, and how much comes from random noise.

In psychometric theory[3], the *reliability*[4] 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

[3] 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* [4]. 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.

[4] Reliability has a completely unrelated meaning in survival analysis, and I believe yet another meaning

"observed score" is W. Recalling the definition of a correlation,

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

we have the reliability of the measurement W equal to

$$\begin{aligned}
\rho &= \left(\frac{Cov(X,W)}{SD(X)SD(W)}\right)^2 \\
&= \left(\frac{\sigma_X^2}{\sqrt{\sigma_X^2}\sqrt{\sigma_X^2+\sigma_e^2}}\right)^2 \\
&= \frac{\sigma_X^4}{\sigma_X^2(\sigma_X^2+\sigma_e^2)} \\
&= \frac{\sigma_X^2}{\sigma_X^2+\sigma_e^2}. \qquad (5)
\end{aligned}$$

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

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 7 at the end of this section.

Since reliability represents quality of measurement, estimating it is an important goal. Using the definition directly is seldom possible. Reliability is the squared correlation between a latent variable and its observable counterpart, but by definition, values of the latent variable cannot be observed. This means another approach is needed.

On rare occasions and perhaps with great expense, it may be possible to obtain perfect or near-perfect measurements on a subset of the sample; the term *gold standard* is sometimes applied to such measurements. In that case, the reliability of the usual measurement can be estimated by a squared sample correlation between the usual measurement and the gold standard measurement. But even measurements that are called gold standard are seldom truly free of measurement error. Consequently, reliabilities that are estimated by correlating imperfect gold standards and ordinary measurements are biased downward: See Exercise 4 at the end of this section.

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

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

in statistical quality control.

where $E(X) = \mu$, $Var(X) = \sigma^2_X$, $E(e_1) = E(e_2) = 0$, $Var(e_1) = Var(e_2) = \sigma^2_e$, and X, e_1 and e_2 are all independent. Because $Var(e_1) = Var(e_2)$, W_1 and W_2 are called *equivalent measurements*. That is, they are contaminated by error to the same degree.

It turns out that the correlation between W_1 and W_2 is exactly equal to the reliability, and this opens the door to reasonable methods of estimation. The calculation (like many throughout this course) is greatly simplified by the following fact. *Calculating variances and covariances can be greatly simplified by assuming that all expected values are zero, even though they may not be. The answer will be the same.* For the proof, see formula (A.11) and the discussion that follows in Section A.3.1 of Appendix A.

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

$$\begin{aligned}
Corr(W_1, W_2) &= \frac{Cov(W_1, W_2)}{SD(W_1)SD(W_2)} \\
&= \frac{E(W_1W_2)}{\sqrt{\sigma^2_X + \sigma^2_e}\sqrt{\sigma^2_X + \sigma^2_e}} \\
&= \frac{E(X + e_1)(X + e_2)}{\sigma^2_X + \sigma^2_e} \\
&= \frac{E(X^2) + 0 + 0 + 0}{\sigma^2_X + \sigma^2_e} \\
&= \frac{\sigma^2_X}{\sigma^2_X + \sigma^2_e}, \qquad (6)
\end{aligned}$$

which is the reliability.

The calculation above is the basis of *test-retest reliability*[5], 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 long enough apart so they forget how they answered the first time.

Correlated measurement error Notice that if participants remembered their wrong answers or lucky guesses from the first time they took an educational test and just gave the same answer the second time, the result would be a positive correlation between the measurement errors e_1 and e_2. This would mess everything up. Throughout this course we will return again and again to the issue of correlated errors of measurement. For now, just notice how careful planning of the data collection (in this case, the time lag between the two administrations of the test) can eliminate or at least reduce the correlation between

[5] 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.

errors of measurement. In general, the best way to take care of correlated measurement error is with good research design.

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

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

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

$$\begin{array}{rcl} R_n & = & \frac{\sum_{i=1}^n (W_{i,1} - \overline{W}_1)(W_{i,2} - \overline{W}_2)}{\sqrt{\sum_{i=1}^n (W_{i,1} - \overline{W}_1)^2}\sqrt{\sum_{i=1}^n (W_{i,2} - \overline{W}_2)^2}} \\ & = & \frac{\sum_{i=1}^n W_{i,1}W_{i,2} \; - \; n\overline{W}_1\overline{W}_2}{\sqrt{\sum_{i=1}^n W_{i,1}^2 - n\overline{W}_1^2}\sqrt{\sum_{i=1}^n W_{i,2}^2 - n\overline{W}_2^2}} \\ & = & \frac{\left(\frac{1}{n}\sum_{i=1}^n W_{i,1}W_{i,2}\right) \; - \; \overline{W}_1\overline{W}_2}{\sqrt{\left(\frac{1}{n}\sum_{i=1}^n W_{i,1}^2\right) - \overline{W}_1^2}\sqrt{\left(\frac{1}{n}\sum_{i=1}^n W_{i,2}^2\right) - \overline{W}_2^2}}, \end{array} \qquad (7)$$

where the subscript on the sample correlation coefficient R_n emphasizes that it is a function of the sample size n. By the Strong Law of Large Numbers (see Appendix A.5), we have the following:

$$\begin{array}{rcl} \frac{1}{n}\sum_{i=1}^n W_{i,1}W_{i,2} & \overset{a.s.}{\rightarrow} & E(W_{i,1}W_{i,2}) = Cov(W_{i,1}, W_{i,2}) + E(W_{i,1})E(W_{i,2}) = \sigma^2_X + \mu^2 \\ \overline{W}_1 & \overset{a.s.}{\rightarrow} & E(W_{i,1}) = \mu \\ \overline{W}_2 & \overset{a.s.}{\rightarrow} & E(W_{i,2}) = \mu \\ \frac{1}{n}\sum_{i=1}^n W_{i,1}^2 & \overset{a.s.}{\rightarrow} & E(W_{i,1}^2) = Var(W_{i,1}) + (E\{W_{i,1}\})^2 = \sigma^2_X + \sigma^2_e + \mu^2 \\ \frac{1}{n}\sum_{i=1}^n W_{i,2}^2 & \overset{a.s.}{\rightarrow} & E(W_{i,2}^2) = Var(W_{i,2}) + (E\{W_{i,2}\})^2 = \sigma^2_X + \sigma^2_e + \mu^2. \end{array}$$

Now, since R_n is a continuous function of the various sample moments in (7) and almost sure convergence can be treated like an ordinary limit,

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

So R_n is a strongly consistent estimator of the reliability. That is, for a large enough sample size, R_n will get arbitrarily close to the true reliability, and this happens with probability one. Notice that this was a limits problem and not a variance-covariance computation, so there was no assumption of zero expected values – even though the limit calculation also works out for that restricted case.

Exercises 0.2.1

1. Calculate expression (5) 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 (4) 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 (5), 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 $E(X) = \mu$, $Var(X) = \sigma^2_X$, $E(e_1) = E(e_2) = 0$, $Var(e_1) = \sigma^2_1$, $Var(e_2) = \sigma^2_2$ 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. Show your work.

The idea here is that the squared *population* correlation[6] 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. 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 $E(X) = \mu$, $Var(X) = \sigma^2_X$, $E(e_1) = E(e_2) = 0$, $Var(e_1) = Var(e_2) = \sigma^2_e$, and X, e_1 and e_2 are all independent. 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 $k \to \infty$?

[6] 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.

7. Suppose we have two equivalent measurements with *correlated* measurement error:

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

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

0.3 The consequences of ignoring measurement error in regression

This section will show what happens in multiple regression when measurement error in the independent variables is ignored. It turns out that measurement error in the dependent variable is a less serious problem, and will be dealt with later.

0.3.1 One Independent Variable

Example 0.3.1.1 *Independently for $i = 1, \ldots, n$,*

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

where $Var(X_i) = \sigma^2_X$, $Var(e_i) = \sigma^2_e$, and X_i, e_i, ϵ_i are all independent.

Unfortunately, the independent variable X_i canot 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, the least squares estimate $\widehat{\beta}_1$ is

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

That is when the fuzzy independent variable W_i is used instead of the real thing, $\widehat{\beta}_1$ converges to the true regression coefficient, but multiplied by the reliability. That it it's biased, even as the sample size approaches infinity, but biased toward zero because reliability is between zero and one. More discussion is needed here.

- No asymptotic bias when $\beta = 0$
- No inflation of Type I error rate
- Loss of power when $\beta \neq 0$

Measurement error just makes relationship seem weaker than it is. This seems reassuring, but watch out!

0.3.2 Two Independent Variables

In this version there are two independent variables measured with error.

Example 0.3.2.1 *Independently for $i = 1, \ldots, n$,*

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

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

$$Var \left[\begin{array}{c} X_{i,1} \\ X_{i,1} \end{array} \right] = \left[\begin{array}{cc} \phi_{11} & \phi_{12} \\ \phi_{12} & \phi_{22} \end{array} \right].$$

Again, because the actual indepenent varibles $X_{i,1}$ 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*

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

The interest is in testing the relationship of X_2 to Y *controlling for* X_1. The null hypothesis is $H_0 : \beta_2 = 0$. When this null hypothesis is true, we have

$$\begin{aligned} \widehat{\beta}_2 &\overset{a.s.}{\rightarrow} \frac{\beta_1 \phi_{1,2} \omega_1}{(\phi_{1,1} + \omega_1)(\phi_{2,2} + \omega_2)} \\ &= \left(\frac{\omega_1}{\phi_{1,1} + \omega_1} \right) \left(\frac{\beta_1 \phi_{1,2}}{\phi_{2,2} + \omega_2} \right) \end{aligned}$$

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

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

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

0.3.3 A large scale simulation study

This was covered in lecture.

0.4 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 must be dealt with almost at every turn: parameter identifiability. For comparison, first consider a regression model without measurement error, where everything is nice. This is not quite the standard model, because the independent variables are random variables. General principles arise right away, so definitions will be prvided as we go.

0.4.1 Unconditional regression without measurement error

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

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

where

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

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

$$E\left(\begin{array}{c} X_i \\ Y_i \end{array}\right) = \boldsymbol{\mu} = \left(\begin{array}{c} \mu_1 \\ \mu_2 \end{array}\right) = \left(\begin{array}{c} \mu_x \\ \beta_0 + \beta_1 \mu_x \end{array}\right),$$

and variance covariance matrix

$$V\left(\begin{array}{c} X_i \\ Y_i \end{array}\right) = \boldsymbol{\Sigma} = [\sigma_{i,j}] = \left[\begin{array}{cc} \phi & \beta_1\phi \\ \beta_1\phi & \beta_1^2\phi + \psi \end{array}\right].$$

Definition 0.4.1 Moments *of a distribution are quantities such $E(X)$, $E(Y^2)$, $Var(X)$, $E(X^2Y^2)$, $Cov(X,Y)$, and so on.*

Definition 0.4.2 Moment structure equations *are a set of equations expressing moments of the distribution of the data in terms of the model parameters. If the moments involed are limited to variances and covariances, the moment structure equations are called* covariance structure equations.

For the regression Model (8), the moments structure equations are

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

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

$$\begin{aligned}\mu_x &= \mu_1 \\ \beta_0 &= \mu_2 - \frac{\sigma_{1,2}}{\sigma_{1,1}}\mu_1 \\ \beta_1 &= \frac{\sigma_{1,2}}{\sigma_{1,1}} \\ \phi &= \sigma_{1,1} \\ \psi &= \sigma_{2,2} - \frac{\sigma_{1,2}^2}{\sigma_{1,1}}.\end{aligned} \tag{10}$$

The existence of this nice solution is quite revealing. It tells us that the parameters of the normal regression Model (8) 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.

This finding extends to multivariate multiple regression – that is, to linear regression with multiple independent variables and multiple dependent variables. Setting this aside for the present, let us admit that X_i is probably measured with error in Model (8).

0.4.2 A first try at including measurement error

The following is basically the true model of Example 0.3.1.1, with everything normally distributed. Independently for $i = 1, \ldots, n$, let

$$\begin{aligned}Y_i &= \beta_0 + \beta_1 X_i + \epsilon_i \\ W_i &= \nu + X_i + e_i,\end{aligned} \tag{11}$$

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 (11) are just the pairs (W_i, Y_i) for $i = 1, \ldots, n$. The true independent 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

$$E\left(\begin{array}{c} W_i \\ Y_i \end{array}\right) = \boldsymbol{\mu} = \left(\begin{array}{c} \mu_1 \\ \mu_2 \end{array}\right) = \left(\begin{array}{c} \mu_x + \nu \\ \beta_0 + \beta_1\mu_x \end{array}\right),$$

and variance covariance matrix

$$V\left(\begin{array}{c} W_i \\ Y_i \end{array}\right) = \boldsymbol{\Sigma} = [\sigma_{i,j}] = \left[\begin{array}{cc} \phi+\omega & \beta_1\phi \\ \beta_1\phi & \beta_1^2\phi + \psi \end{array}\right].$$

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

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

It is impossible to solve these five equations for the seven model parameters[7]. 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

[7] That's a strong statement, and a strong Theorem is coming.

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

$$\boldsymbol{\Sigma} = \left[\begin{array}{cc} 4 & 2 \\ 2 & 5 \end{array} \right],$$

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 (11) are not *identifiable*. This calls for a brief discussion of identifiability, a topic to which we shall return again and again.

0.4.3 Parameter Identifiability

Definition 0.4.3 *A* Statistical Model *is a set of assertions that partly*[8] *specify the probability distribution of a set of observable data.*

Definition 0.4.4 *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.

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

In Figure 1, θ_1 and θ_2 are two distinct sets of parameter values for which the distribution of the observable data is the same. Let T_n be a estimator that is consistent for both 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.

Figure 1: Two parameters values yielding the same probability distribution

θ_1
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

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

probability distribution must be concentrated in the neighborhood around θ_1, *and* it must be concentrated in the neighborhood around θ_2. This is impossible, since the two regions do not overlap, and 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.

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

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

Definition 0.4.5 *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.

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 0.4.6 *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 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 (12) that come from the regression Model (11). 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 (11), the moment structure equations (12) 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[9] that provides the necessary details.

Theorem 2 *Let*

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

where $\mathbf{x} = (x_1, \ldots, x_p)' \in \mathbb{R}^p$ *and* $\mathbf{y} = (y_1, \ldots, y_q)' \in \mathbb{R}^q$. *If the functions* $f_1, \ldots, f_q$ *are analytic (posessing a Taylor expansion) and* $p > q$, *the set of* $\mathbf{x}$ *values where the system 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 **Counting Rule**.

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

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

0.4.4 Double measurement

Suppose we had a second, independent measurement of the independent variable; "independent" means that the measurment errors are statistically independent of one another.

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

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

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_i + \epsilon_i \\ W_{i,1} &= \nu_1 + X_i + e_{i,1} \\ W_{i,2} &= \nu_2 + X_i + e_{i,2}, \end{aligned} \tag{13}$$

where

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

The model implies that the triples $(W_{i,1}, W_{i,2}, Y_i)$ are independent multivarate normal with

$$E \left(\begin{array}{c} W_{i,1} \\ W_{i,1} \\ Y_i \end{array} \right) = \boldsymbol{\mu} = \left(\begin{array}{c} \mu_1 \\ \mu_2 \\ \mu_3 \end{array} \right) = \left(\begin{array}{c} \mu_x + \nu_1 \\ \mu_x + \nu_2 \\ \beta_0 + \beta_1 \mu_x \end{array} \right),$$

and variance covariance matrix

$$V \left(\begin{array}{c} W_{i,1} \\ W_{i,1} \\ Y_i \end{array} \right) = \boldsymbol{\Sigma} = [\sigma_{i,j}] = \left[\begin{array}{ccc} \phi + \omega_1 & \phi & \beta_1 \phi \\ & \phi + \omega_2 & \beta_1 \phi \\ & & \beta_1^2 \phi + \psi \end{array} \right].$$

Here are some comments.

- There are now nine moment structure equations in nine unknown parameters. This model passes the test of the Counting Rule, meaning that identifiability is possible, but not guaranteed.
- Notice that the model dictates $\sigma_{1,3} = \sigma_{2,3}$. This *model-induced constraint* upon $\boldsymbol{\Sigma}$ is testable. If $H_0 : \sigma_{1,3} = \sigma_{2,3}$ is rejected, this calls the correctness of the model into question. Philosophers of science agree that *falsifiability* – the possibility that a scientific model can be challenged by empirical data – is a very good thing.
- For those model parameters appearing in the covariance matrix, the additional measurement of the independent variable appears to have done the trick. It is striaghtforward to solve for the parameters $\phi, \beta_1, \omega_1, \omega_2$ 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, except that the lack of unique values for μ_x, ν_1, ν_2 and β_0 would mess things up. The 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 $\boldsymbol{\mu}$ lack meaning and interest[10], but we can estimate them with $\overline{\mathbf{X}}_n$ and focus on the parameters in the covariance matrix.

Here is the multivariate normal likelihood from Appendix A.3.2, simplified so that it's clear that it depends on the data only through the MLEs $\overline{\mathbf{X}}_n$ and $\widehat{\boldsymbol{\Sigma}}$. This is just a reproduction of expression (A.15).

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

Notice how for *any* positive definite $\boldsymbol{\Sigma}$, the likelihood is maximized when $\boldsymbol{\mu} = \overline{\mathbf{x}}$, and in that case the last term just disappears, leaving us free to conduct inference on the model parameters in $\boldsymbol{\Sigma}$.

Exercises 0.4

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. Recall the *invariance principle* of maximum likelihood estimation. Let the parameter of a model be θ_1, and $\theta_2 = g(\theta_1)$; then $\widehat{\theta_2} = g(\widehat{\theta_1})$. For models where $\boldsymbol{\Sigma}$ is not restricted by the model (that is, for "saturated" models) and $\boldsymbol{\theta} = g(\boldsymbol{\Sigma})$, one can use the invariance principle to obtain $\widehat{\boldsymbol{\theta}}$ in closed form, with no need for numerical approximation.

   So, consider the simple regression model

   $$Y = \beta X + \epsilon,$$

   where β is an unknown constant, $X \sim N(0, \phi)$, $\epsilon \sim N(0, \psi)$ and the random variables X and ϵ are independent. X and Y are observable variables.

[10] 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-paramterized model is a garbage can.

(a) What is the parameter vector $\boldsymbol{\theta}$ for this model? It has three elements.

(b) What is the distribution of the data vector $(X, Y)'$? 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. This gives you the function g in $\boldsymbol{\theta} = g(\boldsymbol{\Sigma})$.

(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 $\widehat{\boldsymbol{\Sigma}}$. 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) Obtain the formula for $\widehat{\gamma}$ and simplify. Show your work.

(g) Give the formula for $\widehat{\phi}$.

(h) Obtain the formula for $\widehat{\psi}$ and simplify. Show your work.

4. 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 $m \times 1$ random vector of observable dependent variables, so the regression can be multivariate; there are m dependent variables.

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

$\boldsymbol{\beta}$ is an $m \times p$ matrix of unknown constants. These are the regression coefficients, with one row for each dependent variable and one column for each independent variable.

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

Are the parameters of this model identifiable? Show your work.

5. Consider the following simple regression through the origin with measurement error

in both the independent and dependent variables. Independently for $i = 1, \ldots, n$,

$$\begin{array}{lcl} Y_i & = & \beta X_i + \epsilon_i \\ W_{i,1} & = & X_i + e_{i,1} \\ W_{i,2} & = & X_i + e_{i,2} \\ V_i & = & Y_i + e_{i,3} \end{array}$$

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) Are the parameters of this model identifiable? Answer Yes or No and prove your answer.

(b) Is just the parameter β (a function of the parameter vector) identifiable?

(c) Suppose we were to *re-parameterize* the model by letting $\sigma^2 = \psi + \omega$. Would the re-parametrized model be identified? Does this seem like a good idea?