Paper for Presentation next week: Graddy, The Fulton Fish market, Journal of Economic Perspectives. Pay particular attention to the regression analysis, and the identification issues that are confronted and (one would hope) solved.

Intro Comments:*

The purpose of this lecture is to run through basic econometrics with a more applied commentary than is normally the case in a econometrics course.

For the rest of the course I recommend reading Manski's "Identification Problems in the Social Sciences". In fact if you do nothing else in this course, read this book. It is short and awesome. It also provides the conceptual basis for almost everything that happens from now on.

Another book I ecourage you to have a look at is Hayashi's "Econometrics". I find it is the best exposition of basic econometric tools avaliable in a textbook. It is exceptionally modern, insofar as everything is hung off a GMM structure. This structure is how many younger applied researchers think (Allan and I included).

*Co-written by Allan Collard-Wexler and John Asker using Joerg Stoye's notes quite closely.

Regression

The definition of a regression is finding the probability distribution of dependent variable y conditional on independent (or covariates) x:

$$P(y|x)$$

Note that this includes linear and non-linear models we will look at, finding expectation or quantiles and so on.

Often we will be interested in the expectation: $E(y|x)$, I will give Savage's motivation for this.

Statistics is Decision Theory

1. Suppose I am drilling for Oil. Output of oil y given geological features x is given by the probability distribution $P(y|x)$. If the firm I am working for is risk-neutral, what features of $P(y|x)$ do I need to know to make my decision?

   The firm is going to solve the problem:

   $$\max_x \pi(x) = \int_0^\infty yP(y|x)dy$$

   and the left hand side is just $E(y|x)$, which is a sufficient statistics in order to choose the best site x.

2. Define the *Loss Function* L that the decision maker is using as a function of the difference between θ and y. For instance, the decision maker may have a minmax loss function (he cares about minimizing the worse possible outcome):

   $$L(\theta) = (\max |y - \theta|)$$

   What is the best way to pick θ in this case given x?

   $$\min_{\theta|x} L(\theta) = \min_\theta (\max |y - \theta|)$$

From here it is clear that if the support of $P(y|x)$ is bounded by $\underline{y}$ and $\bar{y}$, then $\theta = \underline{y} + \frac{1}{2}(\bar{y} - \underline{y})$, i.e. the middle of the support.

3. What about a loss function of $L(\theta) = |y - \theta|$.

   This is equivalent to minimizing

$$\int_{-\infty}^{\theta} (\theta - y) P(y|x) dy + \int_{\theta}^{\infty} (y - \theta) P(y|x) dy$$

   So the first order condition of L, $\frac{\partial L}{\partial \theta} = 0$ is :

$$0 = \int_{-\infty}^{\theta} P(y|x) dy + (\theta - y) + - \int_{\theta}^{\infty} P(y|x) dy + (y - \theta)$$

$$0 = \int_{-\infty}^{\theta} P(y|x) dy - \int_{\theta}^{\infty} P(y|x) dy$$

   Or

$$\int_{-\infty}^{\theta} P(y|x) dy = \int_{\theta}^{\infty} P(y|x) dy$$

   Here it is clear that $\theta(x) = Med[y|x]$.

4. Finally, what about the loss function $L(\theta) = (y - \theta)^2$?

   We want to find $\theta(x) = argmin \int_y (y-\theta)^2 P(y|x) dy$.

Taking the first derivative with respect to θ and setting it to zero, we get:

$$\int_y -2(y-\theta)P(y|x)dy = 0$$

$$\int_y yP(y|x)dy - \theta \int_y P(y|x)dy = 0$$

$$E(y|x) = \theta$$

So this is a reason to focus on the expected value.

OLS and Variations

OLS: Definition and Finite Sample Properties

We model the relation between a dependent variable y and regressors x. Realizations of these will be denoted by subscripts i, i.e. (y_i, x_i). x_i may be a vector $\mathbf{x}_i = (x_{i1}, \ldots x_{iK})$. A model specifies the relation between these random variables up to some unknown components. These components specifically include parameters of interest. We will attempt to estimate the parameters of interest by means of a size n sample of realizations $(y_i, \mathbf{x}_i)$.

In this first section only, we will review finite sample analysis.

It can be helpful to think in terms of data matrices

$$\mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, \mathbf{X} = \begin{bmatrix} x_{11} & \cdots & x_{1K} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nK} \end{bmatrix}, \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

that stack sample realizations.

Assume that the true relationship between x and y is linear:

Assumption 1: Linearity

$$y_i = \mathbf{x}_i\beta + \varepsilon_i$$

or equivalently,

$$\mathbf{y} = \mathbf{X}\beta + \varepsilon.$$

Of course, this assumption is restrictive only in conjunction with some assumption on the random variable ε_i; else it could be read as definition of ε_i. Note that the assumption captures all models that can be *transformed* into linear ones. A well known example is Cobb-Douglas production functions:

$$y_i = A_i L_i^{\beta} K_i^{1-\beta}$$

To begin, we furthermore assume that $\mathbf{x}_i$ is strictly exogenous:

Assumption 2: Strict Exogeneity

$$\mathbb{E}(\varepsilon_i|\mathbf{X}) = 0$$

or equivalently

$$\mathbb{E}(\varepsilon|\mathbf{X}) = \mathbf{0}.$$

Notice that in the presence of a constant regressor, setting the expectation equal to zero is a normalization. Strict exogeneity is a very strong assumption that we will relax later on. For one example, it cannot be fulfilled when the regressors include lagged dependent variables.

Assumption 3: Rank Condition

$$rank(\mathbf{X}) = K \text{ a.s.}$$

If this assumption is violated, one regressor is linearly dependent on the others (at least with positive probability). It is intuitively clear that in this case, we cannot disentangle the regressors' individual effects on y. Indeed, we will later think of this assumption as an *identification condition*.

Assumption 4: Spherical Error

$$\begin{aligned} \mathbb{E}(\varepsilon_i^2|\mathbf{X}) &= \sigma^2 > 0 \\ \mathbb{E}(\varepsilon_i\varepsilon_j|\mathbf{X}) &= 0 \end{aligned}$$

or equivalently,

$$\mathbb{E}\left(\varepsilon\varepsilon'|\mathbf{X}\right) = \sigma^2\mathbf{I}_n, \sigma^2 > 0.$$

Thus we assume the error process to be conditionally independent and homoskedastic.

Our setup treats $\mathbf{x}_i$ as a random variable. This leads to us conditioning statements on $\mathbf{x}_i$ which is a feature you might not have seen this before. This is because the early (and some textbooks') development of OLS assumes that regressors are fixed. This assumption makes life very slightly easier but is generally inappropriate for nonexperimental data.

The OLS estimate of β is derived as

$$\begin{aligned}\mathbf{b} &\equiv \arg\min_{\beta} \sum_i (y_i - \mathbf{x}_i'\beta)^2 \\ &= \arg\min_{\beta} (\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta)\end{aligned}$$

where the sum to be minimized is known as *sum of squared residuals*.

We can solve for $\mathbf{b}$ in closed form:

$$\begin{aligned}(\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta) &= \mathbf{y}'\mathbf{y} - (\mathbf{X}\beta)'\mathbf{y} - \mathbf{y}'\mathbf{X}\beta + \beta'\mathbf{X}'\mathbf{X}\beta \\ &= \mathbf{y}'\mathbf{y} - 2\mathbf{y}'\mathbf{X}\beta + \beta'\mathbf{X}'\mathbf{X}\beta \\ \Longrightarrow \frac{d}{d\beta}(\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta) &= -2\mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{X}\beta\end{aligned}$$

leading to

$$\mathbf{b} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y}$$

where $(\mathbf{X}'\mathbf{X})^{-1}$ exists (a.s.) because of our full rank assumption.

We will later follow Hayashi and express similar estimators in sample moment notation, here: $\mathbf{b} = \mathbf{S}_{\mathbf{xx}}^{-1}\mathbf{s}_{\mathbf{xy}}$, where $\mathbf{S}_{\mathbf{xx}} = \frac{1}{n}\sum_i \mathbf{x}_i\mathbf{x}_i'$ and $\mathbf{s}_{\mathbf{xy}} = \frac{1}{n}\sum_i \mathbf{x}_i y_i$. This notation is helpful for deriving large sample results.

Some quantities of interest are the *fitted value* $\widehat{y}_i = \mathbf{x}_i'\mathbf{b}$ and the *residual* $\widehat{\varepsilon}_i = y_i - \widehat{y}_i$.

We are now ready to state some important properties of $\mathbf{b}$.

Theorem

(i) $\mathbb{E}(\mathbf{b}|\mathbf{X}) = \beta$.

(ii) $Var(\mathbf{b}|\mathbf{X}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$

(iii) $\mathbf{b}$ is efficient among linear unbiased estimators, that is, $\mathbb{E}(\widetilde{\mathbf{b}}|\mathbf{X}) = \beta$ implies $Var(\widetilde{\mathbf{b}}|\mathbf{X}) \geq \sigma^2 (\mathbf{X}'\mathbf{X})^{-1}$ for any data-dependent $\widetilde{\mathbf{b}}$.

In particular, the same statements are implied but without conditioning on $\mathbf{X}$. This is the form in which you may know them and is more relevant for economics, but it is implied here. Notice, though, that stating the results unconditionally is sometimes a symptom of considering $\mathbf{X}$ fixed.

It is important to note that these are finite sample properties. In fact, we have not introduced asymptotic analysis yet. Furthermore, the theorem does not use a normality assumption – indeed, we have yet to make such an assumption.

Proof

See Hayashi's textbook for a proof based on this notation and assumptions.

We finally note that if we assume normality of errors, we can construct exact hypothesis tests. Specifically, assume:

Assumption 5: Normality

$$\varepsilon|\mathbf{X} \sim N(\mathbf{0}, \sigma^2\mathbf{I}_n).$$

Strictly speaking, the only new aspect of the assumption is that ε is normal. The assumption implies immediately that

$$(\mathbf{b} - \beta)|\mathbf{X} \sim N(\mathbf{0}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}),$$

which inspires our test statistics. Specifically, we state without proof the following:

Theorem: Finite Sample Hypothesis Tests

Let β_k denote the k^{th} component of β. Then if $H_0 : \beta_k = \overline{\beta}_k$ is true, one has

$$\text{t-ratio} = t \equiv \frac{b_k - \overline{\beta}_k}{\left[s^2\left((\mathbf{X}'\mathbf{X})^{-1}\right)_{kk}\right]^{1/2}} \sim t_{n-K},$$

the (Student) t-distribution. Here, $s^2 \equiv \frac{(\mathbf{y}-\mathbf{X}\mathbf{b})'(\mathbf{y}-\mathbf{X}\mathbf{b})}{n-K}$ is the usual standard error.

Let $H_0 : \mathbf{R}\beta = \mathbf{r}$ hold, where $\mathbf{R}$ has full rank $\#\mathbf{r}$, then

$$\text{F-statistic} = F \equiv \frac{(\mathbf{R}\mathbf{b} - \mathbf{r})'\left(\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'\right)^{-1}(\mathbf{R}\mathbf{b} - \mathbf{r})}{s^2 \cdot \#\mathbf{r}} \sim F_{\#\mathbf{r},n-K},$$

the F-distribution.

Both distributions can be looked up in books. We will not really use these exact distributions because we will only rarely impose normality. However, both test statistics will recur (with normal approximations to their distributions) under large sample analysis, and you should certainly be aware what the t- and F-distributions are about.

Large Sample Properties of OLS

We now change our focus and consider situations in which we cannot compute finite sample distributions. Specifically, we will drop Normality as well as i.i.d. assumptions (assumption 5 above was such an assumption). The price is that beyond unbiasedness and Gauss-Markov, we can only make statements about limits of sample distributions. We are, strictly speaking, not saying anything about finite sample performance. Of course, the idea is that approximations will work reasonably well in finite samples. In modern econometrics, it is standard to corroborate this by simulation exercises.

Introductory treatments typically assume that samples are i.i.d. We can generalize this assumption at very limited cost in terms of difficulty, because much weaker assumptions are sufficient to generate laws of large numbers and central limit theorems, the two main tools we need for asymptotic analysis. Our standard setting will be characterized by the following concepts:

Definition: $\{\mathbf{w}_i\}$ is *(strictly) stationary* if for any finite set of integers $\{j_1, j_2, \ldots, j_n\}$, the distribution of

$$\left(\mathbf{w}_{i+j_1}, \mathbf{w}_{i+j_2}, \ldots, \mathbf{w}_{i+j_n}\right)$$

does not depend on i.

Definition: Let $\mathcal{R}_{\mathbf{w}}$ denote the range of $\mathbf{w}_i$. A stationary random process $\{\mathbf{w}_i\}$ is *ergodic* if, for any two bounded functions $f : \mathcal{R}_{\mathbf{w}}^{k+1} \to \mathbb{R}$ and $g : \mathcal{R}_{\mathbf{w}}^{l+1} \to \mathbb{R}$ and any i,

$$\begin{aligned}\lim_{n\to\infty} &\left|\mathbb{E} f(w_i, \ldots, w_{i+k}) g(w_{i+n}, \ldots, w_{i+n+l})\right| \\ &= \left|\mathbb{E} f(w_i, \ldots, w_{i+k})\right| \left|\mathbb{E} g(w_i, \ldots, w_{i+l})\right|.\end{aligned}$$

Intuitively, this says that as two realizations move away from each other in the sequence, they approach independence.

Stationarity and ergodicity are weaker than i.i.d. In particular, neighboring observations need not be independent.

Example: AR(1)

Let the process $\{x_i\}$ be

$$x_i = \alpha + \rho x_{i-1} + \varepsilon_i,$$

where ε_i is i.i.d and $|\rho| < 1$. This process is stationary and ergodic but not i.i.d.

Why do we get away with "only" imposing these two things? Because they suffice to invoke a law of large numbers.

Ergodic Theorem

Let the process $\{\mathbf{w}_i\}$ be stationary and ergodic with $\mathbb{E}\mathbf{w}_i = \mu$. Then

$$\frac{1}{n}\sum_{i=1}^{n} \mathbf{w}_i \overset{a.s.}{\to} \mu.$$

Now we consider Assymptotic Normaility: For this we need a CLT

Definition: A scalar random process $\{x_i\}$ is a *martingale with respect to* $\{\mathbf{w}_i\}$ if

$$\mathbb{E}(x_i|\mathbf{w}_{i-1}, \ldots, \mathbf{w}_1) = x_{i-1}.$$

It is just called a martingale if it is a martingale with respect to itself.

Definition: A random process $\{\mathbf{w}_i\}$ with $\mathbb{E}(\mathbf{w}_i) = \mathbf{0}$ is a *martingale difference sequence* if

$$\mathbb{E}(\mathbf{w}_i|\mathbf{w}_{i-1}, \ldots, \mathbf{w}_1) = \mathbf{0}.$$

Every martingale difference sequence is the difference sequence of a martingale (hence the name).

All in all, the requirement that a process is a martingale difference sequence is again a weakening of i.i.d. requirements. Intuitively, it requires that there is no memory in the variable's first moment, but there might be memory in the higher moments. So for example, the process might be conditionally heteroskedastic but in a way that has memory.

Example: ARCH(1)

Let the process $\{x_i\}$ be

$$x_i = -\sqrt{\alpha + \rho x_{i-1}^2} \cdot \varepsilon_i,$$

where ε_i is i.i.d. standard normal (say). Under regularity conditions, this process is ergodic, stationary, and a martingale difference sequence, yet it is not i.i.d.

Why will we get away with "only" imposing ergodic stationarity and m.d.s.? Because they suffice to invoke a central limit theorem.

Ergodic Stationary Martingale Difference Sequence CLT:

Let $\{\mathbf{w}_i\}$ be stationary, ergodic, and a martingale difference sequence with finite second moments $\mathbb{E}(\mathbf{w}_i\mathbf{w}_i') \equiv \mathbf{\Sigma}$. Then

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \mathbf{w}_i \xrightarrow{d} \mathbf{N}(\mathbf{0}, \mathbf{\Sigma}).$$

We are now in a position to reconsider Ordinary Least Squares. We will mostly use sample moment notation; recall that in this notation, $\mathbf{b} = \mathbf{S}_{\mathbf{xx}}^{-1}\mathbf{s}_{\mathbf{xy}}$, where $\mathbf{S}_{\mathbf{xx}} = \frac{1}{n}\sum_i \mathbf{x}_i\mathbf{x}_i'$ and $\mathbf{s}_{\mathbf{xy}} = \frac{1}{n}\sum_i \mathbf{x}_i y_i$. Impose the following assumptions:

Assumption 1: Linearity

$$y_i = \mathbf{x}_i'\beta + \varepsilon_i.$$

Assumption 2: Ergodic Stationarity

The process $\{y_i, \mathbf{x}_i\}$ is jointly stationary and ergodic.

This implies that $\{\varepsilon_i\}$ is stationary and hence that the error term is unconditionally homoskedastic. However, assumption 2 is consistent with conditional heteroskedasticity, i.e. the possibility that $Var(\varepsilon_i|\mathbf{x}_i)$ varies with $\mathbf{x}_i$.

Assumption 3: Predetermined Regressors

$$\mathbb{E}(\mathbf{x}_i \varepsilon_i) = \mathbf{0}.$$

Notice that $\mathbf{0}$ is a vector here. Obviously this assumption is quite a bit weaker than the strict exogeneity assumption from the previous subsection; ε_i may now comove with past or future regressors. Also, we get away with merely restricting the expectation of $\mathbf{x}_i \varepsilon_i$ − and hence the correlation between the two − because the model is linear (look carefully through the consistency proof to see why). Estimation of nonlinear models will typically require stronger independence assumptions.

Also, we cast this in terms of a time series application (ie. past or future observations), but there may be some other natural ordering of observations coming from physical distance, position in some societal structure, or similar. The indexing is with respect to this natural ordering (if one exists).

Assumption 4: Rank Condition (Identification)

$$\mathbf{\Sigma}_{\mathbf{xx}} \equiv \mathbb{E}(\mathbf{x}_i \mathbf{x}_i') \text{ is nonsingular.}$$

Assumption 5: Error Process

$\{\mathbf{x}_i\varepsilon_i\}$ is a martingale difference sequence, and $\mathbf{S} \equiv \mathbb{E}\left(\mathbf{x}_i\varepsilon_i\left(\mathbf{x}_i\varepsilon_i\right)'\right)$ is nonsingular.

Assumption 5 implies assumption 3; we list them separately because assumption 5 is invoked only for parts of the analysis. In particular, assumption 5 excludes correlation of ε_i with *past* (in addition to contemporaneous) regressors.

Theorem

(i) Under assumptions 1-4, $\mathbf{b}$ is consistent:

$$\mathbf{b} \xrightarrow{p} \beta.$$

(ii) Under assumptions 1-5, $\mathbf{b}$ is asymptotically normal:

$$\sqrt{n}\left(\mathbf{b} - \beta\right) \xrightarrow{d} N\left(0, \boldsymbol{\Sigma}_{\mathbf{xx}}^{-1}\mathbf{S}\boldsymbol{\Sigma}_{\mathbf{xx}}^{-1}\right).$$

(iii) Let $\widehat{\mathbf{S}} \to \mathbf{S}$ and presume assumption 2, then

$$\mathbf{S}_{\mathbf{xx}}^{-1}\widehat{\mathbf{S}}\mathbf{S}_{\mathbf{xx}}^{-1} \xrightarrow{p} \boldsymbol{\Sigma}_{\mathbf{xx}}^{-1}\mathbf{S}\boldsymbol{\Sigma}_{\mathbf{xx}}^{-1}.$$

(iv) Let $\mathbb{E}\varepsilon_i^2 < \infty$, then under assumptions 1-4,

$$s^2 \equiv \frac{1}{n-K}\sum_i (y_i - \mathbf{x}_i'\mathbf{b})^2 \xrightarrow{p} \mathbb{E}\varepsilon_i^2.$$

The point of (iii) and (iv) is that we will need these estimators to construct hypothesis tests and confidence regions.

Proof

Here I give the proof for parts 1 and 2 only.

(i) The object of interest is really $(\mathbf{b} - \beta)$. We will first show that it vanishes under assumptions 1-4. Write

$$\begin{aligned}
\mathbf{b} - \beta &= \left(\frac{1}{n} \sum_i \mathbf{x}_i \mathbf{x}_i' \right)^{-1} \frac{1}{n} \sum_i \mathbf{x}_i y_i - \beta \\
&= \left(\frac{1}{n} \sum_i \mathbf{x}_i \mathbf{x}_i' \right)^{-1} \frac{1}{n} \sum_i \mathbf{x}_i (\mathbf{x}_i' \beta + \varepsilon_i) - \beta \\
&= \left(\frac{1}{n} \sum_i \mathbf{x}_i \mathbf{x}_i' \right)^{-1} \frac{1}{n} \sum_i \mathbf{x}_i \varepsilon_i .
\end{aligned}$$

But now notice that $\left(\frac{1}{n} \sum_i \mathbf{x}_i \mathbf{x}_i' \right)^{-1} \xrightarrow{p} \mathbf{\Sigma}_{\mathbf{xx}}^{-1}$ because $\frac{1}{n} \sum_i \mathbf{x}_i \mathbf{x}_i' \xrightarrow{p} \mathbf{\Sigma}_{\mathbf{xx}}$ by the Ergodic Theorem (using assumption 2), preservation of convergence in probability under continuous transformation, and nonsingularity of $\mathbf{\Sigma}_{\mathbf{xx}}$ (assumption 4). Again by the Ergodic Theorem, $\frac{1}{n} \sum_i \mathbf{x}_i \varepsilon_i \xrightarrow{p} \mathbb{E} \mathbf{x}_i \varepsilon_i$, which is zero by assumption. Again by standard results for convergence in probability, it follows that $\mathbf{b} - \beta \xrightarrow{p} \mathbf{\Sigma}_{\mathbf{xx}}^{-1} \cdot \mathbf{0} = \mathbf{0}$.

(ii) Write

$$\sqrt{n} \, (\mathbf{b} - \beta) = \left(\frac{1}{n} \sum_i \mathbf{x}_i \mathbf{x}_i' \right)^{-1} \frac{1}{\sqrt{n}} \sum_i \mathbf{x}_i \varepsilon_i .$$

Assumption 5 and the ergodic stationary martingale CLT imply that $\frac{1}{\sqrt{n}} \sum_i \mathbf{x}_i \varepsilon_i \xrightarrow{d} N(\mathbf{0}, \mathbf{S})$. Also, $\left(\frac{1}{n} \sum_i \mathbf{x}_i \mathbf{x}_i' \right)^{-1} \xrightarrow{p} \mathbf{\Sigma}_{\mathbf{xx}}^{-1}$ as before. The claim then follows by standard facts about weak convergence.

We now reconsider hypothesis testing. The following result is near immediate from what we just proved.

Theorem

(i) Let $H_0 : \beta^k = \overline{\beta}^k$ hold, then

$$t \equiv \frac{\sqrt{n}\left(b_k - \overline{\beta}_k\right)}{\sqrt{\left(\mathbf{S}_{\mathbf{xx}}^{-1}\widehat{\mathbf{S}}\mathbf{S}_{\mathbf{xx}}^{-1}\right)_{kk}}} \xrightarrow{d} N(0,1).$$

Proof

(i) Recall from the previous theorem that $\sqrt{n}\left(\mathbf{b} - \beta\right) \xrightarrow{d} N\left(0, \boldsymbol{\Sigma}_{\mathbf{xx}}^{-1}\mathbf{S}\boldsymbol{\Sigma}_{\mathbf{xx}}^{-1}\right)$ and $\mathbf{S}_{\mathbf{xx}}^{-1}\widehat{\mathbf{S}}\mathbf{S}_{\mathbf{xx}}^{-1} \xrightarrow{p} \boldsymbol{\Sigma}_{\mathbf{xx}}^{-1}\mathbf{S}\boldsymbol{\Sigma}_{\mathbf{xx}}^{-1}$. Restricting attention to the k^{th} component, it follows that $\sqrt{n}\left(\mathbf{b}_k - \beta_k\right) \xrightarrow{d} N\left(0, \left(\boldsymbol{\Sigma}_{\mathbf{xx}}^{-1}\mathbf{S}\boldsymbol{\Sigma}_{\mathbf{xx}}^{-1}\right)_{kk}\right)$ and $\left(\mathbf{S}_{\mathbf{xx}}^{-1}\widehat{\mathbf{S}}\mathbf{S}_{\mathbf{xx}}^{-1}\right)_{kk} \xrightarrow{p} \left(\boldsymbol{\Sigma}_{\mathbf{xx}}^{-1}\mathbf{S}\boldsymbol{\Sigma}_{\mathbf{xx}}^{-1}\right)_{kk}$. The claim then follows by standard facts.

Notice that the above test statistic is not quite a straightforward generalizations of the statistics we saw in the previous chapter. Specifically, we previously assumed conditional homoskedasticity, but we did not so here. As a result, variance-covariance matrices got more complicated, and the test statistics we derived are in fact the *robust* (or White) test statistics that you can find in the output of any OLS program. If we assume homoskedasticity, some expressions accordingly simplify:

Proposition: How things change with conditional homoskedasticity

Let $\mathbb{E}(\varepsilon_i^2|\mathbf{x}_i) = \sigma^2$, then:

(i) $\sqrt{n}\,(\mathbf{b} - \beta) \xrightarrow{d} N\left(0, \sigma^2 \mathbf{\Sigma}_{\mathbf{xx}}^{-1}\right)$.

(ii) $s^2 \mathbf{S}_{\mathbf{xx}}^{-1} \xrightarrow{p} \sigma^2 \mathbf{\Sigma}_{\mathbf{xx}}^{-1}$.

(iii) The hypothesis tests can be asymptotically conducted using the finite-sample t-ratio respectively $\#\mathbf{r}$ times the finite sample F-statistic.

Here, part (ii) follows from the Ergodic Theorem. We therefore see that under conditional homoskedasticity, large sample analysis of OLS leads to the same practical conclusions as finite sample analysis.

Part (ii) of the proposition shows that under homoskedasticity, estimation of $\mathbf{S}$ is not an issue. In contrast, we have not yet provided an estimator os $\mathbf{S}$ for the general case. Indeed, this requires an additional assumption.

Assumption 6: Finite Fourth Moments

$\mathbb{E}(x_{ik}x_{ij})^2$ exists and is finite for all k, j.

Proposition

Let assumption 6 hold, then $\widehat{\mathbf{S}} \equiv \frac{1}{n}\sum_i (y_i - \mathbf{x}_i'\mathbf{b})^2 \mathbf{x}_i\mathbf{x}_i'$ is consistent for $\mathbf{S}$.

Assumption 6 is needed because the variance of $\widehat{\mathbf{S}}$ comes from the fourth moment of $\mathbf{x}_i$, hence this moment must be finite.

Endogenous Variables and Identification

OLS critically relies on the assumption that regressors are in some sense exogenous, the weakest such sense being predeterminedness. There are many contexts in which this fails. For example, imagine a researcher who wants to figure out the returns to schooling by estimating a wage equation

$$\ln wage_i = \alpha + \beta \ln schooling_i + \varepsilon_i.$$

Is the assumption that $\mathbb{E}(schooling_i \varepsilon_i) = 0$ compelling? It would be if members of the population were assigned to schooling at random, but this is clearly not the case. More realistically, schooling selects for ability, i.e. people with higher ability also tend to have more schooling. As a result, we expect schooling and ability to exhibit positive correlation. Assuming that ability posively impacts wages, then with ability not "controlled for," the schooling variable will pick up some of its effect, and β will be overestimated. This problem is known as *endogeneity bias*. In the specific example, it can also be seen as *omitted variable bias*, caused by the omission of ability (which we cannot usually observe) from the equation.

Imagine now that we also observe a random variable x_i with the properties that $\mathbb{E}(x_i schooling_i) \neq 0$ but $\mathbb{E}(x_i \varepsilon_i) = 0$. Intuitively, x_i can be thought of as shifting $schooling_i$ without affecting ε_i and therefore inducing exogenous variation in $schooling_i$. x_i is called an *instrument*. In the specific example, it has been argued that the Vietnam draft number is an instrument: The number was allocated at random, and citizens with low draft numbers tended to enrol in college to avoid the draft.

Given x_i, we can consistently estimate β by

$$b_{IV} \equiv \left(\frac{1}{n} \sum_i x_i schooling_i \right)^{-1} \frac{1}{n} \sum_i x_i wage_i,$$

the instrumental variables estimator. An intuition for b_{IV} is that we estimate β using only the variation in schooling that is attributable to variation in x and, therefore, exogeneous.

We will not develop the theory of IV in any detail because it is a special case of the immediate next section. It is helpful, though, to remind ourselves of some other classic examples of endogeneity that may be amenable to IV analysis. Both are elaborated algebraically by Hayashi.

Probably the most classic example is the problem of simultaneously estimating supply and demand functions:

$$\begin{aligned} q_i &= q_i^d = \alpha_0 + \alpha_1 p_i + \varepsilon_i \\ q_i &= q_i^s = \beta_0 + \beta_1 p_i + \eta_i. \end{aligned}$$

This problem goes back to the 1920's. If ε_i has a positive realization, then demand is shifted upward, hence equilibrium price goes up. As a consequence, p_i and ε_i are positively correlated. This problem is also known as *simultaneity bias*. It is potentially solved by observable supply shifters, which could act as instruments.

The point is that you have to think about the structure of the error term every time you run a regression. Ask the following questions:

1. What is the data generating process? (aka, what kind of model do you have in mind for how these data are generated?)

2. What parts of it do I observe and what do I not observe?

3. The unobserved stuff is what econometricians call the error. What is the interpretation of the error in this context?

4. Given that I now know what my observed and unobserved stuff is, are they correlated?

5. If so, why? If not, why not?

6. if a problem exists, how do I solve it? (The best answer is always to make unobserved stuff observed, ie. get better data if possible. Econometric tricks to get around having crappy data are always less desirable than just having good data)

If you can't answer these question (at least the first 5) by the time you defend your proposal then I will likely want to fail you. That said, if you can't deal with these questions after this course, Allan and I will want to fail you well before that point.

As an aside, 50 per cent of the questions I ask in seminars are basically variants of this list.

Generalized Method of Moments

The Generalized Method of Moments (GMM) organizes many tools that you will have seen before, including anything preceding in this lecture, and many more that you will encounter elsewhere. Many econometricians think of it as major unifying principle of modern econometrics. Its name is due to a famous paper by Hansen (1982).

GMM starts by postulating *moment conditions* which are supposed to be true about the population. The reason for the word 'generalized' is that there may be more moment conditions than free parameters.

Reminder: Method of Moments

Let $\mathbf{w}_i$ be a vector of random variables of interest, and let θ denote the parameter we are after, then a moment condition looks like this:

$$\mathbb{E}(\mathbf{g}(\mathbf{w}_i, \theta)) = \mathbf{0}.$$

Here, $\mathbf{g}$ is some function of $\mathbf{w}$ and θ, and the fact that its expectation is zero reflects something that we know or assume about the model.

The methods of moments estimator is constructed by solving the sample analogs of the moment conditions, i.e. by setting

$$\frac{1}{n}\sum_{i=1}^{n} \mathbf{g}(\mathbf{w}_i, \widehat{\theta}) = \mathbf{0}.$$

(Thus, MM estimators are special cases of *analog estimators*.)

Example 1

Let

$$\mathbf{w}_i = (y_i, \mathbf{z}_i),$$

where y_i is a scalar that we want to predict/explain and $\mathbf{z}_i$ is a vector of regressors.

We believe that y_i is generated by a linear process:

$$y_i = \mathbf{z}_i'\theta + \varepsilon_i$$

and also that $\mathbf{z}_i$ is exogenous, i.e. the error term ε_i is uncorrelated with $\mathbf{z}_i$:

$$\mathbb{E}\mathbf{z}_i\varepsilon_i = \mathbf{0}.$$

This can be written as a moment condition. Set $\mathbf{g}(\mathbf{w}_i, \theta) = \mathbf{z}_i\left(y_i - \mathbf{z}_i'\theta\right)$, then the moment condition is that

$$\mathbb{E}\left(\mathbf{z}_i\left(y_i - \mathbf{z}_i'\theta\right)\right) = \mathbf{0}.$$

Of course, these are some of the assumption underlying OLS. If we solve the above equation's sample analog for θ, then we reconstruct OLS. So OLS is a method of moments estimator.

Example 2

Let

$$\mathbf{w}_i = (y_i, \mathbf{x}_i, \mathbf{z}_i),$$

where y_i is again a scalar and $\mathbf{x}_i$ and $\mathbf{z}_i$ are vectors of regressors, not necessarily disjoint. We still impose that

$$y_i = \mathbf{z}_i'\theta + \varepsilon_i,$$

but $\mathbf{z}_i$ might be endogenous, i.e. correlated with the errors. However, we are willing to believe that $\mathbf{x}_i$ is exogenous:

$$\mathbb{E}\mathbf{x}_i\varepsilon_i = \mathbf{0}.$$

This can again be written as a moment condition:

$$\mathbb{E}\left(\mathbf{x}_i\left(y_i - \mathbf{z}_i'\theta\right)\right) = \mathbf{0}.$$

If $\mathbf{x}_i$ correlates with $\mathbf{z}_i$, then the model is identified.

These are the assumptions behind linear IV estimation, e.g. estimation of linear supply/demand systems. If $\mathbf{z}_i$ and $\mathbf{x}_i$ have the same number of components, solving the sample analog for θ will reconstruct the IV estimator. But the GMM estimator is also defined if there are more instruments than regressors. Thus, we can in principle, use all instruments that we can think of (although see weak instruments in subsequent section).

Example 3

Let $\mathbf{w}_i$ as before but now consider a Poisson regression,

$$y_i = \exp(\mathbf{z}_i'\theta) + \varepsilon_i,$$

then one could write down a moment condition

$$\mathbb{E}\left(\mathbf{x}_i\left(y_i - \exp(\mathbf{z}_i'\theta)\right)\right) = \mathbf{0}.$$

GMM immediately handles this case as well. If $\mathbf{z}_i$ and $\mathbf{x}_i$ have the same number of components, it will among other things reconstruct *Nonlinear Least Squares*. We will not look closely at this case (at least for the moment) because we will restrict ourselves to linear moment conditions. But I mention it here to illustrate the flexibility of the GMM approach.

Linear GMM: An Overview

We want to estimate a linear equation.

$$y_i = \mathbf{z}_i'\theta + \varepsilon_i.$$

We assume that the random variables

$$\mathbf{w}_i = (y_i, \mathbf{x}_i, \mathbf{z}_i)$$

are jointly stationary and ergodic.

We have the moment conditions

$$\mathbb{E}\left(\mathbf{x}_i\left(y_i - \mathbf{z}_i'\theta\right)\right) = \mathbf{0}.$$

The aim is to estimate θ. The idea will be to do this by evaluating the sample analogs of the moment conditions.

Notice in particular that we will not assume exact identification, i.e. that $\mathbf{x}_i$ and $\mathbf{z}_i$ have the same number of components. Also, we will henceforth think of

$$\mathbf{x}_i = \mathbf{z}_i$$

as the special case in which all regressors are their own instruments.

Identification

Let

$$\begin{aligned} K &= \text{number of moment conditions} \\ L &= \text{number of parameters,} \end{aligned}$$

and also impose the following rank assumption:

$$\underset{K \times L}{\boldsymbol{\Sigma}_{\mathbf{xz}} \equiv \mathbb{E}\mathbf{x}_i \mathbf{z}_i'} \text{ has full column rank, i.e. rank } L.$$

Intuitively, this says that no instrument/moment condition is redundant.

Then the above model is

$$\begin{aligned} \text{underidentified if } K &< L, \\ \text{just identified if } K &= L, \\ \text{overidentified if } K &> L. \end{aligned}$$

"Underidentified" we already know. It plainly means there are more unknowns than equations, so even if we knew the distribution of $\mathbf{w}$, we could not solve for the parameters. Underidentification continues to be an insurmountable problem, and we will not further think about it at this point.

"Just identified" is what we have been dealing with so far. It will here emerge as special case.

The new aspect is that GMM is able to deal with "overidentified."

The Case of Overidentification

$K > L$ means that there are more linear equations than unknowns. Such a system of equations has generically no solution.

Of course, if our moment conditions are true, then the system is not generic in this sense. If we knew the true distribution of $\mathbf{w}_i$, the system would turn out to have a solution at the true parameter point. As a result, we could also test our conditions: If no solution exists, some of the moment conditions must be wrong.

The remaining problem is estimation. The sampling distribution of $\mathbf{w}$ will not be the population distribution, and thus, attempting to solve the moment conditions' sample analogs will generically lead to a contradiction.

So how can we estimate θ? While we will not have

$$\frac{1}{n}\sum_{i=1}^{n}\left(\mathbf{x}_i\left(y_i-\mathbf{z}_i'\theta\right)\right)=\mathbf{0}$$

even for large n, we would expect some law of large numbers to yield

$$\frac{1}{n}\sum_{i=1}^{n}\left(\mathbf{x}_i\left(y_i-\mathbf{z}_i'\theta\right)\right)\longrightarrow\mathbf{0}.$$

If θ is identified, we will furthermore find

$$\mathbb{E}\mathbf{x}_i\left(y_i-\mathbf{z}_i'\widetilde{\theta}\right)\neq\mathbf{0}$$

and consequently

$$\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\left(y_i-\mathbf{z}_i'\widetilde{\theta}\right)\rightarrow\mathbb{E}\left(\mathbf{x}_i\left(y_i-\mathbf{z}_i'\widetilde{\theta}\right)\right)\neq\mathbf{0}$$

for any $\widetilde{\theta}\neq\theta$.

Thus, a natural estimator for θ is

$$\widehat{\theta}\equiv\arg\min_{\theta\in\Theta}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\left(y_i-\mathbf{z}_i'\theta\right)\right)'\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\left(y_i-\mathbf{z}_i'\theta\right)\right).$$

Under our maintained assumptions, $\widehat{\theta}_n\rightarrow\theta$.

The core thing we overlooked is that by just squaring $\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\left(y_i-\mathbf{z}_i'\widehat{\theta}_n\right)$, we weighted all the moment conditions equally. But some of them might be more informative than others, for example by relating to random variables with a smaller sampling variation. We will therefore allow for a general weighting scheme. This immediately raises the question of optimal weighting, which we shall discuss.

Linear GMM: Formal Statement

Define the sample analog of $\mathbb{E}(\mathbf{g}(\theta, \mathbf{w}_i))$ by

$$\mathbf{g}_n(\theta) \equiv \frac{1}{n} \sum_{i=1}^{n} \mathbf{g}(\mathbf{w}_i, \theta).$$

Also fix any symmetric and positive definite weighting matrix $\mathbf{W}$ and let $\widehat{\mathbf{W}}$ be an estimator of $\mathbf{W}$, i.e. $\widehat{\mathbf{W}} \to_p \mathbf{W}$ as n grows large. (This allows for $\widehat{\mathbf{W}}$ to be data-dependent, but a constant pre-assigned $\widehat{\mathbf{W}} = \mathbf{W}$ is a possibility too.)

The GMM estimator $\widehat{\theta}(\widehat{\mathbf{W}})$ is

$$\begin{aligned} \widehat{\theta}(\widehat{\mathbf{W}}) &\equiv \arg\min_{\theta \in \Theta} J(\theta, \widehat{\mathbf{W}}), \\ J(\theta, \widehat{\mathbf{W}}) &\equiv n \cdot \mathbf{g}_n(\theta)' \widehat{\mathbf{W}} \mathbf{g}_n(\theta). \end{aligned}$$

Specialization to Linear Moment Conditions

With a linear model, we can solve for the GMM estimator in closed form.

Define the sample moments

$$\begin{aligned} \mathbf{s_{xy}} &\equiv \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i y_i, \\ \mathbf{S_{xz}} &\equiv \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i \mathbf{z}_i'. \end{aligned}$$

Then we can write

$$\begin{aligned} \mathbf{g}_n(\theta) &= \frac{1}{n}\sum_{i=1}^{n}\mathbf{g}(\mathbf{w}_i,\theta) \\ &= \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i(y_i - \mathbf{z}_i'\theta) \\ &= \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i y_i - \left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{z}_i'\right)\theta \\ &= \mathbf{s_{xy}} - \mathbf{S_{xz}}\theta \\ &= 0. \end{aligned}$$

The GMM objective function becomes

$$J(\theta, \widehat{\mathbf{W}}) = n \cdot (\mathbf{s_{xy}} - \mathbf{S_{xz}}\theta)'\widehat{\mathbf{W}}(\mathbf{s_{xy}} - \mathbf{S_{xz}}\theta).$$

The GMM objective function becomes

$$J(\theta, \widehat{\mathbf{W}}) = n \cdot (\mathbf{s_{xy}} - \mathbf{S_{xz}}\theta)' \widehat{\mathbf{W}} (\mathbf{s_{xy}} - \mathbf{S_{xz}}\theta).$$

Minimizing this with respect to θ leads to a first-order condition as follows:

$$\begin{aligned} 2(\mathbf{s_{xy}} - \mathbf{S_{xz}}\theta)' \widehat{\mathbf{W}} \mathbf{S_{xz}} &= 0 \\ \Longrightarrow \mathbf{s}_{\mathbf{xy}}' \widehat{\mathbf{W}} \mathbf{S_{xz}} &= (\mathbf{S_{xz}}\theta)' \widehat{\mathbf{W}} \mathbf{S_{xz}} \\ \Longrightarrow \mathbf{S}_{\mathbf{xz}}' \widehat{\mathbf{W}} \mathbf{s_{xy}} &= \mathbf{S}_{\mathbf{xz}}' \widehat{\mathbf{W}} \mathbf{S_{xz}} \theta. \end{aligned}$$

For this to have a unique solution, we need $\mathbf{S_{xz}}$ to be of full column rank. But this is given (eventually) since $\mathbf{S_{xz}} \stackrel{a.s.}{\to} \mathbf{\Sigma_{xz}}$ by the ergodic theorem, and $\mathbf{\Sigma_{xz}}$ has full column rank by assumption. Since also $\widehat{\mathbf{W}}$ is (eventually) positive definite by the same argument, $\mathbf{S}_{\mathbf{xz}}' \widehat{\mathbf{W}} \mathbf{S_{xz}}$ is invertible. Hence the GMM estimator is

$$\widehat{\theta}(\widehat{\mathbf{W}}) = \left(\mathbf{S}_{\mathbf{xz}}' \widehat{\mathbf{W}} \mathbf{S_{xz}} \right)^{-1} \mathbf{S}_{\mathbf{xz}}' \widehat{\mathbf{W}} \mathbf{s_{xy}}.$$

We will now look at this estimator's asymptotic properties.

Assumptions

Assumption 1: Linear Model

$$y_i = \mathbf{z}_i'\theta + \varepsilon_i.$$

Assumption 2: Ergodic Stationarity

$\mathbf{w}_i = (y_i, \mathbf{z}_i, \mathbf{x}_i)$ is jointly stationary and ergodic.

Assumption 3: Moment Conditions

$$\mathbb{E}\left(\mathbf{x}_i(y_i - \mathbf{z}_i'\theta)\right) = \mathbf{0}.$$

Assumption 4: Rank Condition

$\mathbf{\Sigma_{xz}} \equiv \mathbb{E}(\mathbf{x}_i\mathbf{z}_i')$ is of full column rank.

Assumption 5: Regularity Conditions on Errors

$\{\mathbf{x}_i\varepsilon_i\}$ is a martingale difference sequence.
$\mathbf{S} \equiv \mathbb{E}\left(\mathbf{x}_i\varepsilon_i\left(\mathbf{x}_i\varepsilon_i\right)'\right)$ is nonsingular.

ε_i being distributed i.i.d. and independent of $\mathbf{x}_i$ is sufficient for this to hold.

Theorem: Limiting Distribution of the GMM Estimator

(i)

$$\widehat{\theta}(\widehat{\mathbf{W}}) \xrightarrow{p} \theta.$$

(ii)

$$\sqrt{n}\left(\widehat{\theta}(\widehat{\mathbf{W}}) - \theta\right) \xrightarrow{d} N\left(\mathbf{0}, Avar\left(\widehat{\theta}(\widehat{\mathbf{W}})\right)\right)$$

$$Avar\left(\widehat{\theta}(\widehat{\mathbf{W}})\right) = \left(\boldsymbol{\Sigma}_{\mathbf{xz}}'\mathbf{W}\boldsymbol{\Sigma}_{\mathbf{xz}}\right)^{-1}\boldsymbol{\Sigma}_{\mathbf{xz}}'\mathbf{W}\mathbf{S}\mathbf{W}\boldsymbol{\Sigma}_{\mathbf{xz}}\left(\boldsymbol{\Sigma}_{\mathbf{xz}}'\mathbf{W}\boldsymbol{\Sigma}_{\mathbf{xz}}\right)^{-1}.$$

(iii) Let $\widehat{\mathbf{S}} \xrightarrow{p} \mathbf{S}$, then

$$\widehat{V} \equiv \left(\mathbf{S}_{\mathbf{xz}}'\widehat{\mathbf{W}}\mathbf{S}_{\mathbf{xz}}\right)^{-1}\mathbf{S}_{\mathbf{xz}}'\widehat{\mathbf{W}}\widehat{\mathbf{S}}\widehat{\mathbf{W}}\mathbf{S}_{\mathbf{xz}}\left(\mathbf{S}_{\mathbf{xz}}'\widehat{\mathbf{W}}\mathbf{S}_{\mathbf{xz}}\right)^{-1} \xrightarrow{p} Avar\left(\widehat{\theta}(\widehat{\mathbf{W}})\right).$$

Efficient GMM

If the model is just identified, the sample moment conditions can be solved, and this should yield the same estimator for any weighting matrix. Indeed, if $K = L$, then $\mathbf{S_{xz}}$ is square and hence (by the rank condition) invertible, and we can write

$$\begin{aligned}
\widehat{\theta}(\widehat{\mathbf{W}}) &= \left(\mathbf{S}_{\mathbf{xz}}'\widehat{\mathbf{W}}\mathbf{S}_{\mathbf{xz}}\right)^{-1}\mathbf{S}_{\mathbf{xz}}'\widehat{\mathbf{W}}\mathbf{s}_{\mathbf{xy}} \\
&= \mathbf{S}_{\mathbf{xz}}^{-1}\widehat{\mathbf{W}}^{-1}\mathbf{S}_{\mathbf{xz}}'^{-1}\mathbf{S}_{\mathbf{xz}}'\widehat{\mathbf{W}}\mathbf{s}_{\mathbf{xy}} \\
&= \mathbf{S}_{\mathbf{xz}}^{-1}\mathbf{s}_{\mathbf{xy}},
\end{aligned}$$

the usual IV estimator, which you may also know as $\widehat{\beta}_{IV} \equiv (\mathbf{X}'\mathbf{Z})^{-1}\mathbf{X}'\mathbf{y}$ or similar (keeping in mind that many texts use $\mathbf{X}$ and $\mathbf{Z}$ the other way round from these notes).

But in an overidentified model, $\widehat{\theta}(\widehat{\mathbf{W}})$ will nontrivially depend on $\widehat{\mathbf{W}}$. The obvious question is whether some $\widehat{\mathbf{W}}$ is optimal in a well-defined sense. As you can prove, any symmetric, positive definite matrix $\widehat{\mathbf{W}}$ would ensure consistency. But an intuition is to give more weight to moment conditions that are "less noisy" in the data generating process.

Recalling that

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathbf{x}_i\varepsilon_i \longrightarrow N(\mathbf{0},\mathbb{E}\left(\mathbf{x}_i\varepsilon_i\left(\mathbf{x}_i\varepsilon_i\right)'\right)) \equiv N(\mathbf{0},\mathbf{S}),$$

one might conjecture that an estimator of $\mathbf{S}^{-1}$ would make for a good weighting matrix.

This is indeed the case.

Proposition: Efficient GMM

The GMM estimator's asymptotic variance is bounded as follows:

$$Avar\left(\widehat{\theta}(\widehat{\mathbf{W}})\right) \geq \left(\mathbf{\Sigma}_{\mathbf{xz}}'\mathbf{S}^{-1}\mathbf{\Sigma}_{\mathbf{xz}}\right)^{-1},$$

and this bound is achieved whenever $\widehat{\mathbf{W}}$ is a consistent estimator of $\mathbf{S}^{-1}$.

This raises the question of how to estimate $\mathbf{S}$. As before, impose:

Assumption 6: Finite Fourth Moments

The matrix

$$\mathbb{E}\begin{bmatrix} (\mathbf{x}_{i1}\mathbf{z}_{i1})^2 & \cdots & (\mathbf{x}_{i1}\mathbf{z}_{iL})^2 \\ \vdots & \ddots & \vdots \\ (\mathbf{x}_{iK}\mathbf{z}_{i1})^2 & \cdots & (\mathbf{x}_{iK}\mathbf{z}_{iL})^2 \end{bmatrix}$$

exists and is finite.

Proposition: Estimator of $\mathbf{S}$

Let the above assumption hold and let $\widehat{\theta}$ be a consistent estimator of θ. Then

$$\begin{aligned} \widehat{\mathbf{S}} &\equiv \frac{1}{n}\sum_{i=1}^{n}\widehat{\varepsilon}_i^2\mathbf{x}_i\mathbf{x}_i' \\ \widehat{\varepsilon}_i &\equiv y_i - \mathbf{z}_i'\widehat{\theta} \end{aligned}$$

is a consistent estimator of $\mathbf{S}$.

Thus we find that if all of the above assumptions hold, then an efficient GMM estimator is

$$\widehat{\theta}\left(\widehat{\mathbf{S}}^{-1}\right) \equiv \left(\mathbf{S}_{\mathbf{xz}}'\widehat{\mathbf{S}}^{-1}\mathbf{S}_{\mathbf{xz}}\right)^{-1}\mathbf{S}_{\mathbf{xz}}'\widehat{\mathbf{S}}^{-1}\mathbf{s}_{\mathbf{xy}}$$

with asymptotic variance

$$Avar\left(\widehat{\theta}\left(\widehat{\mathbf{S}}^{-1}\right)\right) = \left(\mathbf{\Sigma}_{\mathbf{xz}}'\mathbf{S}^{-1}\mathbf{\Sigma}_{\mathbf{xz}}\right)^{-1}$$

that can be estimated by

$$\widehat{\mathbf{V}} \equiv \left(\mathbf{S}_{\mathbf{xz}}'\widehat{\mathbf{S}}^{-1}\mathbf{S}_{\mathbf{xz}}\right)^{-1}.$$

From the above, one example of an efficient GMM estimator is the *optimal/two-step GMM estimator* that can be constructed as follows:

$$\begin{aligned}
\widehat{\theta} &\equiv \widehat{\theta}(\mathbf{S}_{\mathbf{xx}}^{-1}) \\
\widehat{\varepsilon_i} &\equiv y_i - \mathbf{z}_i'\widehat{\theta} \\
\widehat{\mathbf{S}} &\equiv \frac{1}{n}\sum_{i=1}^{n} \widehat{\varepsilon}_i^2 \mathbf{x}_i\mathbf{x}_i' \\
\widehat{\theta}_{2SGMM} &\equiv \widehat{\theta}\left(\widehat{\mathbf{S}}^{-1}\right).
\end{aligned}$$

The first stage estimator effectively presumes homoskedasticity. Indeed, it has an interpretation of its own; as we will see below, it is the 2SLS estimator.

As might be expected, the efficient GMM estimator also makes for the asymptotically most powerful tests. However, this does not prove that the 2SIV (or another efficient) estimator should be used in small samples. In fact, one might be worried about finite sample performance because $\widehat{\mathbf{S}}$ uses sample information about second moments. This question has been examined in Monte-Carlo studies (e.g., Altonji/Segal, 1996). The upshot is that in small samples, one-step procedures may lead to more efficient point estimators. Also, tests may be well below their nominal levels in small samples.

I've seen examples where efficient GMM in finite samples greates horrific amounts of noise...

Recall that we tend to hope and pray that assymptotic properties approximate finite sample properties (for a 'big enough' sample size)

As an aside, much of the testing/inference theory requires efficient GMM. In particular Liklihood ratio tests require efficient GMM whereas Wald Stats do not. Given the above, I prefer Wald stats.

GMM with Conditional Homoskedasticity

Up to this point, we did not impose homoskedasticity. This is important because you will very often want to allow for heteroskedasticity. But GMM specializes in interesting ways when homoskedasticity is imposed.

Assumption 7: Conditional Homoskedasticity

$$\mathbb{E}(\varepsilon_i^2|\mathbf{x}_i) = \sigma^2.$$

It immediately follows that

$$\mathbf{S} = \sigma^2\mathbb{E}(\mathbf{x}_i\mathbf{x}_i') \equiv \sigma^2\mathbf{\Sigma}_{\mathbf{xx}}.$$

As a result of this, we do not need assumption 6 (finite fourth moments) any more. Let $\widehat{\sigma}^2$ be a consistent estimator of σ^2, then by the ergodic theorem and limit theorems for continuous functions,

$$\widehat{\mathbf{S}} \equiv \widehat{\sigma}^2\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i' \equiv \widehat{\sigma}^2\mathbf{S}_{\mathbf{xx}}$$

is a consistent estimator for $\mathbf{S}$.

The efficient GMM estimator now becomes

$$\begin{aligned}
\widehat{\theta}\left(\widehat{\mathbf{S}}^{-1}\right) &\equiv \left(\mathbf{S}_{\mathbf{xz}}'\left(\widehat{\sigma}^2\mathbf{S}_{\mathbf{xx}}\right)^{-1}\mathbf{S}_{\mathbf{xz}}\right)^{-1}\mathbf{S}_{\mathbf{xz}}'\left(\widehat{\sigma}^2\mathbf{S}_{\mathbf{xx}}\right)^{-1}\mathbf{s}_{\mathbf{xy}} \\
&= \left(\mathbf{S}_{\mathbf{xz}}'\mathbf{S}_{\mathbf{xx}}^{-1}\mathbf{S}_{\mathbf{xz}}\right)^{-1}\mathbf{S}_{\mathbf{xz}}'\mathbf{S}_{\mathbf{xx}}^{-1}\mathbf{s}_{\mathbf{xy}} \\
&= \widehat{\theta}\left(\mathbf{S}_{\mathbf{xx}}^{-1}\right) \\
&\equiv \widehat{\theta}_{2SLS}.
\end{aligned}$$

In short, the first step of the two-step estimation process that generated efficient GMM is redundant. We call this estimator $\widehat{\theta}_{2SLS}$ because it historically predates GMM under the name of "two-step least squares estimator." (We also encountered it before, namely as the first step in the two-step estimation.)

We notice that if $\mathbb{E}(\mathbf{z}_i\mathbf{z}_i')$ exists and is finite, then

$$\widehat{\sigma}^2 \equiv \frac{1}{n}\sum_{i=1}^{n}(y_i - \mathbf{z}_i'\widehat{\theta}_{2SLS})^2$$

is a consistent estimator.

We now find the following simplifications:

$$Avar\left(\widehat{\theta}_{2SLS}\right) = \sigma^2 \cdot \left(\mathbf{\Sigma}_{\mathbf{xz}}'\mathbf{\Sigma}_{\mathbf{xx}}^{-1}\mathbf{\Sigma}_{\mathbf{xz}}\right)^{-1},$$

which can be estimated by

$$\widehat{\mathbf{V}} = \widehat{\sigma}^2 \cdot \left(\mathbf{S}_{\mathbf{xz}}'\mathbf{S}_{\mathbf{xx}}^{-1}\mathbf{S}_{\mathbf{xz}}\right)^{-1}.$$

Remarks on Weak Instruments

Since GMM allows us to use as many instruments as we like, we could be tempted to use all instruments we can think of. And while a naive asymptotic analysis would indeed come to this conclusion, it is intutively clear that there must be a cost to instruments as well. The cost lies in the fact that they introduce noise.

Mathematically, imagine the extreme case of $\mathbf{z}_i$ and $\mathbf{x}_i$ being unrelated, then $\mathbf{\Sigma_{xz}}$ will not have full rank, and we will lose identification. On the other hand, whenever $\mathbf{\Sigma_{xz}}$ has full rank, our results go through. This looks like near-singular $\mathbf{\Sigma_{xz}}$ and singular $\mathbf{\Sigma_{xz}}$ are two different worlds, but reality is smoother than that. As $\mathbf{\Sigma_{xz}}$ approaches singularity, convergence will become extremely slow; for any fixed sample size, standard errors will diverge. Mathematically speaking, the problem is that IV estimation is consistent over the entire set of nonsingular $\mathbf{\Sigma_{xz}}$, but not *uniformly* so, as can be seen from inspecting the efficient GMM estimator's asymptotic variance: $\left(\mathbf{\Sigma}'_{\mathbf{xz}}\mathbf{S}^{-1}\mathbf{\Sigma_{xz}}\right)^{-1}$.

Analysis of weak instruments is the subject of ongoing research, but some signs of weak instruments are well known. Two things we should always look out for are: Do the instruments predict the regressors only badly, e.g. in terms of F-statistics and R^2? Do the IV estimates come with dramatically higher

standard errors than the OLS estimates? If the answer to either question is "yes," our instruments are probably weak. In papers using IV techniques, it is good practice to report answers to these questions. Also, appropriate techniques for estimation and inference with weak instruments were recently developed, but we will not get into them.

Summary

We studied the general idea of GMM and analyzed in detail the case of linear single-equation GMM.

In this rather specific form, GMM is equivalent to generalizing instrumental variables estimation in two directions: Firstly, by allowing for more instruments than regressors; secondly, by allowing for heteroskedasticity. The first of these generalizations had been considered before GMM was recognized as organizing principle, and hence, the specialization of GMM to homoskedastic errors reconstructed the two-step least squares method.

But it is also important to remember the generality: Many of the techniques predating GMM assumed homoskedasticity to keep things manageable. In GMM, there is no difficulty in allowing for heteroskedasticity, and similarly, we never used i.i.d. assumptions. Moreover, the restriction to linear moment conditions is really for simplicity.

A More Formal Take on Identification

As a prelude tothe rest of the course, we take a more formal look at identification. In linear models, identification is usually verified by checking some rank and order conditions. But not all models work like this. We will therefore take a brief look at what identification is substantively about.

The core intuition for identification can be phrased in the following ways:

"If a model is identified, the mapping from parameters of interest to distributions of observables is invertible."

"In an identified GMM model, if I know the population moment conditions, I know the parameters of interest."

"The parameter θ is identified if no two values $\theta \neq \theta_0$ are observationally equivalent."

These are related to a statement that one frequently reads:

"If a model is not identified, parameters cannot be estimated consistently."

This latter statement is correct, but its converse isn't.

Identification from Moment Conditions

Say a model is defined by orthogonality conditions of the form

$$\mathbb{E}(\mathbf{g}(\mathbf{w}_i, \theta_0)) = \mathbf{0}.$$

Here we changed notation a little, identifying θ_0 with the true value of θ. The symbol θ without decorations will henceforth refer to an arbitrary parameter value. (Hayashi changes the notation analogously at the beginning of chapter 7.)

This model is identified if knowledge of the population distribution $F_0(\mathbf{w}_i)$ would imply knowledge of θ. In general, this is equivalent to saying that the implicit function $F(\mathbf{w}_i) \mapsto \theta$ characterized by the above moment conditions exists and is single-valued at θ_0. More straightforwardly, the model is identified only if

$$\mathbb{E}(\mathbf{g}(\mathbf{w}_i, \theta)) = \mathbf{0} \Longrightarrow \theta = \theta_0,$$

i.e. the moment conditions have θ_0 as unique solution.

In nonlinear GMM, one usually has to impose just that. If $\mathbf{g}$ is linear, the condition is equivalent to more primitive restrictions. Consider the case of OLS, then

$$\mathbf{\Sigma_{ZZ}}\theta = \sigma_{\mathbf{zy}}$$

is solved by a unique θ_0 iff $\mathbf{\Sigma_{ZZ}}$ is nonsingular.

Identification when Likelihoods are Specified

Many models other than moment-based ones have the feature that the probability distribution of observables as function of parameters (the "likelihood function") $f(\mathbf{w}_i; \theta)$ is specified. For example, this is true in an OLS setup with normal errors, but also in setups that do not easily lend themselves to GMM treatment and also in Bayesian econometrics.

In this case, θ_0 is identified iff no other parameter value θ induces the same likelihood function. More technically, we have:

Definition: Likelihood Identification

Consider a model that specifies $f(\mathbf{w}_i; \theta)$. Then θ_0 is identified within $\mathbf{\Theta}$ iff for all $\theta \in \mathbf{\Theta}$ with $\theta \neq \theta_0$,

$[\mathbf{w}_i \in A \Rightarrow f(\mathbf{w}_i; \theta) \neq f(\mathbf{w}_i; \theta_0)]$, some event A with $\Pr(A) > 0$

or equivalently,

$$\Pr(f(\mathbf{w}_i; \theta) \neq f(\mathbf{w}_i; \theta_0)) \neq 0,$$

where the probability is with respect to sampling under the true parameter value.

We will now use a salient example to practise the most general proof technique for establishing identification (when you want to do it formally - which, be warned, is not the focus of this course).

Example: Nonparametric Regression

Consider the model

$$y_i = m(\mathbf{x}_i) + \varepsilon_i$$

with the regularity conditions that $\mathbf{x}_i$ is supported on $\mathbb{R}^k$, that m is continuous, and that ε is independent of $\mathbf{x}_i$.

This model is *not* identified. To see this, let F_ε denote the c.d.f. of ε_i and denote the true values of (m, F_ε) by (m_0, F_{ε_0}). Also define

$$\begin{aligned} \underset{\varepsilon}{\widetilde{F}}(e) &= F_{\varepsilon_0}(e-1) \\ \widetilde{m}(\mathbf{x}) &= m_0(\mathbf{x}) - 1. \end{aligned}$$

In plain English, $m(\mathbf{x})$ decreases by 1, but $\widetilde{\varepsilon}$ is distributed as $\varepsilon + 1$. Then it is intuitively clear that (m_0, F_{ε_0}) and $(\widetilde{m}_0, \widetilde{F}_{\varepsilon_0})$ are observationally equivalent. Formally, write

$$\begin{aligned} \Pr\left(y_i \leq y | \mathbf{x} = \mathbf{x}_i\right) &= \Pr\left(m_0(\mathbf{x}_i) + \varepsilon_i \leq y\right) \\ &= \Pr(\varepsilon_i \leq y - m_0(\mathbf{x}_i)) \\ &= F_{\varepsilon_0}(y - m_0(\mathbf{x}_i)) \\ &= F_{\varepsilon_0}(y - (m_0(\mathbf{x}_i) - 1) - 1) \\ &= \underset{\varepsilon}{\widetilde{F}}(y - \widetilde{m}(\mathbf{x})) \\ &= \Pr\left(\widetilde{m}(\mathbf{x}) + \widetilde{\varepsilon}_i \leq y\right), \end{aligned}$$

hence identification fails.

Assume now the additional condition that $\mathbb{E}(\varepsilon|\mathbf{x}_i) = 0$. Is the model then identified? Yes. This is easily seen by observing that

$$\mathbb{E}(y_i|\mathbf{x}_i) = \mathbb{E}(m_0(\mathbf{x}_i) + \varepsilon_i|\mathbf{x}_i) = m_0(\mathbf{x}_i) + \mathbb{E}(\varepsilon_i|\mathbf{x}_i) = m_0(\mathbf{x}_i),$$

hence m_0 is pinned down by the distribution of $(\mathbf{x}_i, y_i)$.

A more general proof technique for identification goes by assuming that $\theta \neq \theta_0$ and concluding that $\Pr(f(\mathbf{w}_i;\theta) \neq f(\mathbf{w}_i;\theta_0)) > 0$. In the present example, the argument goes as follows:

Let $\widetilde{m} \neq m_0$ and fix any random variable $\widetilde{\varepsilon}_i$ s.t. $\mathbb{E}(\widetilde{\varepsilon}_i|\mathbf{x}_i) = 0$. Then there exists $\mathbf{x}^*$ s.t. $\widetilde{m}(\mathbf{x}^*) \neq m_0(\mathbf{x}^*)$, say (w.l.o.g.) $\widetilde{m}(\mathbf{x}^*) > m_0(\mathbf{x}^*)$. Since attention is restricted to continuous functions, there exists a neighborhood $B(\mathbf{x}^*, \epsilon)$ s.t. $\widetilde{m}(\mathbf{x}) > m_0(\mathbf{y})$ for any $\mathbf{x}, \mathbf{y} \in B(\mathbf{x}^*, \epsilon)$. It follows that

$$\begin{aligned} \mathbb{E}(m_0(\mathbf{x}_i) + \varepsilon_i|\mathbf{x}_i \in B(\mathbf{x}^*, \epsilon)) &< \mathbb{E}(\widetilde{m}(\mathbf{x}_i) + \varepsilon_i|\mathbf{x}_i \in B(\mathbf{x}^*, \epsilon)) \\ &= \mathbb{E}(\widetilde{m}(\mathbf{x}_i) + \widetilde{\varepsilon}_i|\mathbf{x}_i \in B(\mathbf{x}^*, \epsilon)), \end{aligned}$$

where the last equality uses $\mathbb{E}(\varepsilon_i|\mathbf{x}_i) = \mathbb{E}(\widetilde{\varepsilon}_i|\mathbf{x}_i) = 0$. (Notice that independence of ε_i and $\mathbf{x}_i$ was not used; it is not needed to get identification.) But since $\mathbf{x}_i$ is supposed to have full support, $B(\mathbf{x}^*, \epsilon)$ is an event of nonzero probability.

The Most General Identification Condition

The moment-based identification condition can be seen as a case of likelihood identification: θ is not mapped onto a single distribution $f(\mathbf{w}_i;\theta)$, but onto a set of distributions that is characterized by the moment conditions:

$$\theta \mapsto \Gamma(\theta) \equiv \{f(\mathbf{w}_i) : \mathbb{E}\mathbf{g}(\mathbf{w}_i,\theta) = \mathbf{0}\} = \{f(\mathbf{w}_i) : \mathbf{\Sigma_{ZZ}}\theta = \sigma_{\mathbf{zy}}\}$$

(the last expression is the specialization to linear GMM). To accommodate this case, the definition of likelihood identification can be generalized to the requirement that

$$\theta \neq \theta_0 \Longrightarrow \Gamma(\theta) \cap \Gamma(\theta_0) =,$$

where $\Gamma(\theta)$ is the set of distributions $f(\mathbf{w}_i)$ that are consistent with parameter value θ. The condition says that the mapping Γ may be set-valued but that its inverse must be a function. In the example, identification then obtains iff $\mathbf{\Sigma_{ZZ}}$ is nonsingular, just as before.

This generalization of likelihood identification nests identification from moment conditions. Indeed, it is the most general identification criterion in that *any* identification condition can be seen as appropriate specialization of it.