CHAPTER 7

PROXY SEARCHES

Variables that are used in theoretical statements often are not directly observable. When a researcher wishes to discriminate empirically among a set of theories, he must describe precisely the observable differences in the theories. In particular, hypothetical variables must be linked at least probabilistically to observable phenomena. In this situation there is a tendency among empirical workers to dismiss the apparent failure of a theory as merely a breakdown in the link between a hypothetical variable and an observed variable. One might report that "the low R^2 can be interpreted to mean that we have yet to find the appropriate proxy variable." If a theory is thereby completely protected from falsification, we might naturally ask if it is completely protected from verification as well. The goal of this chapter is to answer this question.

We would like to determine the extent to which it is possible to make inferences about theoretical parameters when the hypothetical variables are measured with error. As an extreme possibility, the theoretical parameters may be taken as known and observations used only to determine the accuracy of measurement. A less extreme procedure is to identify several possible ways of measuring the hypothetical variable and to select the proxy that yields the "best" results. This procedure is called a "proxy-variable search." At least when the number of proxy variables is finite, this method apparently spends part of the data evidence to pick a proxy variable but leaves part of the evidence to make inferences about the theoretical parameters.

To give an example of a proxy search in economics, the lifetime budget constraint makes it almost tautological to say that "permanent" consumption depends on "permanent" income. A great deal of empirical work has sought to determine the best way to measure these hypothetical constructs without directly questioning the underlying theory.

The basic statistical model we use is summarized by the equations

$$Y_t = \alpha + \chi_t\beta + z_t\gamma + u_t, \tag{7.1}$$

$$\mathbf{x}_t = \boldsymbol{\theta} + \boldsymbol{\delta}\chi_t + \boldsymbol{\varepsilon}_t, \tag{7.2}$$

$$\chi_t = \eta + \omega z_t + e_t. \tag{7.3}$$

Equation (7.1) describes the theoretical dependence of an observable variable Y_t on an observable variable z_t and an unobservable variable χ_t. Equation (7.2) describes the process that yields the vector of measurements $\mathbf{x}_t$ of the unobservable χ_t, and Equation (7.3) indicates the relevant part of the joint distribution of χ_t and z_t.

We could have treated both Y_t and z_t as unobservable as well. Measurement error in Y_t of the sort described by (7.2) has obvious consequences implied by rewriting (7.1) to allow for the measurement error

$$Y_t = \theta_Y + \delta_Y(\alpha + \chi_t\beta + z_t\gamma + u_t) + \varepsilon_{Yt},$$

where θ_Y, δ_Y, and ε_{Yt} describe the measurement error in Y_t. In such an equation, even if χ_t were observable, we could only determine the coefficients (β,γ) up to the scale factor δ_Y. If the measurement error amplification δ_Y is known, we can estimate β and γ by a regression of Y_t on $\delta_Y\chi_t$ and $\delta_Y z_t$. Conversely, if one of the regression coefficients, β or γ, is known, we may use the same estimates to solve for the other two parameters. Thus we have the choice between spending the evidence to estimate the theoretical coefficients or spending the evidence to estimate the measurement error. Intermediate cases would be implied by assigning proper prior distributions to δ_Y, β, and γ, which though conceptually straightforward seems to be mathematically intractable. A possibility not discussed here is multiple methods of measuring Y_t, each with different inherent biases. The multivariate process that results has proportionality restrictions across equations. For a discussion of maximum likelihood and other estimates see Jöreskog and Goldberger (1975).

Measurement error in the explanatory variables presents problems that are *not* conceptually straightforward. Except in Section 7.6 we have chosen to deal with the case in which one variable is subject to measurement error and the other is not. This is intended to approximate either the situation in which one variable is known to be measured with relative accuracy or the

situation in which one is interested in the process conditional on the measurable variable z_t rather than its theoretical counterpart.

This raises the question of why we should be interested in the parameters of the theoretical process defined conditional on the unobservables $\mathbf{X}_t$ instead of the obviously estimable parameters of the observable process defined conditional only on the observables $\mathbf{x}_t$. For example, a conditional prediction problem in which Y_t is predicted as a function of other observable variables surely requires only the latter parameters. I think the answer to this question has to do with the problem of pooling information from different sources. "Pure" prior information may apply to the theoretical parameters, and even if interest centers on the other parameters it is necessary to know their relationships in order to make use of prior information. When prior information comes from a different experiment with a different measurement device, the two sources of information can be pooled only by identifying what they have in common—the theoretical parameters. Even if pooling is not the immediate goal, it would be terribly unwieldy to have a hundred sets of parameters, all corresponding to a different measuring device, and we thus hypothesize a single set of parameters implied by perfect measurement.

This chapter is designed to proceed step by step toward a discussion of the proxy variable model and its extensions involving many proxies. We first consider simple normal sampling models in which there are inadequate numbers of observations to estimate the unknown parameters. In the Bayesian framework this means that the likelihood function is not integrable and some form of prior density is necessary to compute a proper posterior distribution. The value of analyzing these simple models is that we are able to define concepts and explore peculiarities characteristic of the proxy-variable model in problems that have clearer intuitive resolutions. For example, we learn that maximum likelihood estimates are not defined for some of these models in the sense that at the apparent maximum likelihood point the likelihood function is peculiarly behaved.

In the second section we review the simple errors-in-variables model with $\gamma=0$, $\theta=0$, $\delta=1$, $\omega=0$. No statistical model in the econometric literature has led to so many confusing and erroneous statements as this simple errors-in-variables model. Textbooks tend to suggest that inferences are precluded by the lack of identification; yet intuitively, the observed correlation between $\mathbf{Y}$ and $\mathbf{x}$ seems to contain information about β. In fact, β may be bracketed on one side by the direct least-squares estimate and on the other by the reverse regression estimate equal to the inverse of the regression of $\mathbf{x}$ on $\mathbf{Y}$. The likelihood function attains its maximum along a line corresponding to these values of β and suitably chosen values of the other parameters.

A comprehensive summary of the literature on the errors-in-variables model is provided by Moran (1971). I have selected from that literature only the material that I regard to be most useful. For example, exact prior information about various parameters has been suggested to break the identification log jam, but since such precise prior information is unlikely to be available, I have not included a discussion of it.

The rest of this chapter deals with natural extensions of the simple errors-in-variables model. In the third section we explore the single-proxy-variable model implied by Equations (7.1) to (7.3), and in the fifth section the multiple-proxy-variable model. For each of these, it is possible to compute bounds for γ analogous to the errors-in-variables bound, actually computed by direct and reverse regressions. Without reference to prior information, it is not possible to say anything about β, however. An instrumental variables model is discussed in Section 7.4. This model is a hybrid of the errors-in-variables model, in that there is both a measurement of the unobservable variable and also an independent proxy.

In describing the likelihood function of all these models we must decide first whether Equation (7.3) is part of the "model" or merely a description of one's prior belief about χ_t. Such a distinction to a Bayesian is, of course, meaningless—the "model" is itself merely a description of one's prior belief. But Equation (7.3) with a normally distributed error e_t may be such a special and unlikely "prior" that it may be better to analyze how the posterior distribution is influenced by the model as defined by Equation (7.1) and (7.2) alone. As will be shown, this is not an easy task.

A confusing terminology has been developed to distinguish the model that makes use of (7.3) from the model that does not. The unobservables χ_t that affect the distributions of specific observations are called *incidental parameters*, and the others are called *structural parameters*. The *structural form* of the model makes use of (7.3) to integrate out the incidental parameters and only structural parameters remain. By default, the model consisting only of Equations (7.1) and (7.2) is called the *functional form*. In place of functional and structural form, I would suggest the words conditional and marginal.

One other bit of terminology is used here. Let the joint likelihood function of two parameters be $L(\theta_1,\theta_2)$. A *marginal likelihood* function makes use of a probability distribution for θ_2 to integrate θ_2 from the function: $L^m(\theta_1)=\int L(\theta_1,\theta_2)\, f(\theta_2)d\theta_2$. A *concentrated likelihood* function maximizes the joint likelihood for each value of θ_1: $L^c(\theta_1)=\max_{\theta_2} L(\theta_1,\theta_2)$. [Note that the value of θ_2 depends on θ_1, $\hat{\theta}_2=\hat{\theta}_2(\theta_1)$, and $L^c(\theta_1)=L(\theta_1,\hat{\theta}_2(\theta_1))$.]

The principal conclusion of this chapter is that "reverse" regression should be a part of standard operating procedure. The choice of "depen-

dent" variable for least-squares regression has nothing to do with metaphysical notions of causality. The "left-hand side" variable should be the one measured most inaccurately.

Two important mathematically related questions remain unanswered: What confidence intervals should be used for these models? How do priors effect the decision whether to allocate the evidence to inference about theoretical parameters versus inference about measurement error parameters?

7.1 Inferences with Inadequate Observations

In this section we consider the inferential puzzles that arise in several simple models whose common feature is an excess of uncertain parameters relative to the number of observations. This discussion is intended to provide insights into the more complicated proxy variable problems to be discussed in later sections.

MODEL 1. $x \sim N(\chi, \sigma^2)$. Suppose that a single measurement x is made of some unknown quantity χ with a measurement device that generates normally distributed measurement errors with mean zero and variance σ^2. A single observation from a normal distribution does yield the point estimate x, but without resort to other information, it does not seem possible to say anything about how close x is likely to be to χ. Somewhat surprisingly, this proposition is not transparently obvious on examination of the likelihood function.

The likelihood function given the data x may be written

$$L(\chi, \sigma^2; x) \propto (\sigma^2)^{-1/2} \exp\left[-\frac{1}{2\sigma^2}(x-\chi)^2 \right]. \tag{7.4}$$

We may attempt to maximize this function by setting the logarithmic derivatives to zero

$$0 = \frac{\partial \log L(\chi, \sigma^2; x)}{\partial \chi} = \frac{-(x-\chi)}{\sigma^2},$$

$$0 = \frac{\partial \log L(\chi, \sigma^2; x)}{\partial \sigma^2} = -\frac{1}{2\sigma^2} + \frac{(x-\chi)^2}{2\sigma^4}.$$

The apparent solution to these equations is

$$\chi = x, \qquad \sigma^2 = (\chi - x)^2 = 0,$$

which suggests, contrary to intuition, that the datum favors $\sigma^2 = 0$.

However, since the exponential term in the function (7.4) involves the ratio of two zeroes at the apparent maximum, a more careful examination

of the function is in order. In fact, the function is not properly defined at $(x, 0)$ since within any neighborhood of the point, the function takes on any positive value whatsoever. This can be demonstrated by identifying the lines on which the likelihood function is constant. Setting the logarithm of (7.4) to a constant c we obtain

$$-\frac{1}{2}\log\sigma^2 - \frac{1}{2}\frac{(x-\chi)^2}{\sigma^2} = c$$

which can be rewritten as

$$(x-\chi)^2 = -\sigma^2 \log\sigma^2 - 2c\sigma^2.$$

As σ^2 goes to zero, the right-hand side of this expression goes to zero regardless of the value of c. Thus *every* line of constant likelihood goes through the point $(x, 0)$. Such a point may be called an *essential singularity*.

Although the behavior of the likelihood function in the neighborhood of $(x, 0)$ is peculiar, it remains to be demonstrated that this point is inferentially uninteresting, since there are points close to $(x, 0)$ that are more "likely" than most other points in the parameter space. But, in treating a likelihood function like a probability distribution (a posterior with a diffuse prior), we are interested in the behavior of the function only to the extent that it generates volume under it. Thus, for example, a likelihood function that is uniform between zero and one is inferentially equivalent to a likelihood function that is the same except at the point .5, where it is enormous. Unless the prior allocates positive probability to the point .5, both functions imply the same posterior probabilities.

In an analogous fashion, unless the prior allocates positive probability to the line $\chi = x$, the point $\sigma^2 = 0$ will not be an "unusually" interesting value. Let us take as our prior for χ a normal distribution with mean m^* and finite variance σ_χ^2; then the density of x conditional on σ^2 but marginal with respect to χ is normal with mean m^* and variance $\sigma^2 + \sigma_\chi^2$, and the (marginal) likelihood of σ^2 is

$$L^m(\sigma^2; x) \propto \left(\sigma^2 + \sigma_\chi^2\right)^{-1/2} \exp\left[-\frac{(x-m^*)^2}{2\left(\sigma^2 + \sigma_\chi^2\right)} \right] \tag{7.5}$$

The mode of this function occurs at σ^2 satisfying $\sigma^2 + \sigma_\chi^2 = (x - m^*)^2$, provided that $(x - m^*) > \sigma_\chi^2$. Otherwise, the mode is at the origin, and the marginal likelihood at $\sigma^2 = 0$ takes on the bounded value $(\sigma_\chi^2)^{-1/2} \exp[-\frac{1}{2}(x-m^*)^2/\sigma_\chi^2]$. In words, if the datum x is far from the prior mean m^* in units of the prior standard error, then the datum favors some value of σ^2 greater than zero. If x and m^* are close, the datum favors $\sigma^2 = 0$; but $\sigma^2 = 0$ is never a singular point of the marginal likelihood function, even though it is a singular point of the concentrated likelihood function $\max_\chi L(\chi, \sigma^2; x)$.

Similarly, unless the prior allocates positive probability to the line $\sigma^2=0$, the value $\chi=x$ will not be an "unusually" interesting value. Let us take as our prior for σ^{-2} a gamma distribution with location and scale parameters s^{*2} and ν^*. Then, by referring to properties of a normal-gamma distribution, the marginal likelihood becomes the Student function

$$L^m(\chi;x)\propto f_S(\chi|x,s^*,\nu^*)$$

$$\propto\left[\nu^*+\frac{(\chi-x)^2}{s^{*2}}\right]^{-(\nu^*+1)/2} \tag{7.6}$$

Although this is a function that has a maximum at $\chi=x$, for no value of $\nu^*>0$ is the function unbounded at that point.

The conclusion that is appropriate from this discussion is that the likelihood function is difficult to interpret by itself. We have suggested the slogan "the mapping is the message" to indicate that the evidential content of the data is a mapping from priors into posteriors. Sometimes that mapping is obvious from examination of the likelihood function alone. In this case it is not, and overt reference to prior information is required to determine the values of σ^2 that are favored by the datum.

In analogy with the errors-in-variable terminology, the statement that x is distributed normally with mean χ and variance σ^2 could be called the "functional" form, and the statement that x is distributed normally with mean zero and variance $\sigma^2+\sigma_\chi^2$ could be called the "structural" form. The preceding paragraph then concludes that inferences about σ^2 ought to be made in the context of the structural form of the model since the functional form may lead to erroneous conclusions.[1]

[1]It is interesting also to consider the consequences of diffuse prior distributions. The usual degenerate prior for σ^2 is implied by $\nu^*=0$, and the marginal likelihood (7.6) becomes $|\chi-x|^{-1}$, which has a nonintegrable singularity at $\chi=x$ and is, furthermore, nonintegrable in the tails of the distribution. The usual uniform prior for χ would imply a uniform marginal likelihood for σ^2 (integrate (7.4) with respect to χ) which is nonintegrable in the tail. Thus if you desire a proper posterior distribution for χ, you need a proper prior for σ^2, and if you desire a posterior distribution for σ^2 that is different from the prior, you need a proper prior for χ.

This brings up the question of whether the parameter σ^2 is identified or not. It is true that no two sets of parameters imply the same data density. For this reason, given a proper prior, it cannot be the case that the prior and posterior probabilities of any measurable subset will necessarily coincide. Nonetheless, if the prior for χ is uniform, the posterior and prior on σ^2 will necessarily coincide. It does seem intuitively clear that without some knowledge of χ, the datum contains no interpretable information about σ^2. We may wish to enlarge the definition of identification to include this circumstance. Kadane (1975) provides further discussion.

MODEL 2. $x_i\sim N(\chi,\sigma_i^2)$, $i=1,2$. A single observation from a normal distribution yields an estimate but no meaningful measure of the reliability of the estimate. Suppose next that two independent measurements of χ are made with different measuring devices that may have different variances. In this case, some information about the variances may be derived from the difference between the measurements.

The likelihood function for this model is

$$L(\chi,\sigma_1^2,\sigma_2^2;x_1,x_2)\propto(\sigma_1\sigma_2)^{-1}\exp\left[-\frac{(\chi-x_1)^2}{2\sigma_1^2}-\frac{(\chi-x_2)^2}{2\sigma_2^2}\right]$$

The same pathology as above occurs on the lines

$$(\chi,\sigma_1^2,\sigma_2^2)=(x_1,0,\sigma_2^2)\qquad\text{and}\qquad(\chi,\sigma_1^2,\sigma_2^2)=(x_2,\sigma_1^2,0).$$

It is instructive in this case to concentrate the likelihood function by selecting the value $\chi(\sigma_1^2,\sigma_2^2)$ that maximizes the function for a given σ_1^2 and σ_2^2:

$$\chi(\sigma_1^2,\sigma_2^2)=(\sigma_1^{-2}+\sigma_2^{-2})^{-1}(\sigma_1^{-2}x_1+\sigma_2^{-2}x_2).$$

The concentrated likelihood function is then

$$L(\chi(\sigma_1^2,\sigma_2^2),\sigma_1^2,\sigma_2^2;x_1,x_2)\propto\frac{1}{\sigma_1\sigma_2}\exp\left[-\frac{(x_2-x_1)^2}{2(\sigma_1^2+\sigma_2^2)}\right].$$

In terms of the ratio $r^2=\sigma_1^2/\sigma_2^2$ and the sum $d^2=\sigma_1^2+\sigma_2^2$, this function can be written

$$L^c(r^2,d^2;x_1,x_2)\propto\frac{r^2+1}{r}d^{-2}\exp\left[-\frac{1}{2d^2}(x_2-x_1)^2\right].$$

On any ray out of the origin (r fixed) this function attains its maximum at $d^2=(x_2-x_1)^2$ independent of r. Holding d fixed the function attains a minimum at $r^2=1$ and is unbounded at $r^2=0$ and $r^2=\infty$. The point $r^2=1$, $d^2=(x_2-x_1)^2$, $\chi=(x_1+x_2)/2$ is thus a *saddle point* of the likelihood function.

There is one parameter that seems to be unambiguously "estimable" from these data; it is $\sigma_1^2+\sigma_2^2$ with "estimate" $(x_1-x_2)^2$. This is a reasonable estimate, since x_1-x_2 is normal with mean zero and variance $\sigma_1^2+\sigma_2^2$. The value of knowledge of $d^2=\sigma_1^2+\sigma_2^2$ is that it implies the constraints $\sigma_1^2<d^2$, $\sigma_2^2<d^2$, which in turn may be useful in constraining confidence intervals. To make this clear write the likelihood function in terms of $d^2=\sigma_1^2+\sigma_2^2$

and $r^2=\sigma_1^2/\sigma_2^2$:

$$L(\chi,d^2,r^2;x_1x_2) \propto f_N\left(\chi|m(r^2),v(r^2,d^2)\right) (d^2)^{-1/2}\exp\left[-\frac{1}{2d^2}(x_1-x_2)^2\right] \tag{7.7}$$

where

$$m(r^2)=\frac{x_1+r^2x_2}{1+r^2}$$

$$v(r^2,d^2)=\frac{d^2r^2}{(1+r^2)^2}.$$

In words, the likelihood function is the product of a conditional normal distribution on χ times a function independent of r^2. The second factor is just the likelihood function of d^2 given the observation of x_1-x_2 distributed normally with mean zero and variance d^2.

If we use a diffuse prior for d^2, $f(d^2)\propto d^{-2}$, we may integrate this likelihood function to obtain a Student distribution on χ conditional on r^2,

$$L^m(\chi,r^2;x_2,x_2) \propto f_S\left(\chi|m(r^2),r^2(x_1-x_2)^2/(1+r^2)^2,1\right).$$

Although a Student function with one degree of freedom has no moments, it is possible to compute shortest size-α confidence intervals, which will be located at $m(r^2)$ and have length proportional to the square root of $(x_1-x_2)^2r^2/(1+r^2)^2$.

A marginal posterior distribution on χ requires us to integrate this function with respect to a prior distribution on r^2. Personally and/or publicly acceptable priors for r^2 are unlikely to be available. An alternative is to describe the mapping of one-point priors into posteriors, that is, to compute posterior intervals for χ conditional on r^2 for all values of r^2. Referring to the formulas above, as we vary r^2 from zero to infinity, we vary the location of the interval from x_1 to x_2, and we vary the length of the interval from zero (see Fig. 7.1) to a maximum at $r^2=1$ and back to zero. Thus although it is impossible to compute precise posterior credible intervals, it is possible to give a very reasonable *class* of posterior intervals. Incidentally, the union of these intervals contains χ with probability in excess of α regardless of the prior for r^2.

MODEL 3. $x_{it}\sim N(\chi_t,\sigma^2)$; $t=1,\ldots,T$; $i=1,2$. Whenever possible, it is desirable actually to compute the marginal posterior distribution of the parameter of interest. A model due to Neyman and Scott (1951) provides a dramatic demonstration of the need to marginalize a likelihood function

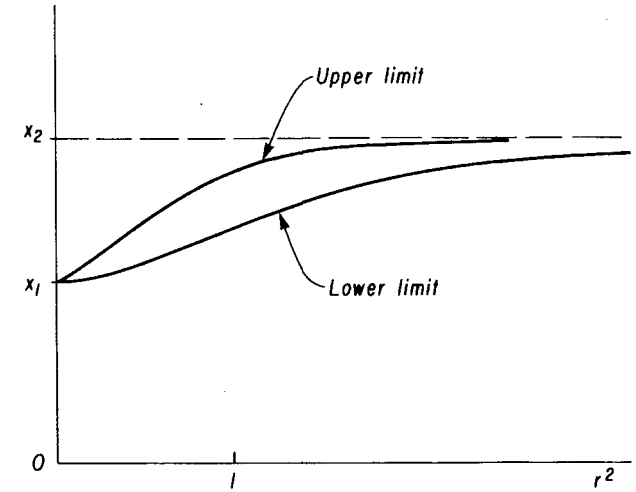

Fig. 7.1 Confidence intervals for χ; $x_1\sim N(\chi,\sigma_1^2)$, $x_2\sim N(\chi,\sigma_2^2)$, $r^2=\sigma_1^2/\sigma_2^2$.

and/or the posterior distribution. Suppose for each of T quantities, $\chi_t (t=1,\ldots,T)$, we obtain two measurements x_{1t} and x_{2t} distributed independently with mean χ_t and variance σ^2. The likelihood function may then be written as

$$L(\chi_1,\ldots,\chi_t,\sigma^2;\mathbf{X})\propto(\sigma^{-2})^T\exp\left[-\frac{1}{2\sigma^2}\sum_{t=1}^{T}\left([x_{1t}-\chi_t]^2+[x_{2t}-\chi_t]^2\right)\right]$$

$$\propto(\sigma^{-2})^T\exp\left[-\frac{1}{\sigma^2}\sum_{t=1}^{T}\left\{(\chi_t-\bar{x}_t)^2+\frac{1}{4}(x_{1t}-x_{2t})^2\right\}\right] \tag{7.8}$$

where $\mathbf{X}$ is the $t\times 2$ matrix of observations and $\bar{x}_t=(x_{1t}+x_{2t})/2$. Maximizing this function with respect to the parameters yields the estimates $\chi_t=\bar{x}_t$ and

$$\hat{\sigma}^2=\sum\frac{(x_{1t}-x_{2t})^2}{4T}.$$

Curiously enough, however, the expected value of $(x_{1t}-x_{2t})^2$ is $2\sigma^2$, and it is easy to show that the estimate just reported converges in probability to $2\sigma^2/4=\sigma^2/2$ as $T\to\infty$.

A marginal likelihood computed by integrating (7.8) with respect to a diffuse distribution for $\chi_1,\chi_2,\ldots,\chi_T$ is easily found to be

$$L^m(\sigma^2;x_1,x_2,\ldots,x_T)\propto(\sigma^{-2})^{T/2}\exp\left[-\frac{1}{4\sigma^2}\sum_t(x_{1t}-x_{2t})^2\right]$$

which has a mode at $\Sigma(x_{1t}-x_{2t})^2/2T$ which does converge in probability to σ^2.

This peculiar example is not easily made sense of. The following is an attempt. The likelihood function implied by two observations from a normal distribution is

$$L(\chi,\sigma^2|x_1,x_2)\propto(\sigma^2)^{-1}\exp\left[-\frac{1}{\sigma^2}(\chi-\bar{x})^2-\frac{1}{4\sigma^2}(x_1-x_2)^2\right]$$

with a mode at $(\chi,\sigma^2)=(\bar{x},(x_1-x_2)^2/2)$. The contours are as depicted in Figure 7.2. Although the function attains its maximum at $\sigma^2=(x_1-x_2)^2/2$, most of the mass is located at values above the maximum. In fact, the usual degrees-of-freedom adjustment would imply the estimate $(x_1-x_2)^2$, thereby implicitly allowing for the relative "thinness" of the likelihood hill at the maximum. If the number of observations is increased, holding χ fixed, this peculiar shape of the likelihood hill corrects itself, and the joint maximum appropriately indicates the point favored by the data. If, as in the example being discussed here, the mean changes, one never gets to the large-sample situation, and the likelihood hill, in fact, becomes increasingly thinner at the maximum. Thus, loosely speaking, in the limit the function approximates its maximum on a set of measure zero.

By the way, this example has had a significant impact on this author's thinking. I used to be relatively uninterested in the difference between

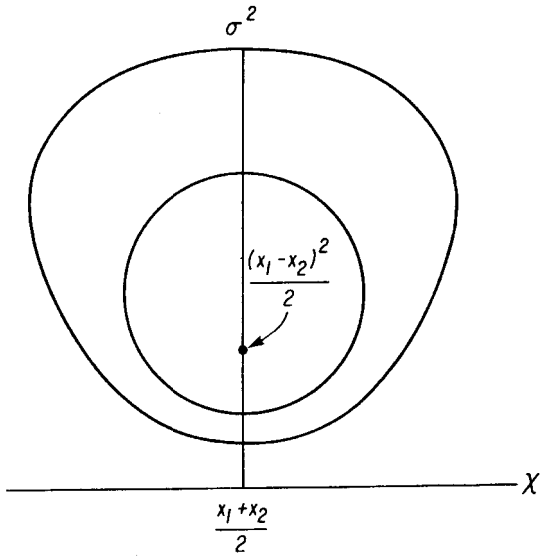

Fig. 7.2 Likelihood contours: $x_i\sim N(\chi,\sigma^2)$, $i=1,2$.

marginal and joint modes, and I would have flippantly asked, "Who chose the (prior) probability distribution that was used to marginalize the likelihood function?" Although that question remains, it is now clear that at least for some problems, marginalization seems essential.

MODEL 4. $x_{it}\sim N(\theta_i+\delta_i\chi_t,\sigma_i^2)$; $t=1,\ldots,T$; $i=1,\ldots,N$. Certain measurement instruments may generate systematic biases. Is it possible by observation of the measurements to determine the extent of the bias? Probably not. In vector notation, a model with biased measurement is

$$\mathbf{x}_t=\boldsymbol{\theta}+\boldsymbol{\delta}\chi_t+\boldsymbol{\varepsilon}_t,\qquad t=1,\ldots,T,$$

where $\mathbf{x}_t$ is a vector of N measurements of the quantity χ_t, $\boldsymbol{\theta}$ and $\boldsymbol{\delta}$ are $N\times 1$ vectors describing the bias, and $\boldsymbol{\varepsilon}_t$ has a multivariate normal distribution with mean vector zero and variance matrix $\mathbf{D}=\text{diag}\{\sigma_1^2,\sigma_2^2,\ldots,\sigma_N^2\}$. We observe in passing that this is the factor analysis model, with a single factor χ_t. (See Harmon, 1960; Jöreskog, 1963; and Lawley and Maxwell, 1964.) As usual, the joint likelihood function has essential singularities, implied by the restrictions that the ith measurement (any i) is errorless, $\sigma_i^2=0$ and $x_{it}=\theta_i+\delta_i\chi_t$. Maximizing the rest of the likelihood function subject to this restriction requires us merely to regress all of the other measurements on $\mathbf{x}_i$, thereby estimating the equations $x_{jt}=\theta_j+\delta_j\chi_t=\theta_j+\delta_j(x_{it}-\theta_i)/\delta_i$.

There is also an identification problem, since by altering the scale of the unobservable χ_t with an offsetting change in the scale of $\boldsymbol{\delta}$ we do not alter the distribution of the observables $\mathbf{x}$. Thus $\boldsymbol{\delta}$ can be determined only up to a scale factor: equivalently, we may only conclude that some instruments give relatively high readings and others relatively low readings, but it is not possible to say which is right, if any.

The structural form of this model assumes that the quantities χ_t come from a normal distribution with mean $\bar{\chi}$ and variance σ_χ^2. The vector $\mathbf{x}_t$ is thereby assumed to be drawn from a multivariate normal distribution with mean $\boldsymbol{\theta}+\boldsymbol{\delta}\bar{\chi}$ and variance $\boldsymbol{\delta}\sigma_\chi^2\boldsymbol{\delta}'+\mathbf{D}$. As in the structural model $\boldsymbol{\delta}$ can be multiplied by any constant, and with a suitable rescaling of σ_χ^2 and $\bar{\chi}$, the distribution of the observables is unchanged. We might as well proceed with the assumption that $\sigma_\chi^2=1$ and $\bar{\chi}=0$, keeping in mind that $\boldsymbol{\delta}$ is determinable only up to a scale factor. Maximum likelihood estimation then requires maximization of

$$L(\boldsymbol{\delta},\mathbf{D},\boldsymbol{\theta};\mathbf{x}_1,\ldots,\mathbf{x}_T)\propto|\boldsymbol{\delta}\boldsymbol{\delta}'+\mathbf{D}|^{-T/2}$$

$$\times\exp\left[-\tfrac{1}{2}\sum_t(\mathbf{x}_t-\boldsymbol{\theta})'(\boldsymbol{\delta}\boldsymbol{\delta}'+\mathbf{D})^{-1}(\mathbf{x}_t-\boldsymbol{\theta})\right].$$

Algorithms for the maximization of such a function are discussed in the factor-analysis literature. The work of Jöreskog (1966) especially should be mentioned here.

7.2 The Errors-in-Variables Problem

The errors-in-variables problem has the feature of all the models just discussed: the ratio of observations to parameters is unhappily small. We must accordingly be alert to the possibility that direct examination of the likelihood surface may be misleading. We nonetheless attempt to explore the likelihood surface directly, a task which may be easier now that we are armed with the knowledge of its potential pitfalls. In this section, and in the sections to follow, we first explore the joint likelihood function of the functional form of the model in which the incidental parameters are treated like the other parameters. The salient feature of this function is its essential singularities. We then explore the (marginal) likelihood function of the structural form of the model in which the incidental parameters are integrated out of the function with respect to a probability function with possibly uncertain hyperparameters.

The errors-in-variables model is mathematically equivalent to Model 4 just discussed with $N=2$, but with a known normalization. In its simplest form we may write the model as

$$Y_t = \beta\chi_t + u_t \tag{7.9}$$

$$x_t = \chi_t + \varepsilon_t \tag{7.10}$$

to indicate that an observation Y_t is linked by a linear process to an unobservable χ_t, which is measured by x_t subject to measurement error ε_t. In effect, we have two measurements of χ_t, one unbiased and the other subject to amplification β.

The least-squares estimate of β suffers from the errors-in-variables "attenuation"—it is biased toward the origin. The bias does not disappear as sample size increases. This sampling property has its counterpart in the likelihood function which in vector notation is

$$L(\boldsymbol{\chi},\beta,\sigma_u^2,\sigma_\varepsilon^2;\mathbf{Y},\mathbf{x}) \propto (\sigma_u^2)^{-T/2}\exp\left[-\frac{1}{2\sigma_u^2}(\mathbf{Y}-\boldsymbol{\chi}\beta)'(\mathbf{Y}-\boldsymbol{\chi}\beta)\right]$$

$$(\sigma_\varepsilon^2)^{-T/2}\exp\left[-\frac{1}{2\sigma_\varepsilon^2}(\boldsymbol{\chi}-\mathbf{x})'(\boldsymbol{\chi}-\mathbf{x})\right].$$

This function, as shown by Solari (1969), has essential singularities at the points satisfying $\sigma_\varepsilon^2=0$ and $\boldsymbol{\chi}=\mathbf{x}$ or $\sigma_u^2=0$ and $\mathbf{Y}=\boldsymbol{\chi}\beta$. Minimizing the nonpathological part of the function, subject to these two pairs of constraints, yields, respectively, the "two regressions" $\hat{\beta}^D=(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\mathbf{Y}$ and

$\hat{\beta}^R=[(\mathbf{Y}'\mathbf{Y})^{-1}\mathbf{Y}'\mathbf{x}]^{-1}$, the first being the simple regression of $\mathbf{Y}$ on $\mathbf{x}$ and the second being the inverse of the regression of $\mathbf{x}$ on $\mathbf{Y}$. These are the extreme estimates of β analogous to the estimates x_1 and x_2 of χ with $x_1\sim N(\chi,\sigma_1^2)$ and $x_2\sim N(\chi,\sigma_2^2)$. These are not maximum likelihood estimates, since the function is not defined at these points. As mentioned before, there is a tendency for the simple regression $\hat{\beta}^D$ to underestimate β ($\hat{\beta}^D$ is "attenuated"), and to compensate for this the "reverse regression" $\hat{\beta}^R$ yields an estimate larger in absolute value than the direct regression. Using the formula $R^2=(\mathbf{Y}'\mathbf{x})^2/\mathbf{x}'\mathbf{x}\mathbf{Y}'\mathbf{Y}$ we may derive the result

$$R^2\hat{\beta}^R=\hat{\beta}^D \tag{7.11}$$

with $0\leqslant R^2\leqslant 1$. Furthermore, it may be shown that the reverse regression estimate tends to overestimate β, and the two regressions therefore consistently bound β.

As in the two-observations-per-mean model (Model 2), the stable point of the likelihood function is a saddle point (Solari, 1969). The estimate of β at the saddle point is a geometric average of the two regressions, and $\boldsymbol{\chi}$ is a simple weighted average of $\mathbf{x}$ and $\mathbf{Y}/\beta$, completely analogous to the saddle point discussed previously, in which the estimate of χ was a simple compromise between x_1 and x_2. Another feature of this point is the curious relationship among the estimates $\hat{\beta}^2=\hat{\sigma}_u^2/\hat{\sigma}_\varepsilon^2$. This phenomenon has generated a great deal of confusion in the literature. It is not especially surprising if we were to write the process like Model 2 as

$$x_{1t}=\frac{Y_t}{\beta}=\chi_t+\frac{u_t}{\beta}$$

$$x_{2t}=x_t=\chi_t+\varepsilon_t,$$

and observe that because of the symmetries, it is not surprising to find a saddle point at $\operatorname{Var}x_1=\operatorname{Var}x_2$, that is, at $\sigma_u^2/\beta^2=\sigma_\varepsilon^2$. In the former problem the constraint $\sigma_1^2=\sigma_2^2$ seems reasonable by an appeal to symmetry. No such symmetry exists between Y_t/β and x_t, and it is unlikely that the saddle point will be a point of special interest between the two extremes.

A (prior) distribution for the unobservables χ_t may allow us to make clearer inferences. In the structural form of this model, we assume that each χ_t was drawn from the same normal population $\chi_t\sim N(\bar{\chi},\sigma_\chi^2)$. This amounts to assuming that the vector (Y_t,x_t) comes from a bivariate normal distribution with mean $(\beta\bar{\chi},\bar{\chi})$ and variance matrix

$$\boldsymbol{\Sigma}=\begin{bmatrix}\sigma_u^2 & 0\\ 0 & \sigma_\varepsilon^2\end{bmatrix}+\begin{bmatrix}\beta\\ 1\end{bmatrix}\sigma_\chi^2[\,\beta,1\,]$$

$$=\begin{bmatrix}\sigma_u^2+\beta^2\sigma_\chi^2 & \beta\sigma_\chi^2\\ \beta\sigma_\chi^2 & \sigma_\chi^2+\sigma_\varepsilon^2\end{bmatrix}. \tag{7.12a}$$

The fact that the mean of Y_t is $\beta\bar{\chi}$ and the mean of x_t is $\bar{\chi}$ suggests an estimator for β: $\bar{Y}/\bar{x}$, the ratio of the observed means. But if the relationship between Y_t and χ_t included the constant α, $Y_t=\alpha+\beta\chi_t$, the mean vector would be $(\alpha+\beta\bar{\chi},\bar{\chi})$, which would not imply an estimate of β. In fact, we could only solve for α in terms of β, $\alpha=\bar{Y}-\beta\bar{\chi}=\bar{Y}-\beta\bar{x}$. Except in the unlikely event that we know α or have a proper prior for it, inferences about β depend entirely on the covariance matrix of the process and not on the location vector. Henceforth, we proceed as if there were an uncertain α in the theoretical relationship. The likelihood functions reported below have been "concentrated" by setting $\bar{\chi}$ to $\bar{x}$ and α to $\bar{Y}-\beta_x$. This has the effect of removing the means of the observations, $x_t-\bar{x}$ and $Y_t-\bar{Y}$.

Observe in passing that the conditional distribution of Y_t given x_t is normal with mean $x_t\beta\sigma_\chi^2/(\sigma_\chi^2+\sigma_\epsilon^2)$. This suggests that in regressing $\mathbf{Y}$ on $\mathbf{x}$ we obtain the "shrunken" coefficient: $\beta\sigma_\chi^2/(\sigma_\chi^2+\sigma_\epsilon^2)$. Conversely, the regression of $\mathbf{x}$ on $\mathbf{Y}$ yields the coefficient $\beta\sigma_\chi^2/(\sigma_u^2+\beta^2\sigma_\chi^2)=\beta^{-1}(\beta^2\sigma_\chi^2/(\sigma_u^2+\beta^2\sigma_\chi^2))$, β^{-1} times a factor less than one. The first shrinkage factor is close to one for σ_ϵ^2 small relative to σ_χ^2, whereas the second is close to one for σ_u^2 small relative to $\beta^2\sigma_\chi^2$. This should generate some further understanding of the content of the "two regressions."

The likelihood function of the structural form, assuming normality and with variables defined around their means, is

$$L(\beta,\sigma_u^2,\sigma_\chi^2,\sigma_\epsilon^2;\mathbf{Y},\mathbf{x}) \propto |\mathbf{\Sigma}|^{-T/2}\exp\left[-\tfrac{1}{2}\sum_t (Y_t,x_t)\mathbf{\Sigma}^{-1}(Y_t,x_t)'\right]$$

$$=|\mathbf{\Sigma}|^{-T/2}\exp\left[-\tfrac{1}{2}\operatorname{tr}\mathbf{\Sigma}^{-1}\mathbf{S}\right] \qquad (7.12b)$$

where

$$\mathbf{S}=\begin{bmatrix}\mathbf{Y}'\mathbf{Y} & \mathbf{Y}'\mathbf{x}\\ \mathbf{x}'\mathbf{Y} & \mathbf{x}'\mathbf{x}\end{bmatrix}.$$

The following result describes the maximum likelihood region.

THEOREM 7.1 (ERRORS-IN-VARIABLES BOUND). *The likelihood function (7.12b) with $\mathbf{\Sigma}$ defined by (7.12a) attains its maximum at any value of β between the direct regression estimate $\hat{\beta}^D=(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\mathbf{Y}$ and the reverse regression estimate $\hat{\beta}^R=(\mathbf{x}'\mathbf{Y})^{-1}\mathbf{Y}'\mathbf{Y}$.*

Proof: If $\mathbf{\Sigma}$ is unconstrained, this function attains its maximum of $|\mathbf{S}/T|^{-T/2}\exp[-T]$ at $\mathbf{\Sigma}=\mathbf{S}/T$. This is a feasible value of $\mathbf{\Sigma}$ if we can find values of σ_u^2, σ_χ^2, σ_ϵ^2, and β such that $\mathbf{\Sigma}(\sigma_u^2,\sigma_\chi^2,\sigma_\epsilon^2,\beta)=\mathbf{S}/T$. We can, in fact,

do this by selecting a value of β, and solving for the other parameters as

$$\sigma_\chi^2=\frac{\mathbf{x}'\mathbf{Y}}{T\beta}$$

$$\sigma_u^2=\frac{\mathbf{Y}'\mathbf{Y}}{T}-\beta^2\sigma_\chi^2=\frac{\mathbf{Y}'\mathbf{Y}-\mathbf{x}'\mathbf{Y}\beta}{T}$$

$$\sigma_\epsilon^2=\frac{\mathbf{x}'\mathbf{x}}{T}-\sigma_\chi^2=\frac{\mathbf{x}'\mathbf{x}-\mathbf{x}'\mathbf{Y}\beta^{-1}}{T}.$$

The constraints $\sigma_u^2\geqslant 0$ and $\sigma_\epsilon^2\geqslant 0$ imply that not all values of β can be associated with the maximum of the likelihood function. These constraints imply $\mathbf{Y}'\mathbf{Y}\geqslant\mathbf{x}'\mathbf{Y}\beta$ and $\mathbf{x}'\mathbf{x}\geqslant\mathbf{x}'\mathbf{Y}\beta^{-1}$, that is,

$$\frac{\beta}{\mathbf{Y}'\mathbf{Y}(\mathbf{x}'\mathbf{Y})^{-1}}=\frac{\beta}{\hat{\beta}^R}\leqslant 1$$

$$\frac{(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\mathbf{Y}}{\beta}=\frac{\hat{\beta}^D}{\beta}\leqslant 1.$$

In words, any value of β between the direct regression $\hat{\beta}^D$ and the reverse regression $\hat{\beta}^R$ with suitably chosen values of the other parameters imply a likelihood value equal to the maximum.

The general features of the concentrated likelihood function

$$L^c(\beta;\mathbf{Y},\mathbf{x})=\max_{\sigma_u^2>0,\sigma_\epsilon^2>0,\sigma_\chi^2>0} L(\beta,\sigma_u^2,\sigma_\chi^2,\sigma_\epsilon^2;\mathbf{Y},\mathbf{x})$$

are difficult to compute. From the preceding discussion we know it has a plateau of height $|\mathbf{S}/T|^{-T/2}\exp[-T]$ between $\hat{\beta}^D$ and $\hat{\beta}^R$. It is also easy to derive

$$L^c(0;\mathbf{Y},\mathbf{x})=\left[\mathbf{x}'\mathbf{x}\mathbf{Y}'\mathbf{Y}/T^2\right]^{-T/2}\exp[-T]$$

(by observing that given $\beta=0$, $\mathbf{x}$ and $\mathbf{Y}$ are independent with different variances). Similarly,

$$L^c(\beta;\mathbf{Y},\mathbf{x})\geqslant\max_{\sigma_u^2>0,\sigma_\epsilon^2>0,\sigma_\chi^2=0} L(\beta,\sigma_u^2,\sigma_\chi^2,\sigma_\epsilon^2;\mathbf{Y},\mathbf{x})$$

$$=L^c(0;\mathbf{Y},\mathbf{x}).$$

$$L^c(\beta;Y,\mathbf{x})\geqslant\max_{\sigma_u^2>0,\sigma_\epsilon^2=0,\sigma_\chi^2>0} L(\beta,\sigma_u^2,\sigma_\chi^2,\sigma_\epsilon^2;\mathbf{Y},\mathbf{x})$$

$$=\left[(\mathbf{Y}-\mathbf{x}\beta)'(\mathbf{Y}-\mathbf{x}\beta)\right]^{-T/2}\left[\mathbf{x}'\mathbf{x}\right]^{-T/2}T^{T/2}\exp[-T],$$

$$L^c(\beta;\mathbf{Y},\mathbf{x})\geqslant\max_{\sigma_u^2=0,\sigma_\epsilon^2>0,\sigma_\chi^2>0} L(\beta,\sigma_u^2,\sigma_\chi^2,\sigma_\epsilon^2;\mathbf{Y},\mathbf{x})$$

$$=\left[(\mathbf{x}-\mathbf{Y}\beta^{-1})'(\mathbf{x}-\mathbf{Y}\beta^{-1})\right]^{-T/2}\left[\mathbf{Y}'\mathbf{Y}\right]^{-T/2}T^{T/2}\exp[-T].$$

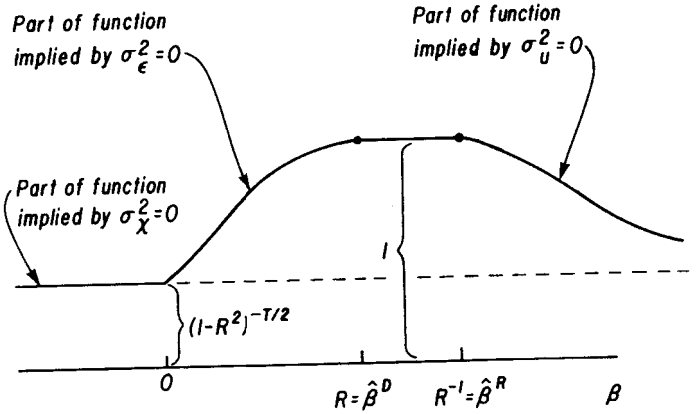

Fig. 7.3 Concentrated likelihood function: the errors-in-variables model.

Note that this last function (whose leading term is associated with the reverse regression) in the limit tends to the value $L^c(0;\mathbf{Y},\mathbf{x})$.

Normalizing so that the likelihood value at the maximum is one and so that $\mathbf{Y}'\mathbf{Y}=\mathbf{x}'\mathbf{x}=1$, and by observing that $|\mathbf{S}|=(\mathbf{x}'\mathbf{x})(\mathbf{Y}'\mathbf{Y})(1-R^2)$, where R^2 is the sample correlation coefficient of $\mathbf{Y}$ and $\mathbf{x}$, we obtain the lower bound for the concentrated likelihood function depicted in Figure 7.3. Note that in addition to having a plateau, this function is uniform for large absolute values of β. To the extent that the concentrated likelihood function is an appropriate one-dimensional description of the evidence about β, we are led to conclude that it is impossible to distinguish one large value of β from another, although values of β with the same sign as $\hat{\beta}^D$ are favored over values with the opposite sign. Decisions that depend on the tails of the distribution thus necessarily are heavily dependent on prior information about β or about the variances.[2]

An approximate posterior distribution for this model has been derived by Lindley and El-Sayyad (1968). They use the ignorance distribution for σ_ϵ^2, proportional to σ_ϵ^{-2}, and show that in a large sample β, conditional on the variance ratio $\lambda=\sigma_u^2/\sigma_\epsilon^2$, is normal with mean β_λ, the root of the quadratic equation (with sign of $\mathbf{x}'\mathbf{Y}$)

$$\beta^2+t\beta-\lambda=0, \qquad t=\frac{\lambda\mathbf{x}'\mathbf{x}-\mathbf{Y}'\mathbf{Y}}{\mathbf{x}'\mathbf{Y}},$$

[2]Note also that the likelihood ratio of $\beta=0$ versus $\beta\neq 0$ is unaffected by the measurement error in χ. It is erroneous to conclude from this fact that a test of the hypotheses $\beta=0$ versus $\beta\neq 0$ is uninfluenced by the measurement error, since, as we have argued in Chapter 4, the relationship between a likelihood ratio and an appropriate hypothesis test is indirect. Nonetheless, it is comforting to know at least that the t value of the regression coefficient has [illegible] that ought to apply to β.

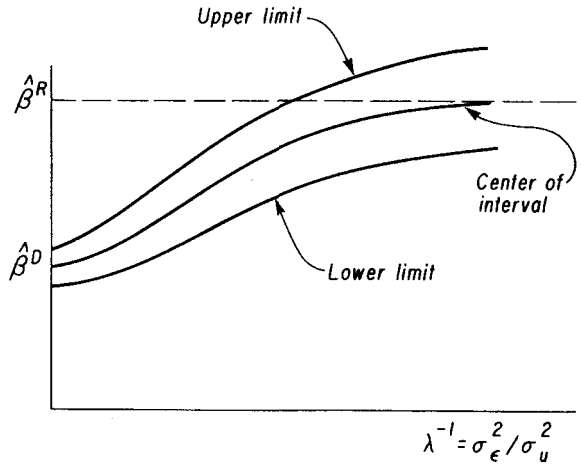

Fig. 7.4 Confidence intervals: the errors-in-variables problem.

and with variance

$$\frac{\beta_\lambda^2\left(\mathbf{x}'\mathbf{x}\mathbf{Y}'\mathbf{Y}-(\mathbf{x}'\mathbf{Y})^2\right)}{T(\mathbf{x}'\mathbf{Y})^2}.$$

There is little sample information about λ, and in the absence of legitimate prior information, it makes sense to consider the class of (approximate) posterior distributions for $\lambda>0$, as we did for Model 2 in the previous section.

The (nonzero) root of the quadratic equation for $\lambda=0$ is $\beta_0=\mathbf{Y}'\mathbf{Y}/\mathbf{x}'\mathbf{Y}$, the reverse regression estimate of β. As λ converges to infinity, the (appropriate) root of the quadratic equation converges to $\beta_\infty=\mathbf{x}'\mathbf{Y}/\mathbf{x}'\mathbf{x}$, the direct regression estimate of β. The standard error is always proportional to β_λ and at $\lambda\to\infty$ the variance takes on the value

$$\left(\frac{\mathbf{x}'\mathbf{Y}}{\mathbf{x}'\mathbf{x}}\right)^2\frac{\mathbf{x}'\mathbf{x}\mathbf{Y}'\mathbf{Y}-[\mathbf{x}'\mathbf{Y}]^2}{T(\mathbf{x}'\mathbf{Y})^2}=\frac{\mathbf{x}'\mathbf{x}\mathbf{Y}'\mathbf{Y}-(\mathbf{x}'\mathbf{Y})^2}{T(\mathbf{x}'\mathbf{x})^2},$$

which happens to be just the usual least-squares estimate of the variance of $\hat{\beta}$. Figure 7.4 indicates this class of intervals which vary from the usual interval located at the direct least-squares $\hat{\beta}^D$ to an interval located at the reverse regression estimate with length expanded by the factor $|\hat{\beta}^R/\hat{\beta}^D|$.

7.3. The Proxy-Variable Problem

A variable x is called a *proxy* for another variable χ if x and χ are positively correlated. Within the normal family we may write $E(x|\chi)=\theta+$

$\delta\chi$, and the assumption of positive correlation is equivalent to $\delta > 0$. In contrast, x is a *measurement* of χ if $E(x|\chi)=\chi$, that is, if $\theta=0$ and $\delta=1$.

The simplest proxy variable problem is described by equations analogous to (7.9) and (7.10)

$$Y_t = \beta\chi_t + u_t \tag{7.13}$$

$$x_t = \delta\chi_t + \varepsilon_t \tag{7.14}$$

where, as in the errors-in-variable problem, we take Y_t and x_t to be defined around their means. The parameter of interest may be considered to be β, in which case x_t is a proxy for the explanatory variable χ_t. This model suffers from the identification problem associated with factor analysis (see Model 4): the vector (β,δ) is unique only up to a scale factor, since the distribution of the observables (Y_t,x_t) is unaffected by a scale change in (β,δ) offset by a scale change in χ_t. A regression of $\mathbf{Y}$ and $\mathbf{x}$ can by itself determine *only* the sign of β (given the sign of δ).

The essential singularities of the joint likelihood function implied by (7.13) and (7.14) occur on the lines $(\sigma_u^2,\boldsymbol{\chi})=(0,\mathbf{Y}/\beta)$ and $(\sigma_\varepsilon^2,\boldsymbol{\chi})=(0,\mathbf{x})$. Maximizing the nonpathological part of the function subject to these constraints yields the estimates $\hat{\beta}/\hat{\delta}=(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\mathbf{Y}$ and $(\hat{\beta}/\hat{\delta})=(\mathbf{x}'\mathbf{Y})^{-1}\mathbf{Y}'\mathbf{Y}$, which are just the direct and reverse regression estimates. Note that the consequence of the identification problem is that the location of these modes depends on the ratio β/δ only.

A more interesting model is described by the equations

$$Y_t = \beta\chi_t + \gamma z_t + u_t \tag{7.13'}$$

$$x_t = \delta\chi_t + \varepsilon_t. \tag{7.14'}$$

This differs from the previous model in allowing another variable, z_t, to affect the dependent variable Y_t. As before, the essential singularities are implied by $\sigma_u^2=0$ or $\sigma_\varepsilon^2=0$. Using these constraints, we may write the model in the first case as $x_t=\delta\chi_t+\varepsilon_t=\delta(Y_t-\gamma z_t)/\beta+\varepsilon_t$ and in the second as $Y_t=(\beta/\delta)x_t+\gamma z_t+u_t$. The first equation implies that γ may be estimated by regressing $\mathbf{x}$ on $\mathbf{Y}$ and $\mathbf{z}$ and by resolving the resulting equation to put $\mathbf{Y}$ on the left-hand side. This may be called the reverse regression estimate of γ and is denoted by $\hat{\gamma}^R$. Referring to the second equation, we see that the essential singularities implied by $\sigma_\varepsilon^2=0$ suggest regressing $\mathbf{Y}$ directly on $\mathbf{x}$ and $\mathbf{z}$. The estimate of γ that results is called the direct estimate and is denoted by $\hat{\gamma}^D$.

If the logic of the errors-in-variable model is extended to this proxy-variable problem, we conjecture that in the structural form of this model the concentrated likelihood function has a plateau between $\hat{\gamma}^D$ and $\hat{\gamma}^R$. This is, in fact, the case. The structural form of this model requires a distribution for χ_t. To make the problem interesting, we allow χ_t and z_t to be correlated. In that case, the simple estimate of γ computed by regressing $\mathbf{Y}$ on $\mathbf{z}$ alone is potentially misleading, since a variable correlated with $\mathbf{z}$ has been omitted from the equation. This creates a pressure on the researcher to find a proxy for the unobservable χ_t. Assuming in particular that χ_t is normal with mean ωz_t and variance σ_χ^2

$$\chi_t \sim N(\omega z_t, \sigma_\chi^2)$$

the means and variances of the observables become

$$E[(Y_t,x_t)|z_t] = [(\beta\omega+\gamma)z_t, \delta\omega z_t]$$

$$\boldsymbol{\Sigma} = V[(Y_t,x_t)|z_t] = \begin{bmatrix} \beta^2\sigma_\chi^2+\sigma_u^2 & \beta\delta\sigma_\chi^2 \\ \beta\delta\sigma_\chi^2 & \delta^2\sigma_\chi^2+\sigma_\varepsilon^2 \end{bmatrix}.$$

By an argument analogous to the material in the previous section, it may be shown that the likelihood function has a plateau between the direct and reverse regressions. The following result analogous to (7.11) can also be derived.

Let $\hat{\gamma}^S$ be the "simple" regression estimate of γ, $(\mathbf{z}'\mathbf{z})^{-1}\mathbf{z}'\mathbf{Y}$, computed by omitting $\mathbf{x}$ from the equation. Let $\hat{\gamma}^D$ and $\hat{\gamma}^R$ be the least-squares estimate and the reverse least-squares estimate. Then it is easy to verify that

$$(\hat{\gamma}^D-\hat{\gamma}^S) = R^2(\hat{\gamma}^R-\hat{\gamma}^S), \tag{7.15}$$

where R^2 is the squared multiple correlation coefficient computed when $\mathbf{Y}$ is regressed on both $\mathbf{x}$ and $\mathbf{z}$. Interpreting this result, we conclude that when an error-ridden proxy variable is included in an equation, the estimates that result are insufficiently far from the estimates when the variable is omitted altogether. The difference should be expanded by a factor no larger than $1/R^2$.

7.4 Instrumental Variables

Suppose next that the variable χ_t is measured with error, but in addition, there is a proxy variable available. The model then consists of the following three equations:

$$Y_t = \beta\chi_t + u_t \tag{7.16}$$

$$x_t = \chi_t + \varepsilon_{1t} \tag{7.17}$$

$$w_t = \delta\chi_t + \varepsilon_{2t}, \tag{7.18}$$

where variables are defined around their means, x_t is a measurement of χ_t, w_t is a proxy variable, and u_t, ε_{1t}, and ε_{2t} are independent normal random variables with zero means and variances σ_u^2, σ_1^2, and σ_2^2, respectively. If the

proxy equation were errorless, that is, if σ_2^2 were zero, we could use the relationship $\chi_t = w_t/\delta$ to determine $Y_t = \beta w_t/\delta + u_t$ and $x_{1t} = w_t/\delta + \varepsilon_{1t}$. Thus in regressing $\mathbf{Y}$ on $\mathbf{w}$ we would estimate β/δ, and in regressing $\mathbf{x}$ on $\mathbf{w}$ we would estimate δ^{-1}. The ratio would then estimate β:

$$\hat{\beta}^{IV} = \frac{\mathbf{Y}'\mathbf{w}}{\mathbf{x}'\mathbf{w}} \tag{7.19}$$

This estimator is known as the "instrumental variables" estimator of β. The variable $\mathbf{w}$ is said to be an "instrument" for $\mathbf{x}$, and the estimator is often derived in the following way. The variable $\mathbf{x}$ is regressed on $\mathbf{w}$ to obtain an estimate equal to $(\mathbf{w}'\mathbf{w})^{-1}\mathbf{w}'\mathbf{x}$ and a "predicted" value of $\mathbf{x}$ equal to $\hat{\mathbf{x}} = \mathbf{w}(\mathbf{w}'\mathbf{w})^{-1}\mathbf{w}'\mathbf{x}$. Then β is estimated by regressing $\mathbf{Y}$ on the predicted $\mathbf{x}$: $(\hat{\mathbf{x}}'\hat{\mathbf{x}})^{-1}\hat{\mathbf{x}}'\mathbf{Y} = \mathbf{Y}'\mathbf{w}/\mathbf{x}'\mathbf{w}$, which is the instrumental variables estimator.

The instrumental variables estimator is due to Reiersol (1945) and to Geary (1949). It is not difficult to demonstrate that $\hat{\beta}^{IV}$ is a consistent estimator of β, for example, see Malinvaud (1970). A careful examination of the likelihood function may help us choose between the consistent estimator $\hat{\beta}^{IV}$ which may have a large small-sample variance and the inconsistent estimators $\hat{\beta}^{D}$ and $\hat{\beta}^{R}$ which may have relatively small variances in small samples.

The likelihood function implied by the functional form of this model has essential singularities on the surfaces $(\sigma^2, \boldsymbol{\chi}) = (0, \mathbf{Y}/\beta)$, $(\sigma_1^2, \boldsymbol{\chi}) = (0, \mathbf{x})$, and $(\sigma_2^2, \boldsymbol{\chi}) = (0, \mathbf{w}/\delta)$. Maximizing the nonpathological part of the likelihood function subject to these constraints yields the estimates $\hat{\beta}^{D} = \mathbf{x}'\mathbf{Y}/\mathbf{x}'\mathbf{x}$, $\hat{\beta}^{R} = \mathbf{Y}'\mathbf{Y}/\mathbf{x}'\mathbf{y}$, and $\hat{\beta}^{IV} = \mathbf{Y}'\mathbf{w}/\mathbf{x}'\mathbf{w}$, respectively. For the models discussed previously, the essential singularities of the functional form are maximum likelihood points of the likelihood function of the structural form. The structural form of the model being discussed here, however, has a single maximum likelihood point, located at one of the three estimates $\hat{\beta}^{D}$, $\hat{\beta}^{R}$, or $\hat{\beta}^{IV}$. It will be shown that if all three estimates have the same sign, then the instrumental variables estimate is maximum likelihood if it lies between $\hat{\beta}^{D}$ and $\hat{\beta}^{R}$, that is, if it satisfies the errors-in-variables bound. Otherwise, the maximum likelihood estimate is one of the end points of the bound, $\hat{\beta}^{D}$ if $\hat{\beta}^{IV}$ is less in absolute value than $\hat{\beta}^{D}$; $\hat{\beta}^{R}$ if $\hat{\beta}^{IV}$ exceeds $\hat{\beta}^{R}$.

The structural form of the model makes use of the assumption that the incidental variables χ_t are independent normal variables with variance σ_χ^2. Equations (7.16), (7.17), and (7.18) then define a trivariate normal process with covariance matrix

$$\boldsymbol{\Sigma} = \begin{bmatrix} \beta^2\sigma_\chi^2 + \sigma_u^2 & \beta\sigma_\chi^2 & \beta\delta\sigma_\chi^2 \\ \beta\sigma_\chi^2 & \sigma_\chi^2 + \sigma_1^2 & \delta\sigma_\chi^2 \\ \beta\delta\sigma_\chi^2 & \delta\sigma_\chi^2 & \delta^2\sigma_\chi^2 + \sigma_2^2 \end{bmatrix} \tag{7.20}$$

and the likelihood function is

$$L(\boldsymbol{\Sigma}; \mathbf{Y}, \mathbf{x}, \mathbf{w}) \propto |\boldsymbol{\Sigma}|^{-T/2} \exp\left[-\frac{1}{2} \operatorname{tr} \boldsymbol{\Sigma}^{-1}\mathbf{S} \right] \tag{7.21}$$

where

$$\mathbf{S} = \begin{bmatrix} \mathbf{Y}'\mathbf{Y} & \mathbf{Y}'\mathbf{x} & \mathbf{Y}'\mathbf{w} \\ \mathbf{x}'\mathbf{Y} & \mathbf{x}'\mathbf{x} & \mathbf{x}'\mathbf{w} \\ \mathbf{w}'\mathbf{Y} & \mathbf{w}'\mathbf{x} & \mathbf{w}'\mathbf{w} \end{bmatrix}.$$

THEOREM 7.2 (INSTRUMENTAL VARIABLES). *Maximization of the likelihood function* (7.21) *subject to the constraint* (7.20) *with the variances* $\sigma_u^2, \sigma_1^2, \sigma_2^2$, *and* σ_χ^2 *nonnegative implies the following estimate of* β:

(*a*) *If* $\hat{\beta}^{IV}$ *and* $\hat{\beta}^{D}$ *are the same sign*

$$\hat{\beta} = \text{median}\,(\hat{\beta}^{D}, \hat{\beta}^{R}, \hat{\beta}^{IV})$$

(*b*) *If* $\hat{\beta}^{IV}$ *and* $\hat{\beta}^{D}$ *are opposite in sign*

if the smallest correlation is

$$\hat{\beta} = \begin{cases} \hat{\beta}^{R} & r_{xw}^2 \\ \hat{\beta}^{D} & r_{yw}^2 \\ \hat{\beta}^{IV} & r_{xy}^2 \end{cases}.$$

Proof: A well-known result is that if $\boldsymbol{\Sigma}$ is unconstrained, maximization of (7.21) with respect to $\boldsymbol{\Sigma}$ yields the estimate

$$\hat{\boldsymbol{\Sigma}} = \frac{\mathbf{S}}{T}.$$

This is the maximum likelihood estimate for the constrained problem if the constraints are not binding, that is, if there exist parameter values such that $\boldsymbol{\Sigma} = \mathbf{S}/T$. The three off-diagonal elements of this matrix determine the following estimates

$$\beta = \frac{\mathbf{Y}'\mathbf{w}}{\mathbf{x}'\mathbf{w}}$$

$$\sigma_\chi^2 = \frac{\mathbf{Y}'\mathbf{x}\mathbf{x}'\mathbf{w}}{T\mathbf{Y}'\mathbf{w}}$$

$$\delta = \frac{\mathbf{Y}'\mathbf{w}}{\mathbf{Y}'\mathbf{x}}.$$

These values together with the diagonal elements of Σ imply

$$\sigma_u^2 = \frac{\mathbf{Y}'\mathbf{Y}}{T} - \beta^2\sigma_x^2 = \frac{\mathbf{Y}'\mathbf{Y} - \mathbf{w}'\mathbf{Y}\mathbf{Y}'\mathbf{x}}{T\mathbf{x}'\mathbf{w}}$$

$$\sigma_1^2 = \frac{\mathbf{x}'\mathbf{x}}{T} - \sigma_x^2 = \frac{\mathbf{x}'\mathbf{x} - \mathbf{Y}'\mathbf{x}\mathbf{x}'\mathbf{w}}{T\mathbf{Y}'\mathbf{w}}$$

$$\sigma_2^2 = \frac{\mathbf{z}'\mathbf{z}}{T} - \delta^2\sigma_x^2 = \frac{\mathbf{w}'\mathbf{w} - \mathbf{Y}'\mathbf{w}\mathbf{w}'\mathbf{x}}{T\mathbf{Y}'\mathbf{x}}.$$

If the estimated variances are all positive, these are the maximum likelihood estimates. The constraint $\sigma_x^2 \geqslant 0$ implies

$$\frac{\mathbf{Y}'\mathbf{x}\mathbf{x}'\mathbf{w}}{\mathbf{Y}'\mathbf{w}} \geqslant 0.$$

This means that $\hat{\beta}^D$ and $\hat{\beta}^{IV}$ must have the same sign. Without loss of generality, we assume that $\mathbf{x}'\mathbf{Y} > 0$. Given $\mathbf{x}'\mathbf{Y} > 0$, the constraints $\sigma_u^2 \geqslant 0, \sigma_1^2 \geqslant 0, \sigma_2^2 \geqslant 0$ imply, respectively,

$$\hat{\beta}^R = \frac{\mathbf{Y}'\mathbf{Y}}{\mathbf{Y}'\mathbf{x}} \geqslant \frac{\mathbf{w}'\mathbf{Y}}{\mathbf{w}'\mathbf{x}} = \hat{\beta}^{IV} \tag{7.22}$$

$$\hat{\beta}^{IV} = \frac{\mathbf{w}'\mathbf{Y}}{\mathbf{w}'\mathbf{x}} \geqslant \frac{\mathbf{Y}'\mathbf{x}}{\mathbf{x}'\mathbf{x}} = \hat{\beta}^D \tag{7.23}$$

$$\mathbf{Y}'\mathbf{x} - \frac{\mathbf{Y}'\mathbf{w}\mathbf{w}'\mathbf{x}}{\mathbf{w}'\mathbf{w}} \geqslant 0. \tag{7.24}$$

If one or more of the four inequalities is violated, it is not possible to have $\hat{\boldsymbol{\Sigma}} = \mathbf{S}/T$. In that case, the maximum occurs on the boundary of the constraint set. It could occur interior to the constraint set only if the likelihood function had more than one local maximum, which it does not, in fact, have. The constrained maximum can be found by imposing the constraints one at a time and selecting the constraint that yields the highest likelihood value, provided no other constraints are violated once one of the constraints is imposed.

Now consider the maximization of the likelihood function subject to one of the constraints $\sigma_u^2 = 0$, $\sigma_1^2 = 0$, or $\sigma_2^2 = 0$. If $\sigma_u^2 = 0$, (7.16) becomes $Y_t = \beta\chi_t$ without error and can be substituted into (7.17) and (7.18):

$$x_t = \frac{Y_t}{\beta} + u_{1t}$$

$$w_t = \frac{Y_t\delta}{\beta} + u_{2t}.$$

Maximization of the likelihood function of this normal process leads to the following likelihood value and estimate of β:

$$\sigma_u^2 = 0 : \hat{\beta} = \frac{\mathbf{Y}'\mathbf{Y}}{\mathbf{x}'\mathbf{Y}}, \qquad L_1 \propto \left[(\mathbf{Y}'\mathbf{Y})(\mathbf{x}'\mathbf{M}_Y\mathbf{x})(\mathbf{w}'\mathbf{M}_Y\mathbf{w})\right]^{-T/2}$$

where $\mathbf{M}_Y = \mathbf{I} - \mathbf{Y}(\mathbf{Y}'\mathbf{Y})^{-1}\mathbf{Y}'$. Similarly, given the other constraints,

$$\sigma_1^2 = 0 : \hat{\beta} = \frac{\mathbf{x}'\mathbf{Y}}{\mathbf{x}'\mathbf{x}} \qquad L_2 \propto \left[(\mathbf{x}'\mathbf{x})(\mathbf{Y}'\mathbf{M}_x\mathbf{Y})(\mathbf{w}'\mathbf{M}_x\mathbf{w})\right]^{-T/2}$$

$$\sigma_2^2 = 0 : \hat{\beta} = \frac{\mathbf{w}'\mathbf{Y}}{\mathbf{w}'\mathbf{x}}, \qquad L_3 \propto \left[(\mathbf{w}'\mathbf{w})(\mathbf{Y}'\mathbf{M}_w\mathbf{Y})(\mathbf{x}'\mathbf{M}_z\mathbf{x})\right]^{-T/2}$$

The constraint $\sigma_x^2 = 0$ implies the obviously inferior likelihood value $[(\mathbf{Y}'\mathbf{Y})(\mathbf{x}'\mathbf{x})(\mathbf{w}'\mathbf{w})]^{-T/2}$, and that constraint need not be considered further.

If we neglect the exponent and multiply by $(\mathbf{Y}'\mathbf{Y})(\mathbf{x}'\mathbf{x})(\mathbf{w}'\mathbf{w})$, the three likelihood values become, respectively, $[(1-r_{xy}^2)(1-r_{yw}^2)]^{-1}$, $[(1-r_{xy}^2)(1-r_{xw}^2)]^{-1}$and $[(1-r_{yw}^2)(1-r_{xw}^2)]^{-1}$. The first of these three numbers is the largest if r_{xw}^2 is the smallest of the squared correlations, the second if r_{yw}^2 is the smallest, and the third if r_{xy}^2 is the smallest. This establishes part (*b*) of the theorem.

All that remains to be shown to prove part (*a*) is that if $\hat{\beta}^{IV}$ and $\hat{\beta}^D$ have the same sign, then the violation of inequality (7.24) implies that r_{xy}^2 is the smallest correlation; and if $|\hat{\beta}^{IV}| < |\hat{\beta}^D|$, then r_{yw}^2 is the smallest correlation; and if $|\hat{\beta}^R| < |\hat{\beta}^{IV}|$, then r_{xw}^2 is the smallest. For convenience, we continue to assume $r_{xy} > 0$. If inequality (7.24) is violated, then in terms of correlations $0 < r_{xy} < r_{yw}r_{wx}$. But using $r_{yw} < 1$ and $r_{wx} < 1$, we obtain $r_{xy} < r_{yw}$ and $r_{xy} < r_{wx}$. Similarly, $|\hat{\beta}^{IV}| < |\hat{\beta}^D|$ implies $0 < r_{yw}/r_{xw} < r_{xy}$, or $r_{yw}^2 < r_{xy}^2 r_{xw}^2$, and r_{yw}^2 is the smallest; $|\hat{\beta}^R| < |\hat{\beta}^{IV}|$ implies $0 < r_{xy}^{-1} < r_{yw}/r_{xw}$ or $r_{xw}^2 < r_{yw}^2 r_{xy}^2$, and r_{xw}^2 is the smallest.

This theorem may be used to evaluate a statement that is often made about the use of instrumental variables estimators. It is often suggested that $\mathbf{w}$ is a "good" instrument if it is highly correlated with $\mathbf{x}$. On the other hand, if $\mathbf{w}$ and $\mathbf{x}$ are highly correlated, the instrumental variables estimate may not be much different from the direct estimate, and it is then difficult to see how the estimate could be an improvement over direct least squares. The result just proved does shed some light on this puzzle. Normalize the data so that the observed vectors all have length one, and let the three observed correlations be the numbers r_{xy}, r_{xw}, r_{yw}. Since the correlation matrix is positive definite, it must be the case that

$$0 \leqslant \det\begin{bmatrix} 1 & r_{xy} & r_{yw} \\ r_{xy} & 1 & r_{xw} \\ r_{yw} & r_{xw} & 1 \end{bmatrix} = 1 + 2r_{xy}r_{xw}r_{yw} - r_{yw}^2 - r_{xw}^2 - r_{xy}^2,$$

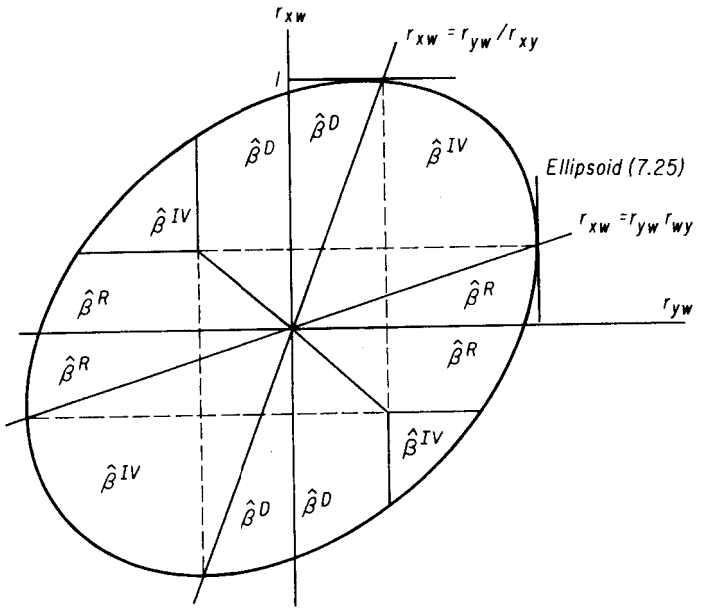

Fig. 7.5 Choice of estimates: the instrumental variables model.

which can be rewritten as

$$[r_{xw}r_{yw}]\begin{bmatrix}1 & -r_{xy}\\ -r_{xy} & 1\end{bmatrix}\begin{bmatrix}r_{xw}\\ r_{yw}\end{bmatrix}\leqslant 1-r_{xy}^2. \tag{7.25}$$

Holding fixed r_{xy}, inequality (7.25) implies Figure 7.5. By inspection of this figure, a high value of r_{xw} does not guarantee that $\hat{\beta}^{IV}$ is maximum likelihood. In fact, an increase in r_{xw} may shift the estimate from $\hat{\beta}^{IV}$ to $\hat{\beta}^{D}$. That is to say, the maximum likelihood estimate may be direct least squares even when r_{xw} is high, if r_{yw} is low. Note, however, that if r_{xw} is one, then the only feasible point is on the ray $r_{xw}=r_{yw}/r_{xy}$, and $\hat{\beta}^{IV}=\hat{\beta}^{D}$. In that sense, the statement about r_{xw} is valid. (Each ray out of the origin corresponds to a different value for $\hat{\beta}^{IV}$, increasing in the clockwise direction.)

By the way, the maximum likelihood estimate is a peculiar discontinuous function of the data if $\hat{\beta}^{IV}$ and $\hat{\beta}^{D}$ are opposite in sign. A small change in the data induces a shift from $\hat{\beta}^{D}$ to $\hat{\beta}^{IV}$, and thus a reversal of the sign of the estimate. Such a property is intuitively nonsensical. May we surmise that measures of uncertainty are in that case enormous, so that the discontinuity is small when measured in units of uncertainty? Or is the mode of the marginal likelihood function the more appropriate point

7.5 Multiple Proxy Variables

A generalization of the preceding is the multiple proxy variable model described by the equations

$$Y_t=\beta\chi_t+\gamma z_t+u_t \tag{7.26}$$

$$\mathbf{x}_t=\boldsymbol{\delta}\chi_t+\boldsymbol{\varepsilon}_t \tag{7.27}$$

$$\chi_t=\omega z_t+e_t. \tag{7.28}$$

Equations (7.26) and (7.28) are unchanged from the discussion in Section 7.3. Equation (7.27) is as before, except that it describes the generation of an m-dimensional vector of proxies rather than a single one. We take u_t, $\boldsymbol{\varepsilon}_t$, and e_t to be independent, normal random variables with means zero and variances σ_u^2, $\boldsymbol{\Omega}$, and σ_χ^2, respectively, where $\boldsymbol{\Omega}$ is an $m\times m$ covariance matrix.

The essential singularities of the joint likelihood function of the functional form occur when σ_u^2 or when the determinant of $\boldsymbol{\Omega}$ is zero. Since I have been unable to connect these points with the maximum likelihood points of the structural form, I do not discuss them further here.

The structural form of the model makes use of (7.28) to integrate out the incidental parameters. In effect, this assumes that $(Y_t,\mathbf{x}_t')$ has a multivariate normal distribution with mean and variance

$$E(Y_t,\mathbf{x}_t')=(\beta\omega z_t+\gamma z_t,\omega z_t\boldsymbol{\delta}')$$

$$V(Y_t,\mathbf{x}_t')=\begin{bmatrix}\beta^2\sigma_\chi^2+\sigma_u^2 & \beta\sigma_\chi^2\boldsymbol{\delta}'\\ \boldsymbol{\delta}\beta\sigma_\chi^2 & \boldsymbol{\delta}\sigma_\chi^2\boldsymbol{\delta}'+\boldsymbol{\Omega}\end{bmatrix}.$$

As usual, the basic identification problem arises. The distribution of the observables is unaffected by a scale change in $(\beta,\boldsymbol{\delta}')$ offset by a scale change in (ω,σ_χ). Note also that there are proportionality constraints on this $(m+1)$-variate distribution, since both the covariance between Y_t and $\mathbf{x}_t$ $(\beta\sigma_\chi^2\boldsymbol{\delta}')$ and the vector of regression coefficients $(\omega\boldsymbol{\delta}')$ are proportional to $\boldsymbol{\delta}$.

Unconstrained maximum likelihood estimate of the parameters of this process leads to the equations

$$\beta\omega+\gamma=(\mathbf{z}'\mathbf{z})^{-1}\mathbf{z}'\mathbf{Y}$$

$$\omega\boldsymbol{\delta}'=(\mathbf{z}'\mathbf{z})^{-1}\mathbf{z}'\mathbf{X}$$

$$\beta\sigma_\chi^2\boldsymbol{\delta}'=\frac{\mathbf{Y}'\mathbf{M}_z\mathbf{X}}{T}$$

$$\beta^2\sigma_\chi^2+\sigma_u^2=\frac{\mathbf{Y}'\mathbf{M}_z\mathbf{Y}}{T}$$

$$\boldsymbol{\delta}\sigma_\chi^2\boldsymbol{\delta}'+\boldsymbol{\Omega}=\frac{\mathbf{X}'\mathbf{M}_z\mathbf{X}}{T}$$

where $M_z = \mathbf{I} - \mathbf{z}(\mathbf{z}'\mathbf{z})^{-1}\mathbf{z}'$ and $\mathbf{X}$ is the $T \times m$ matrix of m proxy variables.

For the sake of producing an analytical result, we assume that the sample moments satisfy the proportionality constraints of the process

$$(\mathbf{z}'\mathbf{z})^{-1}\mathbf{z}'\mathbf{X} \propto \mathbf{Y}'\mathbf{M}_z\mathbf{X}.$$

THEOREM 7.3. *If the sample moments satisfy the proportionality constraints, all "reverse regression" estimates of γ are identical. (A reverse regression estimate is computed by regressing one of the proxies on $\mathbf{Y}$ and $\mathbf{z}$, and then transforming the estimated equation to write $\mathbf{Y}$ as a function of $\mathbf{z}$ and the proxy.)*

Proof: A reverse regression estimate of γ can be written

$$\hat{\gamma}_i^R = (\mathbf{z}'\mathbf{z})^{-1}\mathbf{z}'\mathbf{Y} - (\mathbf{z}'\mathbf{z})^{-1}\mathbf{z}'\mathbf{x}_i\hat{\beta}_i^R$$

where $\hat{\beta}_i^R$ is the reverse regression estimate of β:

$$\hat{\beta}_i^R = (\mathbf{Y}'\mathbf{M}_z\mathbf{x}_i)^{-1}\mathbf{Y}'\mathbf{M}_z\mathbf{Y}.$$

The proportionality constraint is

$$\frac{\mathbf{z}'\mathbf{x}_i}{\mathbf{Y}'\mathbf{M}_z\mathbf{x}_i} = \frac{\mathbf{z}'\mathbf{x}_j}{\mathbf{Y}'\mathbf{M}_z\mathbf{x}_j}$$

which implies $\mathbf{z}'\mathbf{x}_i\hat{\beta}_i^R = \mathbf{z}'\mathbf{x}_j\hat{\beta}_j^R$. Using this in the equation for $\hat{\gamma}_i^R$ produces the desired equality $\hat{\gamma}_i^R = \hat{\gamma}_j^R$. This is true, by the way, not only for any one proxy but for any linear combination as well.

THEOREM 7.4 (MULTIPLE PROXY VARIABLES). *If the sample moments satisfy the proportionality constraints, the values of γ corresponding to the maximum of the likelihood function are bounded on one side by the (unique) reverse regression estimate*

$$\hat{\gamma}^R = (\mathbf{z}'\mathbf{z})^{-1}\mathbf{z}'\mathbf{Y} - (\mathbf{z}'\mathbf{z})^{-1}\mathbf{z}'\mathbf{x}_i\hat{\beta}^R$$

where $\hat{\beta}^R = (\mathbf{Y}'\mathbf{M}_z\mathbf{x}_i)^{-1}\mathbf{Y}'\mathbf{M}_z\mathbf{Y}$, and on the other by

$$\hat{\gamma}^D = (\mathbf{z}'\mathbf{z})^{-1}\mathbf{z}'\mathbf{Y} - (\mathbf{z}'\mathbf{z})^{-1}\mathbf{z}'\mathbf{x}_i\hat{\beta}^R R^2$$

where R^2 is the multiple correlation coefficient and $\hat{\gamma}^D$ is the estimate computed when $\mathbf{Y}$ is regressed on $\mathbf{z}$ and all the proxy variables

$$R^2 = \frac{\mathbf{Y}'\mathbf{M}_z\mathbf{X}(\mathbf{X}'\mathbf{M}_z\mathbf{X})^{-1}\mathbf{X}'\mathbf{M}_z\mathbf{Y}}{\mathbf{Y}'\mathbf{M}_z\mathbf{Y}}.$$

Notice also that $\hat{\gamma}^D$ satisfies the equation

$$(\hat{\gamma}^D - \hat{\gamma}^S) = R^2(\hat{\gamma}^R - \hat{\gamma}^S)$$

Proof: Select any value of β and δ_1 and solve the equations for the other parameters

$$\omega = (\mathbf{z}'\mathbf{z})^{-1}\mathbf{z}'\mathbf{x}_1\delta_1^{-1}$$

$$\gamma = (\mathbf{z}'\mathbf{z})^{-1}\mathbf{z}'\mathbf{Y} - (\mathbf{z}'\mathbf{z})^{-1}\mathbf{z}'\mathbf{x}_1\frac{\beta}{\delta_1}$$

$$\sigma_x^2 = \frac{\mathbf{Y}'\mathbf{M}_z\mathbf{x}_1}{T\beta\delta_1}$$

$$\sigma_u^2 = \frac{\mathbf{Y}'\mathbf{M}_z\mathbf{Y}}{T} - \frac{\mathbf{Y}'\mathbf{M}_z\mathbf{x}_1}{T}\frac{\beta}{\delta_1}$$

$$\boldsymbol{\Omega} = \frac{\mathbf{X}'\mathbf{M}_z\mathbf{X}}{T} - \boldsymbol{\delta}\sigma_x^2\boldsymbol{\delta}'$$

$$= \frac{\mathbf{X}'\mathbf{M}_z\mathbf{X}}{T} - \frac{\mathbf{X}'\mathbf{M}_z\mathbf{Y}\sigma_x^2\mathbf{Y}'\mathbf{M}_z\mathbf{X}}{T^2\sigma_x^4\beta^2}$$

$$= \frac{\mathbf{X}'\mathbf{M}_z\mathbf{X}}{T} - \mathbf{X}'\mathbf{M}_z\mathbf{Y}\left(\frac{\delta_1}{T\beta\mathbf{Y}'\mathbf{M}_z\mathbf{x}_1}\right)\mathbf{Y}'\mathbf{M}_z\mathbf{X}.$$

The constraint $\sigma_u^2 > 0$ implies

$$\frac{\beta}{\delta_1} < \frac{\mathbf{Y}'\mathbf{M}_z\mathbf{Y}}{\mathbf{Y}'\mathbf{M}_z\mathbf{x}_1} = \hat{\beta}^R.$$

The constraint $\boldsymbol{\Omega}$ positive definite implies[3]

$$\left(\frac{\mathbf{Y}'\mathbf{M}_z\mathbf{x}_1\beta}{\delta_1}\right)^{-1} < \left(\mathbf{Y}'\mathbf{M}_z\mathbf{X}(\mathbf{X}'\mathbf{M}_z\mathbf{X})^{-1}\mathbf{X}'\mathbf{M}_z\mathbf{Y}\right)^{-1}$$

that is

$$\frac{\beta}{\delta_1} > \frac{\mathbf{Y}'\mathbf{M}_z\mathbf{X}(\mathbf{X}'\mathbf{M}_z\mathbf{X})^{-1}\mathbf{X}'\mathbf{M}_z\mathbf{Y}}{\mathbf{Y}'\mathbf{M}_z\mathbf{x}_1} = \hat{\beta}^R R^2.$$

These two constraints on β/δ_1 determine the hypothesized constraints on γ.

An interesting implication of Theorem 7.4 is that extra proxies are not enough to identify the coefficients, but the bounds for the coefficients are reduced, depending on the multiple R^2 of $\mathbf{Y}$ on all the proxies. If the proxies were independent of each other, that is, if $\boldsymbol{\Omega}$ were diagonal, extra proxies would be more helpful, however.[4]

[3] $\boldsymbol{\Omega} = \mathbf{S} - \boldsymbol{\delta}\sigma^2\boldsymbol{\delta}'$ is positive definite if for any vector ψ, $0 < \psi'\mathbf{S}\psi - \psi'\boldsymbol{\delta}\sigma^2\boldsymbol{\delta}'\psi$, which in turn requires $\sigma^2 < \max_\psi \psi'\mathbf{S}\psi/\psi'\boldsymbol{\delta}\boldsymbol{\delta}'\psi = 1/\boldsymbol{\delta}'\mathbf{S}^{-1}\boldsymbol{\delta}$.

[4] This does not lend itself to a very tractable analysis and is not discussed here. The constraints that $\boldsymbol{\Omega}$ is proportional to the identity matrix and $\boldsymbol{\delta}$ is the vector of ones does lead

7.6 Errors in Many Variables

Another more general model has all k variables measured with error:

$$Y_t = \beta' \chi_t + u_t \tag{7.29}$$

$$\mathbf{x}_t = \chi_t + \varepsilon_t \tag{7.30}$$

where variables are measured around their means, χ_t is a k-dimensional vector of incidental parameters, $\mathbf{x}_t$ is a vector of measurements of χ_t, and ε_t is a vector of measurement errors distributed normally with mean vector zero and diagonal covariance matrix $\mathbf{D}$. The structural form of this model makes use of the assumption that χ_t is normal with mean zero and covariance Ω.

Thus the vector $(Y_t, \mathbf{x}_t)$ has covariance matrix equal to

$$\Sigma = \begin{pmatrix} \sigma_u^2 + \beta'\Omega\beta & \beta'\Omega \\ \Omega\beta & \mathbf{D} + \Omega \end{pmatrix}.$$

The maximum likelihood estimate can be found as before by setting Σ equal to the sample moments

$$\mathbf{S} = \begin{bmatrix} \mathbf{Y}'\mathbf{Y} & \mathbf{Y}'\mathbf{X} \\ \mathbf{X}'\mathbf{Y} & \mathbf{X}'\mathbf{X} \end{bmatrix} \cdot T^{-1}$$

where $\mathbf{X}$ is the $T \times k$ matrix of measurements of $\chi_t, t = 1, \ldots, T$. Given the diagonal matrix $\mathbf{D}$, we can solve for the other parameters as

$$\hat{\beta} = \frac{\Omega^{-1}\mathbf{X}'\mathbf{Y}}{T} = (\mathbf{X}'\mathbf{X} - T\mathbf{D})^{-1}\mathbf{X}'\mathbf{Y} \tag{7.31}$$

$$\hat{\Omega} = \frac{\mathbf{X}'\mathbf{X}}{T} - \mathbf{D}$$

$$\hat{\sigma}^2 = \frac{\mathbf{Y}'\mathbf{Y}}{T} - \beta'\Omega\beta = \frac{\mathbf{Y}'\mathbf{Y}}{T} - \frac{\mathbf{Y}'\mathbf{X}(\mathbf{X}'\mathbf{X}T^{-1} - \mathbf{D})^{-1}\mathbf{X}'\mathbf{Y}}{T^2}.$$

Any value of $\hat{\beta}$ corresponding to any nonnegative diagonal matrix $\mathbf{D}$ is, therefore, a maximum likelihood point provided that $\hat{\Omega}$ is positive semi-definite and $\hat{\sigma}_u^2$ is nonnegative.

As far as I know, an exact description of the set of maximum likelihood estimates is not known. It is possible to compute $k-1$ "reverse" regressions in which one of the explanatory variables is used as the dependent variables. Each of these reverse regressions is feasible, given a suitably chosen $\mathbf{D}$, as is the direct regression. It is natural to conjecture that the [illegible] is the convex hull of these k regressions. This is correct if all

k regressions are in the same orthant, a result implicitly due to Frisch (1934), proved by Reiersol (1945) and Koopmans (1937), and discussed by Malinvaud (1970, p. 43). When the estimates lie in different orthants, the feasible region is apparently unbounded with sides generated by the k regressions.

It is interesting, in closing, to contrast Equation (7.31) with the matrix-weighted average that is the posterior mean of β given a normal prior with mean zero and precision matrix $\mathbf{D}$. For the prior-dependent problem the least-squares estimate is shrunk toward the origin by adding to the $\mathbf{X}'\mathbf{X}$ matrix a positive diagonal matrix. For the errors-in-variables model the least-squares estimate is blown away from the origin by subtracting from the $\mathbf{X}'\mathbf{X}$ matrix a positive diagonal matrix. I have sometimes chided my colleagues who are enamored of ridge regression that the net effect of the two forces may mean that least squares is just right!

7.7 Priors and Proxies

The interesting problems of the relationship between prior information and proxy variables have yet to be discussed. To give an example, when "bad" estimates are implied by a data set, they are sometimes "explained" with reference to the errors-in-variables model. "The coefficient is smaller than it ought to be because of the errors-in-variable attenuation," or "the theory would have worked better if we could have found more appropriate proxy variables." Are these statements and statements like them appropriate? Can it be that there is a form of prior information that allows us to discard or discount especially unreasonable estimates but to ignore the errors-in-variables issues if the coefficients are relatively consistent with the priors? Are more proxies necessarily better than less? What are the inferential consequences of searching for a proxy until one is found that "works"?

Most of these questions refer to measures of dispersion of posterior distribution. To get close to answers, we must generate approximate measures of dispersion. The last time in this chapter that such measures were discussed was in Section 7.2, in which we reported Lindley and El-Sayyad's (1968) analysis of the errors-in-variables model in which diffuse priors were assumed for the regression parameter. Their analysis could be generalized to the more complicated proxy problems, but that work remains to be done. We report here a few simple results that constitute a minor foray into an intriguing and important field of research.

With reference to the simple proxy variable model discussed in Section 7.3, the following proposition can be explored: "a poor proxy is worse than none." With a proper prior distribution this cannot be the case, just as it is *never* better in an *inference* problem to leave a variable out of a regression, unless there is some cost of observation, or in a more general framework,

cost of constructing a prior. But it may be the case that the regression computed with the proxy variable left out is a better approximation to the location of the posterior distribution than is the regression with the proxy variable in the equation.

In the material on interpretive searches we began with a prior located at the origin and concluded that a variable may be omitted if its estimated coefficient was not far from zero. A proxy variable may be a candidate for omission (a) if its coefficient is small, (b) if its coefficient is far from the value we would have expected if we had perfect measurement, or (c) if the coefficient on $\mathbf{z}$ is closer to the value we expected when the proxy is omitted. Although each has been used in practice and each has a certain appeal, none is an obviously valid procedure. Prior information is clearly the foundation of these procedures, and we now turn to a consideration of how priors affect our estimates.

First we observe that prior information about either β or δ can be used to normalize the vector (β,δ), but otherwise, because of the identification problem, it has no effect on the modes of the posterior. It is not uncommon, however, to have prior information about both β and δ. A prior for δ is likely to be located at one, of course, and it may be that the researcher has some more-or-less distinct ideas about β as well. Let us take the extreme case in which β and δ are known exactly. This implies no constraints on the maximum likelihood estimates if no inequalities are violated. In that event, the appropriate estimate of γ is

$$\tilde{\gamma}=(\mathbf{z}'\mathbf{z})^{-1}\mathbf{z}'\mathbf{Y}-(\mathbf{z}'\mathbf{z})^{-1}\mathbf{z}'\mathbf{x}\frac{\beta}{\delta}, \tag{7.32}$$

which is the least-squares estimate of γ with the constraint that the coefficient on $\mathbf{x}$ is equal to the ratio β/δ.[5]

In the second best world, in which we can choose between regressions with or without the $\mathbf{x}$ variable included, we need only determine which of the two resulting estimates is closer to (7.32). This is equivalent to selecting the regression with $\mathbf{x}$ omitted if the least-squares coefficient on $\mathbf{x}$ is closer to zero than to β/δ. Incidentally, it is necessary only to know the sign of β/δ to know that it is better to omit $\mathbf{x}$ if its coefficient has the wrong sign.

We may now consider situations in which there is prior information about γ (possibly zero). We begin with the presumption that if the least-squares estimate of γ is close to the value we expect, we may conclude first that the coefficient on x is likely to be close to β and second that it is desirable to adjust the prior estimate of γ in the direction of the least-squares estimate. If, on the other hand, the proxy doesn't "work," it produces a coefficient far from the prior estimate. We would be led to conclude that there is little information about β or about γ generated by the experiment.

[5]Note that $\tilde{\gamma}$ is an unbiased estimator of γ.

In fact, it is not enough to have even exact prior information about γ, since the basic identification problem remains; if we multiply the vector (β,δ) by some constant and make an offsetting change in (σ_χ^2,ω), we do not alter the distribution of the observables.[6] The fact that prior information about γ alone is of little value for estimating β is fairly obvious when we consider that $\hat{\gamma}$ may be close to the true value merely if ω is small, that is, if $\boldsymbol{\chi}$ and $\mathbf{z}$ are uncorrelated. It is necessary, but it is not sufficient.

If *both* γ and δ are known, there is no identification problem, and we may solve for the maximum likelihood estimates as

$$\omega=\frac{(\mathbf{z}'\mathbf{z})^{-1}\mathbf{z}'x}{\delta}$$

$$\beta=\frac{(\mathbf{z}'\mathbf{z})^{-1}\mathbf{z}'(\mathbf{Y}-\gamma\mathbf{z})}{\omega}$$

$$=\delta(\mathbf{z}'\mathbf{x})^{-1}\mathbf{z}'(\mathbf{Y}-\gamma\mathbf{z})$$

$$\sigma_\chi^2=\frac{\mathbf{x}'\mathbf{M}_z\mathbf{Y}}{T\beta\delta}$$

$$\sigma_u^2=\frac{\mathbf{Y}'\mathbf{M}_z\mathbf{Y}}{T}-\beta^2\sigma_\chi^2$$

$$\sigma_\varepsilon^2=\frac{\mathbf{x}'\mathbf{M}_z\mathbf{x}}{T}-\delta^2\sigma_\chi^2$$

[6]An interesting observation is that these assumptions determine only a subset of the set of all bivariate normal processes. Denoting the means and variances of $(Y_t-\gamma z_t,x_t)$ by

$$(m_1,m_2)=z_t(\beta\omega,\delta\omega)$$

$$\begin{bmatrix}v_{11} & v_{12}\\ v_{12} & v_{22}\end{bmatrix}=\begin{bmatrix}\beta^2\sigma_\chi^2+\sigma_u^2 & \beta\delta\sigma_\chi^2\\ \beta\delta\sigma_\chi^2 & \delta^2\sigma_\chi^2+\sigma_\varepsilon^2\end{bmatrix},$$

from the fact that the variances are positive we may derive the constraints

$$\frac{v_{12}}{m_1m_2}>0;\quad v_{11}-\frac{m_1}{m_2}v_{12}>0;\quad v_{22}-\frac{m_2}{m_1}v_{12}>0.$$

It can be shown that if the direct and reverse regressions do not bracket γ, then the data favor a bivariate normal process excluded from this class.

provided the variances are positive. Note that β is an instrumental variables estimator, with $\mathbf{z}$ used as an instrument for $\mathbf{x}$ in the regression of $\mathbf{Y} - \gamma\mathbf{z}$ on $\mathbf{x}$.

An appealing proposition is that if $\mathbf{x}$ is a "good" measurement of χ, then when we regress $\mathbf{Y}$ on $\mathbf{x}$ and $\mathbf{z}$, the coefficient on $\mathbf{z}$ should be close to the known value of γ. If it is not, then it seems unlikely that we could obtain much information about β. Intuition thus suggests that the uncertainty in β may be related to the difference between γ and $\hat{\gamma}$. This, like a long list of other interesting questions, will remain unanswered until some appropriate approximations of the posterior dispersion are available.

CHAPTER 8

DATA-SELECTION SEARCHES

Theoretical models are often vague or have nothing at all to say about the choice of particular observations. Even less frequently do they suggest the circumstances in which two or more observations can be regarded as independent pieces of relevant information. It is thus necessary for the empirical worker both to select a subset of potential observations and to determine the extent to which observations are correlated. To put it another way, the researcher must identify observations or transformations of observations that can be considered to be independent replications of an unchanging "experiment." In practice, this may mean estimating coefficients with different subsets or different transformations of the data set and selecting the result that appears best according to some criteria. We call this a data-selection search.

The fact that this process is data dependent obviously has consequences for the interpretation of the final result of a data-selection search. It seems clear that when the data evidence is partly spent to pick a data set, the regression equation that is finally selected to convey the data evidence at least overstates the precision of the evidence and likely distorts it as well. The function of this chapter is thus to describe the inferences that are appropriate when some of the data are discarded or when the data are transformed by a data-dependent function.

In the case of interpretive searches, a constrained regression can at best approximate the location of a