Lecture Notes for Econometrics II

Jörg Stoye

March 5, 2025

Abstract

Please do not distribute further. I borrow extensively from Hayashi's textbook and to a lesser degree, from Amemiya's and Bruce Hansen's textbooks and the Newey/McFadden and Matzkin handbook chapters.

1 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 $X_i = (X_{i1}, \dots X_{ik})'$. A model specifies the relation between these random variables up to some unknown components. These components specifically include parameters of interest (but they might also include so-called nuisance parameters). We will attempt to estimate the parameters of interest by means of a size n sample of realizations (Y_i, X_i).

In this first section only, we will conduct finite sample analysis. For this, it can be helpful to think in terms of data matrices

$$\boldsymbol{Y} = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix}, \boldsymbol{X} = \begin{bmatrix} X_1' \\ \vdots \\ X_n' \end{bmatrix} = \begin{bmatrix} X_{11} & \cdots & X_{1k} \\ \vdots & \ddots & \vdots \\ X_{n1} & \cdots & X_{nk} \end{bmatrix}, \boldsymbol{e} = \begin{bmatrix} e_1 \\ \vdots \\ e_n \end{bmatrix}$$

that stack sample realizations.

Assume that the true relationship between X and Y is linear:

Assumption 1.1 ***Linearity***

$$Y_i = X_i'\beta_0 + e_i$$

or equivalently,

$$\boldsymbol{Y} = \boldsymbol{X}\beta_0 + \boldsymbol{e}.$$

Here and henceforth, the subscript β_0 denotes a true parameter value.

Assumption **??** is restrictive only in conjunction with additional assumption on e_i; else it could be read as definition of e_i. Note that a model does not have to be linear in parameters of economic interest to be a linear statistical model. Well known examples include Cobb-Douglas production functions and certain wage equations that become linear upon taking logs of some quantities.

To begin, we furthermore assume that X_i is strictly exogenous:

Assumption 1.2 ***Strict Exogeneity***

$$\mathbb{E}(e_i|\boldsymbol{X}) = 0$$

or equivalently

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

Notice that in the presence of a constant regressor, setting the expectation equal to zero is a normalization since any other value could be absorbed into the coefficient on the constant. The restrictive part of the assumption is that $\mathbb{E}(e_i|\boldsymbol{X})$ may not depend on either i or $\boldsymbol{X}$. In fact, the second of these restrictions is very strong and will be relaxed later. For one example, it cannot be fulfilled when the regressors include lagged dependent variables.

Assumption 1.3 ***Rank Condition***

$$\text{rank}(\boldsymbol{X}) = k \ \textit{a.s.}$$

If $\text{rank}(\boldsymbol{X}) < k$, one regressor is linearly dependent on the others. It is intuitively clear that in this case, we cannot disentangle the regressors' individual linear effects on Y. Indeed, we will later think of this assumption as an *identification condition*.

The assumption as written is a bit sloppy because it fails in highly relevant cases. (Namely...?) However, because all other assumptions are stated conditionally on $\boldsymbol{X}$, we could condition our analysis on the event that the assumption holds in our sample.

Assumption 1.4 ***Spherical Error***

$$\begin{aligned} \mathbb{E}(e_i^2|\boldsymbol{X}) &= \sigma^2 > 0 \\ \mathbb{E}(e_i e_j|\boldsymbol{X}) &= 0 \end{aligned}$$

or equivalently,

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

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

The Ordinary Least Squares (OLS) estimator $\hat{\beta}$ of β_0 is derived as

$$\begin{aligned}\hat{\beta} &\equiv \arg\min_{\beta} \sum_i (Y_i - X_i'\beta)^2 \\ &= \arg\min_{\beta} (\boldsymbol{Y} - \boldsymbol{X}\beta)' (\boldsymbol{Y} - \boldsymbol{X}\beta),\end{aligned}$$

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

We can solve this in closed form:

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

leading to

$$\hat{\beta} = (\boldsymbol{X}'\boldsymbol{X})^{-1} \boldsymbol{X}'\boldsymbol{Y}$$

where $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ exists (a.s.) because of our full rank assumption. We will later follow Hansan and express similar estimators in sample moment notation, here: $\hat{\beta} = \hat{\boldsymbol{Q}}_{XX}^{-1}\hat{\boldsymbol{Q}}_{XY}$, where $\hat{\boldsymbol{Q}}_{XX} = \frac{1}{n}\sum_i X_i X_i'$ and $\hat{\boldsymbol{Q}}_{XY} = \frac{1}{n}\sum_i X_i Y_i$. This notation is helpful for deriving large sample results.

Some quantities of interest are the *fitted value* $\hat{Y}_i = X_i'\hat{\beta}$ and the *residual* $\hat{e}_i = Y_i - \hat{Y}_i$. It is instructive to consider these in data matrix notation: $\hat{\boldsymbol{Y}} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}$ and $\hat{\boldsymbol{e}} = \boldsymbol{Y} - \hat{\boldsymbol{Y}} = (\boldsymbol{I}_n - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}')\boldsymbol{Y}$. $\boldsymbol{P} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ is the projector that projects any vector into the column space of $\boldsymbol{X}$, whereas $\boldsymbol{I}_n - \boldsymbol{P}$ is the associated annihilator that generates the orthogonal complement. We therefore see that the vector of fitted values $\hat{\boldsymbol{Y}}$ is the projection of $\boldsymbol{Y}$ onto the column space of $\boldsymbol{X}$. It immediately follows that $\hat{\boldsymbol{e}}$ is orthogonal to both that column space and to $\hat{\boldsymbol{Y}}$. This is the well-known geometry of least squares.

We are now ready to state some important properties of $\hat{\beta}$.

Theorem 1.1 ***Finite Sample Behavior of OLS***

Under Assumptions **??**, **??**, *and* **??**, *for any* $\boldsymbol{X}$ *with* $\text{rank}(\boldsymbol{X}) = k$ *we have:*

1. $\mathbb{E}(\hat{\beta}|\boldsymbol{X}) = \beta_0$.
2. $\text{var}(\hat{\beta}|\boldsymbol{X}) = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$
3. $\hat{\beta}$ *is efficient among linear unbiased estimators, that is,* $\mathbb{E}(\tilde{\beta}|\boldsymbol{X}) = \beta_0$ *implies* $\text{var}(\tilde{\beta}|\boldsymbol{X}) \geq \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$ *for any* $\tilde{\beta}$ *that is linear in* $\boldsymbol{Y}$.

Furthermore, also using Assumption **??**, *analogous unconditional statements hold (homework).*

We reiterate that, in contrast to asymptotic approximations that you may have seen before (and will see lots of in this lecture), these are finite sample properties. Furthermore, the theorem does not use a normality assumption, and it will be clear from the proof that not all assumptions are needed for all parts of the Theorem.

Proof.

1.

$$\begin{aligned}
\mathbb{E}(\hat{\beta}|\boldsymbol{X}) &= \mathbb{E}\big((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}|\boldsymbol{X}\big) \\
&= (\boldsymbol{X}'\boldsymbol{X})^{-1}\,\mathbb{E}\,(\boldsymbol{X}'(\boldsymbol{X}\beta_0+\boldsymbol{e})|\boldsymbol{X}) \\
&= \beta_0+(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\mathbb{E}\,(\boldsymbol{e}|\boldsymbol{X})=\beta_0,
\end{aligned}$$

where we used strict exogeneity and the conditioning on $\boldsymbol{X}$.

2.

$$\begin{aligned}
\mathrm{var}(\hat{\beta}|\boldsymbol{X}) &= \mathrm{var}\big((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}|\boldsymbol{X}\big) \\
&= \mathrm{var}\big((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'(\boldsymbol{X}\beta_0+\boldsymbol{e})|\boldsymbol{X}\big) \\
&= \mathrm{var}\big(\beta_0+(\boldsymbol{X}'\boldsymbol{X})^{-1}\,\boldsymbol{X}'\boldsymbol{e}|\boldsymbol{X}\big) \\
&= (\boldsymbol{X}'\boldsymbol{X})^{-1}\,\boldsymbol{X}'\sigma^2\boldsymbol{X}\,(\boldsymbol{X}'\boldsymbol{X})^{-1} \\
&= (\boldsymbol{X}'\boldsymbol{X})^{-1}\,\sigma^2
\end{aligned}$$

where we used strict exogeneity, Assumption **??**, the conditioning on $\boldsymbol{X}$, and the fact that β_0 is not a random variable.

3. By hypothesis, $\tilde{\beta}=\boldsymbol{C}\boldsymbol{Y}$ and hence trivially $\tilde{\beta}=((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'+\boldsymbol{D})\boldsymbol{Y}$, where $\boldsymbol{C}$ and $\boldsymbol{D}$ may depend on $\boldsymbol{X}$ but not on $\boldsymbol{Y}$. Unbiasedness of $\tilde{\beta}$ then implies

$$\begin{aligned}
\beta_0 &= \mathbb{E}\left(((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'+\boldsymbol{D})\boldsymbol{Y}|\boldsymbol{X}\right) \\
&= \beta_0+\mathbb{E}\,(\boldsymbol{D}\boldsymbol{Y}|\boldsymbol{X}) \\
&= \beta_0+\mathbb{E}\,(\boldsymbol{D}(\boldsymbol{X}\beta_0+\boldsymbol{e})|\boldsymbol{X}) \\
&= \beta_0+\mathbb{E}(\boldsymbol{D}\boldsymbol{X}\beta_0|\boldsymbol{X})+\boldsymbol{D}\underbrace{\mathbb{E}(\boldsymbol{e}|\boldsymbol{X})}_{=\mathbf{0}},
\end{aligned}$$

hence $\mathbb{E}(\boldsymbol{D}\boldsymbol{X}\beta_0|\boldsymbol{X})=\mathbf{0}$. Since $\boldsymbol{D}$ can depend only on $\boldsymbol{X}$ and in particular not on β_0, it must be the case that $\boldsymbol{D}\boldsymbol{X}=\mathbf{0}$.

Now write

$$\begin{aligned}
\text{var}\left(((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'+\boldsymbol{D})\boldsymbol{Y}|\boldsymbol{X}\right) &= \text{var}\left(((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'+\boldsymbol{D})(\boldsymbol{X}\boldsymbol{\beta}_0+\boldsymbol{e})|\boldsymbol{X}\right) \\
&= \text{var}\left(((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'+\boldsymbol{D})\boldsymbol{e}|\boldsymbol{X}\right) \\
&= \sigma^2((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'+\boldsymbol{D})((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'+\boldsymbol{D})' \\
&= \sigma^2((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}+\boldsymbol{D}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}+(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{D}'+\boldsymbol{D}\boldsymbol{D}') \\
&\geq \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1},
\end{aligned}$$

where the last step used cancellation, the fact that $\boldsymbol{D}\boldsymbol{X}=\boldsymbol{0}$, and the fact that $\boldsymbol{D}\boldsymbol{D}'$ is positive semidefinite. ■

To check comprehension of the Gauss-Markov theorem and its proof, you should be able to answer the following questions: Is unbiasedness of $\tilde{\beta}$ needed in the hypothesis? If so, where is it used? Why can we not write $\hat{\beta}=\tilde{\beta}+\boldsymbol{D}\boldsymbol{Y}$ and then go through the proof of (iii) to get the exact opposite conclusion?

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

Assumption 1.5 ***Normality***

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

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

$$(\hat{\beta}-\beta_0)|\boldsymbol{X}\sim N(\boldsymbol{0},\sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}),$$

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

Theorem 1.2 ***Finite Sample Hypothesis Tests***

Let β_j denote the j^{th} component of vector β. Then if $H_0:{\beta_0}^{[j]}=\beta_j$ is true, one has

$$t\text{-ratio}=t\equiv\frac{\hat{\beta}_j-\beta_j}{\left(s^2\left[(\boldsymbol{X}'\boldsymbol{X})^{-1}\right]_{jj}\right)^{1/2}}\sim t_{n-k},$$

the (Student) t-distribution. Here, $s^2\equiv\frac{(\boldsymbol{Y}-\boldsymbol{X}\hat{\beta})'(\boldsymbol{Y}-\boldsymbol{X}\hat{\beta})}{n-k}$ is the usual standard error.

Let $H_0:\boldsymbol{R}\beta_0=\boldsymbol{r}$ hold, where $\boldsymbol{R}$ has full rank $\#\boldsymbol{r}$, then

$$F\text{-statistic}=F\equiv\frac{\left(\boldsymbol{R}\hat{\beta}-\boldsymbol{r}\right)'\left(\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}'\right)^{-1}\left(\boldsymbol{R}\hat{\beta}-\boldsymbol{r}\right)}{s^2\ \#\boldsymbol{r}}\sim F_{\#\boldsymbol{r},n-k},$$

the F-distribution.

Both distributions have been tabulated. We will not really use these exact distributions because we will only rarely impose normality. However, both test statistics will recur (with asymptotic approximations to their distributions) under large sample analysis.

2 Large Sample Properties of OLS

We now change our focus and consider situations in which we cannot compute finite sample distributions. Specifically, we drop Normality. We maintain an i.i.d. assumption in order to emphasize conceptual issues, though for most results in this lecture, that assumption is considerably stronger than required. The price of these relaxations is that beyond unbiasedness and Gauss-Markov, we can only make statements about limits of sample distributions. Since these statements do not, in general, hold uniformly, 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. If you develop new estimators, you will probably corroborate this by simulation exercises. (There is also work on uniform asymptotics, but we will not get into that.)

An i.i.d. assumption, together with very modest restrictions on the distribution of (Y_i, X_i), allows us to characterize asymptotic distributions of estimators, test statistics, and the like because of the following results.

Theorem 2.1 ***Strong Law of Large Numbers (Kolmogorov)***

Let the process $\{\boldsymbol{w}_i\}$ be i.i.d. with finite expectation $\mathbb{E}\boldsymbol{w}_i = \boldsymbol{\mu}$. Then

$$\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{w}_i \xrightarrow{a.s.} \boldsymbol{\mu}.$$

Theorem 2.2 ***Multivariate Central Limit Theorem (close to Lindeberg-Lévy)***

Let $\{\boldsymbol{w}_i\}$ be i.i.d. with finite expectation $\mathbb{E}\boldsymbol{w}_i = \boldsymbol{\mu}$ and variance $\mathbb{E}(\boldsymbol{w}_i\boldsymbol{w}_i') - \boldsymbol{\mu}\boldsymbol{\mu}' \equiv \boldsymbol{\Sigma}$. Then

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{w}_i - \boldsymbol{\mu}\right) \xrightarrow{d} \boldsymbol{N}(\boldsymbol{0}, \boldsymbol{\Sigma}).$$

Thus, sample means converge to expectations. Deviations of sample means from the corresponding expectation converge to normal after appropriate rescaling, which in this lecture will always be $\sqrt{n}$. These conclusions actually hold under considerably weaker conditions than stated here.

Let's now reconsider Ordinary Least Squares. We will mostly use sample moment notation; recall that in this notation, $\hat{\beta} = \hat{\boldsymbol{Q}}_{XX}^{-1}\hat{\boldsymbol{Q}}_{XY}$, where $\hat{\boldsymbol{Q}}_{XX} = \frac{1}{n}\sum_i X_i X_i'$ and $\hat{\boldsymbol{Q}}_{XY} = \frac{1}{n}\sum_i X_i Y_i$. Impose the following assumptions:

Assumption 2.1 ***Linearity***

$$Y_i = X_i'\beta_0 + e_i.$$

Assumption 2.2 ***IID***

The process $\{Y_i, X_i\}$ is i.i.d.

This implies that $\{e_i\}$ is also i.i.d., hence that the error term is unconditionally homoskedastic. However, this is a statement about the marginal distribution of e_i. Assumption **??** is consistent with conditional heteroskedasticity, i.e. the possibility that $\text{var}(e_i|X_i)$ varies with X_i.

Assumption 2.3 ***Predetermined Regressors***

$$\mathbb{E}(X_i e_i) = \mathbf{0}.$$

We get away with merely restricting the expectation of $X_i e_i$ – and hence the correlation between the two – because the model is linear. Estimation of nonlinear models will typically require stronger mean independence assumptions. Given our i.i.d. assumption, e_i is also assumed to be independent, hence uncorrelated, with past or future regressors. In contexts where this is a concern, it is possible to weaken it.

Assumption 2.4 ***Rank Condition (Identification)***

$$\boldsymbol{Q}_{XX} \equiv \mathbb{E}(X_i X_i') \textit{ is nonsingular.}$$

Theorem 2.3 ***Asymptotic Behavior of OLS***

Under assumptions 1-4:

1. *The estimator $\hat{\beta}$ is consistent:*
$$\hat{\beta} \xrightarrow{p} \beta_0.$$

2. *The estimator $\hat{\beta}$ is asymptotically normal:*
$$\sqrt{n}\left(\hat{\beta} - \beta_0\right) \xrightarrow{d} N\left(0, \boldsymbol{Q}_{XX}^{-1}\Omega\boldsymbol{Q}_{XX}^{-1}\right).$$

3. *Let $\hat{\Omega} \xrightarrow{p} \Omega$, then*
$$\hat{\boldsymbol{Q}}_{XX}^{-1}\hat{\Omega}\hat{\boldsymbol{Q}}_{XX}^{-1} \xrightarrow{p} \boldsymbol{Q}_{XX}^{-1}\Omega\boldsymbol{Q}_{XX}^{-1}.$$

4. *Let $\mathbb{E}e_i^2 < \infty$, then*
$$s^2 \equiv \frac{1}{n-k}\sum_i (Y_i - X_i'\hat{\beta})^2 \xrightarrow{p} \mathbb{E}e_i^2.$$

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

Proof. Here and henceforth, we freely use Slutsky's Theorem and also the Continuous Mapping Theorem, including its corollary that $X_n \xrightarrow{p} x, Y_n \xrightarrow{p} y \Rightarrow X_n Y_n \xrightarrow{p} xy$.

1. The object of interest really is $(\hat{\beta} - \beta_0)$. We first show that it vanishes under Assumptions **??**-**??**. Thus, write

$$\begin{aligned}
\hat{\beta} - \beta_0 &= \left(\frac{1}{n} \sum_i X_i X_i' \right)^{-1} \frac{1}{n} \sum_i X_i Y_i - \beta_0 \\
&= \left(\frac{1}{n} \sum_i X_i X_i' \right)^{-1} \frac{1}{n} \sum_i X_i (X_i' \beta_0 + e_i) - \beta_0 \\
&= \left(\frac{1}{n} \sum_i X_i X_i' \right)^{-1} \frac{1}{n} \sum_i X_i e_i.
\end{aligned}$$

   But now notice that $\left(\frac{1}{n} \sum_i X_i X_i' \right)^{-1} \xrightarrow{p} \boldsymbol{Q}_{XX}^{-1}$ because $\frac{1}{n} \sum_i X_i X_i' \xrightarrow{p} \boldsymbol{Q}_{XX}$ by the SLLN (using Assumption **??**), the Continuous Mapping Theorem, and nonsingularity of $\boldsymbol{Q}_{XX}$ (Assumption **??**). Also by the SLLN, $\frac{1}{n} \sum_i X_i e_i \xrightarrow{p} \mathbb{E} X_i e_i = \mathbf{0}$.

2. Write

$$\sqrt{n}(\hat{\beta} - \beta_0) = \left(\frac{1}{n} \sum_i X_i X_i' \right)^{-1} \frac{1}{\sqrt{n}} \sum_i X_i e_i.$$

   The CLT now yields $\frac{1}{\sqrt{n}} \sum_i X_i e_i \xrightarrow{d} N(\mathbf{0}, \Omega)$. Also, $\left(\frac{1}{n} \sum_i X_i X_i' \right)^{-1} \xrightarrow{p} \boldsymbol{Q}_{XX}^{-1}$ as before.

3. Trivial.

4. We show the equivalent statement that $\frac{1}{n} \sum_i (Y_i - X_i' \hat{\beta})^2 \xrightarrow{p} \mathbb{E} e_i^2$. Indeed, the adjustment of the denominator is irrelevant from a consistency point of view and has other justifications (namely, finite sample unbiasedness of the resulting estimator). Write

$$\begin{aligned}
\frac{1}{n} \sum_i (Y_i - X_i' \hat{\beta})^2 &= \frac{1}{n} \sum_i (Y_i - X_i' \beta_0 - X_i' (\hat{\beta} - \beta_0))^2 \\
&= \frac{1}{n} \sum_i (e_i - X_i' (\hat{\beta} - \beta_0))^2 \\
&= \frac{1}{n} \sum_i (e_i^2 - 2(\hat{\beta} - \beta_0)' X_i e_i + (\hat{\beta} - \beta_0)' X_i X_i' (\hat{\beta} - \beta_0)) \\
&= \frac{1}{n} \sum_i e_i^2 - \frac{2}{n} (\hat{\beta} - \beta_0)' \sum_i X_i e_i + (\hat{\beta} - \beta_0)' \hat{\boldsymbol{Q}}_{XX} (\hat{\beta} - \beta_0),
\end{aligned}$$

   but we already know that $(\hat{\beta} - \beta_0) \xrightarrow{p} \mathbf{0}$, that $\frac{1}{n} \sum_i X_i e_i \xrightarrow{p} \mathbf{0}$, and that $\hat{\boldsymbol{Q}}_{XX} \xrightarrow{p} \boldsymbol{Q}_{XX}$.

■

The following result will be useful for hypothesis testing.

Assumption 2.5 ***Error Process***

$\Omega \equiv \mathbb{E}\left(X_i e_i \left(X_i e_i\right)'\right)$ *is nonsingular.*

Theorem 2.4 ***Hypothesis Testing***

Let Assumptions **??**-**??** *hold. Then:*

1. *Let* $H_0 : \beta_0{}^{[j]} = \beta_j$ *hold, then*

$$t \equiv \frac{\sqrt{n}\left(\hat{\beta}_j - \beta_j\right)}{\sqrt{\left(\hat{\boldsymbol{Q}}_{XX}^{-1}\hat{\Omega}\hat{\boldsymbol{Q}}_{XX}^{-1}\right)_{jj}}} \xrightarrow{d} N(0,1).$$

2. *Let* $H_0 : \boldsymbol{R}\beta_0 = \boldsymbol{r}$ *hold, where* $\boldsymbol{R}$ *has rank* $\#\boldsymbol{r}$*, then*

$$W \equiv n\left(\boldsymbol{R}\hat{\beta} - \boldsymbol{r}\right)'\left(\boldsymbol{R}\hat{\boldsymbol{Q}}_{XX}^{-1}\hat{\Omega}\hat{\boldsymbol{Q}}_{XX}^{-1}\boldsymbol{R}'\right)^{-1}\left(\boldsymbol{R}\hat{\beta} - \boldsymbol{r}\right) \xrightarrow{d} \chi^2(\#\boldsymbol{r}).$$

Proof.

1. Recall from Theorem **??** that $\sqrt{n}(\hat{\beta}-\beta_0) \xrightarrow{d} N(0, \boldsymbol{Q}_{XX}^{-1}\Omega\boldsymbol{Q}_{XX}^{-1})$ and $\hat{\boldsymbol{Q}}_{XX}^{-1}\hat{\Omega}\hat{\boldsymbol{Q}}_{XX}^{-1} \xrightarrow{p} \boldsymbol{Q}_{XX}^{-1}\Omega\boldsymbol{Q}_{XX}^{-1}$. Restricting attention to the j^{th} component, it follows that $\sqrt{n}(\hat{\beta}_j-\beta_0{}^{[j]}) \xrightarrow{d} N\left(0, \left(\boldsymbol{Q}_{XX}^{-1}\Omega\boldsymbol{\Sigma}_{XX}^{-1}\right)_{jj}\right)$ and $(\hat{\boldsymbol{Q}}_{XX}^{-1}\hat{\Omega}\hat{\boldsymbol{Q}}_{XX}^{-1})_{jj} \xrightarrow{p} (\boldsymbol{Q}_{XX}^{-1}\Omega\boldsymbol{Q}_{XX}^{-1})_{jj}$. Theorem **??** and Assumption **??** imply that $(\hat{\boldsymbol{Q}}_{XX}^{-1}\hat{\Omega}\hat{\boldsymbol{Q}}_{XX}^{-1})_{jj}^{-1} \xrightarrow{p} (\boldsymbol{Q}_{XX}^{-1}\Omega\boldsymbol{Q}_{XX}^{-1})_{jj}^{-1}$, and the claim then follows.

2. Rewrite $W = \sqrt{n}(\boldsymbol{R}\hat{\beta} - \boldsymbol{r})'(\boldsymbol{R}\hat{\boldsymbol{Q}}_{XX}^{-1}\hat{\Omega}\hat{\boldsymbol{Q}}_{XX}^{-1}\boldsymbol{R}')^{-1}\sqrt{n}(\boldsymbol{R}\hat{\beta} - \boldsymbol{r})$. If H_0 holds, then $\sqrt{n}(\boldsymbol{R}\hat{\beta} - \boldsymbol{r}) = \boldsymbol{R}\sqrt{n}(\hat{\beta}-\beta_0)$, but by Theorem **??** (part 2), $\sqrt{n}(\boldsymbol{R}\hat{\beta}-\boldsymbol{r}) \xrightarrow{d} N(\boldsymbol{0}, \boldsymbol{R}\boldsymbol{Q}_{XX}^{-1}\Omega\boldsymbol{Q}_{XX}^{-1}\boldsymbol{R}')$. By Theorem **??** (part 3), $\boldsymbol{R}\hat{\boldsymbol{Q}}_{XX}^{-1}\hat{\Omega}\hat{\boldsymbol{Q}}_{XX}^{-1}\boldsymbol{R}' \xrightarrow{p} \boldsymbol{R}\boldsymbol{Q}_{XX}^{-1}\Omega\boldsymbol{Q}_{XX}^{-1}\boldsymbol{R}'$. The claim then follows.

■

This theorem gives us the tools to test linear hypotheses. Specifically, compute the relevant test statistic and compare it to the corresponding quantile of a Normal respectively Chi-squared distribution. The analog of the F-statistic is called Wald test. Recall that with the t-test, you have to distinguish between one- and two-sided tests! This is not the case with the Wald test. (For a visual intuition, recall that $\chi^2(\#\boldsymbol{r})$ is the distribution of the squared length of a $\#\boldsymbol{r}$-dimensional standard normal random vector.)

Notice also that the above test statistics are not quite 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," after Hal White) test statistics that you can find in the output of any OLS routine. If we assume homoskedasticity, some expressions accordingly simplify:

Corollary 2.5 *Let* $\mathbb{E}(e_i^2|X_i) = \sigma^2$*, then:*

1. $\sqrt{n}(\hat{\beta} - \beta_0) \xrightarrow{d} N(0, \sigma^2 \boldsymbol{Q}_{XX}^{-1})$.
2. $s^2 \hat{\boldsymbol{Q}}_{XX}^{-1} \xrightarrow{p} \sigma^2 \boldsymbol{Q}_{XX}^{-1}$.
3. *The hypothesis tests achieve asymptotic size control using the finite-sample t-ratio respectively* $\#\boldsymbol{r}$ *times the finite sample F-statistic.*

Here, part 2 follows from the SLLN. We therefore see that under conditional homoskedasticity, large sample analysis of OLS leads to the same practical conclusions as finite sample analysis, though of course the justification is weaker.

Part 2 also shows that under homoskedasticity, estimation of Ω is not an issue. In contrast, we have not yet provided an estimator of Ω for the general case. Indeed, this requires an additional assumption.

Assumption 2.6 ***Finite Fourth Moments***

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

Theorem 2.6 ***Estimation of Variance***

Let Assumption **??** *hold, then* $\hat{\Omega} \equiv \frac{1}{n}\sum_i (Y_i - X_i'\hat{\beta})^2 X_i X_i' \xrightarrow{p} \Omega$.

Proof.

$$\begin{aligned}
\hat{\Omega} &= \frac{1}{n}\sum_i \big(Y_i - X_i'\hat{\beta}\big)^2 X_i X_i' \\
&= \frac{1}{n}\sum_i \big(e_i + X_i'(\beta_0 - \hat{\beta})\big)^2 X_i X_i' \\
&= \underbrace{\frac{1}{n}\sum_i e_i^2 X_i X_i'}_{\xrightarrow{p} \mathbb{E}\boldsymbol{e}_i^2 X_i X_i' = \Omega} + \underbrace{\frac{2}{n}\sum_i e_i X_i'(\beta_0 - \hat{\beta}) X_i X_i'}_{\xrightarrow{p} \mathbf{0}} + \underbrace{\frac{1}{n}\sum_i (\beta_0 - \hat{\beta})' X_i X_i' (\beta_0 - \hat{\beta}) X_i X_i'}_{\xrightarrow{p} \mathbf{0}}.
\end{aligned}$$

Let's elaborate the second claim. (The third one is much easier!) Pick an arbitrary cell $X_{ij}X_{ik}$ of the matrix $X_i X_i'$, then

$$\frac{1}{n}\sum_i e_i X_i'(\beta_0 - \hat{\beta}) X_{ij} X_{ik} \le \frac{1}{n}\sum_i \left\| e_i X_i'(\beta_0 - \hat{\beta}) X_{ij} X_{ik} \right\| \le \left(\frac{1}{n}\sum_i \big(e_i X_i'(\beta_0 - \hat{\beta})\big)^2 \frac{1}{n}\sum_i (X_{ij}X_{ik})^2 \right)^{1/2},$$

where the second inequality is the Cauchy-Schwarz inequality. Finally, $\frac{1}{n}\sum_i (X_{ij}X_{ik})^2 \xrightarrow{p} \mathbb{E}(X_{ik}X_{ij})^2 < \infty$ by Assumption **??**, and the quadratic form $\frac{1}{n}\sum_i \big(e_i X_i'(\beta_0 - \hat{\beta})\big)^2 = (\beta_0 - \hat{\beta})' \left(\frac{1}{n}\sum_i e_i^2 X_i X_i'\right)(\beta_0 - \hat{\beta})$ vanishes because the matrix converges to Ω and the vectors vanish. ■

Assumption **??** is needed because the variance of $\hat{\Omega}$ comes from the fourth moment of X_i, hence this moment must be finite. While the proof is not hard, it introduces some patterns of argument that you'll see again.

2.1 Miscellanea

We wrap up our initial treatment of OLS with two asides. They can be skipped without loss of continuity but I will bring them up at this point in the lecture to establish connections to material you may have seen elsewhere.

2.1.1 Best Linear Predictor Justification of OLS

The linear model we analyzed so far mirrors the historical development of OLS and is also required to give OLS estimands a causal interpretation. However, there exists another popular justification for the OLS estimator that especially rationalizes its use as data summary tool, including in papers that go on to do something rather different from OLS. Thus, suppose you know the population distribution of (Y_i, X_i), so there is nothing to estimate; however, you want to predict a future realization of Y_i after having seen the corresponding realization of X_i. Your loss function is square loss, that is, if you predict $\hat{Y}_i$ but the truth turns out to be Y_i, then you lose $(\hat{Y}_i - Y_i)^2$. If you are furthermore restricted to linear prediction – that is, you must specify a prediction rule s.t. $\hat{Y}_i$ is linear in X_i – then your expected loss $\mathbb{E}(\hat{Y}_i - Y_i)^2$, where the expectation is taken with respect to the true distribution of (Y_i, X_i). Simple algebra shows that this is minimized by the *best linear predictor*

$$\beta^* = \mathbb{E}(X_i X_i')^{-1} \mathbb{E} X_i Y_i.$$

Under the conditions maintained here, we will have $\hat{\beta} \xrightarrow{p} \beta^*$, thus OLS regression can be thought of as estimating the best linear predictor under square loss. This gives a meaningful (although not any more causal) interpretation to OLS even if we do not presume the linear model to be true. One flip side is that one might ask why we restrict attention to linear predictors. Note also that if the true conditional expectation $\mathbb{E}(Y_i|X_i)$ is not linear in X_i, then $\mathbb{E}((Y_i - X_i'\beta^*)^2|X_i)$ will vary with X_i even if the true conditional variance of Y_i does not. Thus, this justification of OLS cannot be reasonably combined with assuming homoskedasticity, and standard errors reported with OLS if justified in this manner should always be robust.

2.1.2 Instrumental Variables

OLS critically relies on the assumption that regressors are in some sense exogenous. 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 $\mathbb{E}(\ln schooling_i \varepsilon_i) = 0$ 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 innate ability, so that we expect the two to be positively correlated. Assuming that ability positively impacts wages, then with ability not "controlled for," the schooling variable will pick up some of its effect, and β will be overestimated. This is an example of *omitted variable bias* caused by the omission of ability from the equation, but since innate ability is not directly observable, it cannot be fixed by including the omitted variable.

Imagine now that we also observe a random variable Z_i that is correlated with $\ln schooling$ but not with ε. Intuitively, Z_i can be thought of as shifting $schooling_i$ without affecting ε_i and therefore as inducing exogenous variation in $schooling_i$. Then Z_i is called an *instrument*.

Given x_i, we can consistently estimate β by

$$\hat{\beta}_{IV} \equiv \left(\frac{1}{n} \sum_i Z_i \ln schooling_i \right)^{-1} \frac{1}{n} \sum_i Z_i \ln wage_i,$$

the *instrumental variables estimator*. In this extremely simple case, the estimator can be heuristically motivated by writing

$$\begin{aligned} \operatorname{cov}(Z_i, Y_i) &= \beta_0 \operatorname{cov}(Z_i, \ln schooling_i) + \operatorname{cov}(Z_i, \varepsilon_i) \\ \Rightarrow \beta_0 &= \frac{\operatorname{cov}(Z_i, Y_i)}{\operatorname{cov}(Z_i, \ln schooling_i)}, \end{aligned}$$

showing that $\hat{\beta}_{IV}$ is really a sample analog of β_0. (This also instantly clarifies why we need both positive correlation with the endogenous regressor and zero correlation with the error term.)

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.

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

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

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.

An example going back to the 1940's is the simultaneous equation model from macroeconomics. The very simple such model in Haavelmo (1943) is

$$\begin{aligned} c_i &= \alpha_0 + \alpha_1 y_i + \varepsilon_i \\ y_i &= c_i + i_i, \end{aligned}$$

where y is income, c is consumption, i is investment, and $\alpha_1 \in (0, 1)$ is the marginal propensity to consume. Here, y_i is endogenous in the first equation but can be instrumented for by i_i.

3 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 eventually, e.g. in time series. Its name is due to a famous paper by Hansen (1982). As often with big ideas, some elements already floated around in related work (notably by Amemiya and Sargan).

Like the method of moments, which it generalizes, GMM starts by postulating *moment conditions* which are supposed to be true in population. The difference is that there may be more moment conditions than free parameters. (A "meta-contribution" was to recognize how generally applicable the principle is in economics, e.g. for estimating parameers off Euler equations without specifying a likelihood.)

Reminder: Method of Moments

Let W_i be some random vector and let θ_0 denote the true value of some parameter of interest θ. In this lecture, θ will always be finite dimensional, that is, $\theta \in \Theta \subseteq \mathbb{R}^k$ for some $k < \infty$. Then a moment condition looks like this:

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

Here, g is a known function of 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 $\hat{\theta}$ is constructed by solving the sample analogs of the moment conditions, i.e. it is implicitly defined by

$$\frac{1}{n}\sum_{i=1}^{n} g(W_i, \hat{\theta}) = \mathbf{0}.$$

(Thus, MM estimators are special cases of *analog estimators*.) To be clear, neither existence nor uniqueness of $\hat{\theta}$ are obvious at this level of generality, but they are in many (not all) actual uses of the method.

Example 3.1 ***Ordinary Least Squares***

Let

$$W_i = (Y_i, X_i),$$

where Y_i is a scalar that we want to predict/explain and X_i is a vector of regressors. We suppose that Y_i is generated by a linear process:

$$Y_i = X_i'\theta_0 + e_i$$

and also that X_i is exogenous, i.e. the error term e_i is uncorrelated with X_i:

$$\mathbb{E}X_i e_i = \mathbf{0}.$$

This can be written as a moment condition. Set $g(W_i, \theta) = X_i (Y_i - X_i'\theta)$*, then the moment condition is that*

$$\mathbb{E}\left(X_i \left(Y_i - X_i'\theta_0\right)\right) = \mathbf{0}.$$

Of course, these are some of the assumption underlying OLS. Indeed, it is easy to see that solving the above equation's sample analog (i.e., replacing expectations by sample means) for θ*, we recover the OLS estimator. Hence, OLS is a method of moments estimator.*

Example 3.2 ***Instrumental Variables***

Let

$$W_i = (Y_i, X_i, Z_i),$$

where Y_i *is again a scalar and* X_i *and* Z_i *are vectors of regressors, not necessarily disjoint. We still impose that*

$$Y_i = X_i'\theta_0 + e_i,$$

but X_i *might be endogenous, i.e. correlated with the errors. However, we are willing to believe that* Z_i *is exogenous:*

$$\mathbb{E}Z_i e_i = \mathbf{0}.$$

This can again be written as a moment condition:

$$\mathbb{E}\left(Z_i \left(Y_i - X_i'\theta_0\right)\right) = \mathbf{0}.$$

If Z_i *is correlated with* X_i *and a rank condition holds, then at population level this moment condition is uniquely solved by* θ_0*. Again, it is easy to see that solving the sample analog of the moment condition for* θ *yields the IV estimator.*

Example 3.3 ***Poisson Regression***

Let

$$Y_i = \exp(X_i'\theta_0) + e_i,$$

then one could write down a moment condition

$$\mathbb{E}\left(Z_i \left(Y_i - \exp(X_i'\theta_0)\right)\right) = \mathbf{0}.$$

GMM immediately handles this case as well. If Z_i *and* X_i *have the same number of components, it will among other things recover Nonlinear Least Squares. We will not look more closely at this case for the moment because we initially restrict ourselves to linear moment conditions.*

3.1 Linear GMM: The Helicopter Tour

We next provide a brief overview over linear GMM. We want to estimate a linear equation:

$$Y_i = X_i'\theta_0 + \varepsilon_i.$$

We assume that the random variable

$$W_i = (Y_i, X_i, Z_i)$$

is i.i.d. (Just to be clear, mutual independence of Y_i, X_i, and Z_i is of course not assumed. Also, the method does not crucially rely on an i.i.d. assumption.)

We have the moment conditions

$$\mathbb{E}\left(Z_i\left(Y_i - X_i'\theta_0\right)\right) = \mathbf{0}.$$

The aim is to estimate θ_0. 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. X_i and Z_i need not have the same number of components. Also, we will henceforth think of

$$X_i = Z_i$$

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

3.1.1 Identification

Let

$$\begin{aligned} l &= \text{number of moment conditions (thus } Z_i \in \mathbb{R}^l) \\ k &= \text{number of parameters (thus } X_i \in \mathbb{R}^k), \end{aligned}$$

and also assume that $\underset{l\times k}{\boldsymbol{Q}} = \mathbb{E}(Z_i X_i')$ has maximal rank. Intuitively, this says that no instrument/moment condition is redundant.

Then the above model is

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

Underidentified we already know. It plainly means there are more unknowns than equations, so even if we knew the population distribution of W, we could not solve for θ_0. Econometric analysis of

underidentified models is an active area of research and is in particular what your instructor works on, but 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*. Overidentification means that there are more linear equations than unknowns. Such a system of equations has generically no solution. Of course, by assuming that our moment conditions hold true true, we are assuming that the (population level) system is not generic in this sense. If we knew the true distribution of W, the system would turn out to have an (overdetermined) 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. (I contrast, ina just identified system of equations, the equations always define some θ_0, though if the conditions are misspecified, this so-called *pseudotrue* parameter value may not be of any substantive interest.)

3.1.2 Estimation

A new difficulty arises in estimation of overidentified models. The empirical distribution of W will not be the population distribution, and because overdetermined systems of equations are "typically" inconsistent, we expect that attempting to exactly solve the moment conditions' sample analogs will usually lead to a contradiction.

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

$$\frac{1}{n}\sum_{i=1}^{n} Z_i \left(Y_i - X_i'\theta_0\right) = \mathbf{0}$$

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

$$\frac{1}{n}\sum_{i=1}^{n} Z_i \left(Y_i - X_i'\theta_0\right) \xrightarrow{p} \mathbf{0}.$$

If θ_0 is identified, that is if $\mathbb{E}\left(Z_i \left(Y_i - X_i'\theta\right)\right) = \mathbf{0}$ is uniquely solved by θ_0, we will furthermore find

$$\mathbb{E}\big(Z_i\big(Y_i - X_i'\tilde{\theta}\big)\big) \neq \mathbf{0}$$

and consequently

$$\frac{1}{n}\sum_{i=1}^{n} Z_i\big(Y_i - X_i'\tilde{\theta}\big) \to \mathbb{E}\big(Z_i\big(Y_i - X_i'\tilde{\theta}\big)\big) \neq \mathbf{0}$$

for any $\tilde{\theta} \neq \theta_0$. Thus, a natural estimator of θ_0 is

$$\hat{\theta} \equiv \arg\min_{\theta\in\Theta} \left(\frac{1}{n}\sum_{i=1}^{n} Z_i \left(Y_i - X_i'\theta\right)\right)' \left(\frac{1}{n}\sum_{i=1}^{n} Z_i \left(Y_i - X_i'\theta\right)\right).$$

3.1.3 Testing

Once the model has been estimated, we can test hypotheses about θ_0. But overidentified GMM also allows for directly testing the model's assumptions.

If the moment conditions are correct, we will have $\hat{\theta} \xrightarrow{p} \theta_0$ and

$$\left(\frac{1}{n}\sum_{i=1}^{n} Z_i\big(Y_i - X_i'\hat{\theta}_n\big)\right)' \left(\frac{1}{n}\sum_{i=1}^{n} Z_i\big(Y_i - X_i'\hat{\theta}_n\big)\right) \xrightarrow{p} \mathbb{E}\left(Z_i\left(Y_i - X_i'\theta_0\right)\right)' \mathbb{E}\left(Z_i\left(Y_i - X_i'\theta_0\right)\right) = 0.$$

In contrast, if some of the moment conditions are wrong, the population system of conditions will typically be inconsistent. As a result, we should still expect that

$$\hat{\theta} \xrightarrow{p} \theta^* \equiv \arg\min_{\theta\in\Theta}\big\{\mathbb{E}\left(Z_i\left(Y_i - X_i'\theta\right)\right)' \mathbb{E}\left(Z_i\left(Y_i - X_i'\theta\right)\right)\big\},$$

where θ^* is sometimes called the *pseudotrue* value of θ, but also that

$$\begin{aligned} \mathbb{E}\left(Z_i\left(Y_i - X_i'\theta^*\right)\right) &\neq \mathbf{0} \\ \Longrightarrow \mathbb{E}\left(Z_i\left(Y_i - X_i'\theta^*\right)\right)' \mathbb{E}\left(Z_i\left(Y_i - X_i'\theta^*\right)\right) &> 0 \end{aligned}$$

and hence that $\big(\frac{1}{n}\sum_{i=1}^{n} Z_i\big(Y_i - X_i'\hat{\theta}_n\big)\big)'\big(\frac{1}{n}\sum_{i=1}^{n} Z_i\big(Y_i - X_i'\hat{\theta}_n\big)\big)$ does not converge to zero. This suggests a test: Reject the moment conditions if

$$\left(\frac{1}{n}\sum_{i=1}^{n} Z_i\big(Y_i - X_i'\hat{\theta}_n\big)\right)' \left(\frac{1}{n}\sum_{i=1}^{n} Z_i\big(Y_i - X_i'\hat{\theta}_n\big)\right)$$

is too large.

The core thing we overlooked is that by just squaring $\frac{1}{n}\sum_{i=1}^{n} Z_i\big(Y_i - X_i'\hat{\theta}_n\big)$, 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.

3.2 Linear GMM: Formal Statement

Define the sample analog of $\mathbb{E}g(Y_i, X_i, Z_i, \theta)$ as

$$\bar{g}_n(\theta) \equiv \frac{1}{n}\sum_{i=1}^{n} g(Y_i, X_i, Z_i, \theta).$$

Also fix any symmetric and positive definite weighting matrix $\boldsymbol{W}$ and let $\hat{\boldsymbol{W}}$ be an estimator of $\boldsymbol{W}$, i.e. $\hat{\boldsymbol{W}} \xrightarrow{p} \boldsymbol{W}$. This allows for $\hat{\boldsymbol{W}}$ to be data-dependent, but a constant pre-assigned $\hat{\boldsymbol{W}} = \boldsymbol{W}$ is possible as well.

The GMM estimator $\hat{\theta}(\hat{\boldsymbol{W}})$ is

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

3.2.1 Specialization to Linear Moment Conditions

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

Define the sample moments

$$\begin{aligned}\hat{\boldsymbol{Q}}_{ZY} &\equiv \frac{1}{n}\sum_{i=1}^{n} Z_i Y_i, \\ \hat{\boldsymbol{Q}} &\equiv \frac{1}{n}\sum_{i=1}^{n} Z_i X_i'.\end{aligned}$$

Then we can write

$$\begin{aligned}\bar{g}_n(\theta) &= \frac{1}{n}\sum_{i=1}^{n} g(Y_i, X_i, Z_i, \theta) \\ &= \frac{1}{n}\sum_{i=1}^{n} Z_i(Y_i - X_i'\theta) \\ &= \frac{1}{n}\sum_{i=1}^{n} Z_i Y_i - \left(\frac{1}{n}\sum_{i=1}^{n} Z_i X_i'\right)\theta \\ &= \hat{\boldsymbol{Q}}_{ZY} - \hat{\boldsymbol{Q}}\theta.\end{aligned}$$

The moment conditions' sample analog reduces to

$$\hat{\boldsymbol{Q}}\theta = \hat{\boldsymbol{Q}}_{ZY}.$$

The GMM objective function becomes

$$J(\theta, \hat{\boldsymbol{W}}) = n \cdot (\hat{\boldsymbol{Q}}_{ZY} - \hat{\boldsymbol{Q}}\theta)'\hat{\boldsymbol{W}}(\hat{\boldsymbol{Q}}_{ZY} - \hat{\boldsymbol{Q}}\theta).$$

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

$$\begin{aligned}-2n\hat{\boldsymbol{Q}}'\hat{\boldsymbol{W}}(\hat{\boldsymbol{Q}}_{ZY} - \hat{\boldsymbol{Q}}\theta) &= \mathbf{0} \\ \Rightarrow \hat{\boldsymbol{Q}}'\hat{\boldsymbol{W}}\hat{\boldsymbol{Q}}_{ZY} &= \hat{\boldsymbol{Q}}'\hat{\boldsymbol{W}}\hat{\boldsymbol{Q}}\theta.\end{aligned}$$

For this to have a unique solution, we need $\hat{\boldsymbol{Q}}$ to be of full column rank. But this is given (with probability approaching 1) since $\hat{\boldsymbol{Q}} \xrightarrow{a.s.} \boldsymbol{Q}$ by the LLN and $\boldsymbol{Q}$ has full column rank by assumption. Since also $\hat{\boldsymbol{W}}$ is (w.p.a. 1) positive definite by the same argument, $\hat{\boldsymbol{Q}}'\hat{\boldsymbol{W}}\hat{\boldsymbol{Q}}$ is (w.p.a. 1) invertible. Hence, the GMM estimator

$$\hat{\theta}(\hat{\boldsymbol{W}}) = \left(\hat{\boldsymbol{Q}}'\hat{\boldsymbol{W}}\hat{\boldsymbol{Q}}\right)^{-1}\hat{\boldsymbol{Q}}'\hat{\boldsymbol{W}}\hat{\boldsymbol{Q}}_{ZY}$$

is (w.p.a. 1) well-defined. We will now look at its asymptotic properties.

3.2.2 Consistency and Asymptotic Distribution

We make the following assumptions.

Assumption 3.1 ***Linear Model***

$$Y_i = X_i'\theta_0 + e_i.$$

Assumption 3.2 ***IID***

$$W_i = (Y_i, X_i, Z_i) \text{ is jointly i.i.d.}$$

Assumption 3.3 ***Moment Conditions***

$$\mathbb{E}\left(Z_i(Y_i - X_i'\theta_0)\right) = \mathbf{0}.$$

Assumption 3.4 ***Rank Condition***

$$\boldsymbol{Q} \equiv \mathbb{E}(Z_i X_i') \text{ is of full column rank.}$$

Assumption 3.5 ***Regularity Conditions on Errors***

$$\Omega \equiv \mathbb{E}\left(Z_i e_i \left(Z_i e_i\right)'\right) \text{ is nonsingular.}$$

Theorem 3.1 ***Limiting Distribution of the GMM Estimator***

Under the above assumptions, we have:

1.

$$\hat{\theta}(\hat{\boldsymbol{W}}) \xrightarrow{p} \theta_0.$$

2.

$$\sqrt{n}\big(\hat{\theta}(\hat{\boldsymbol{W}}) - \theta_0\big) \xrightarrow{d} N\big(\mathbf{0}, \operatorname{Avar}\big(\hat{\theta}(\hat{\boldsymbol{W}})\big)\big)$$
$$\operatorname{Avar}\big(\hat{\theta}(\hat{\boldsymbol{W}})\big) = \left(\boldsymbol{Q}'\boldsymbol{W}\boldsymbol{Q}\right)^{-1}\boldsymbol{Q}'\boldsymbol{W}\Omega\boldsymbol{W}\boldsymbol{Q}\left(\boldsymbol{Q}'\boldsymbol{W}\boldsymbol{Q}\right)^{-1}.$$

3. Let $\hat{\Omega} \xrightarrow{p} \Omega$*, then*

$$\hat{\boldsymbol{V}} \equiv \big(\hat{\boldsymbol{Q}}'\hat{\boldsymbol{W}}\hat{\boldsymbol{Q}}\big)^{-1}\hat{\boldsymbol{Q}}'\hat{\boldsymbol{W}}\hat{\Omega}\hat{\boldsymbol{W}}\hat{\boldsymbol{Q}}\big(\hat{\boldsymbol{Q}}'\hat{\boldsymbol{W}}\hat{\boldsymbol{Q}}\big)^{-1} \xrightarrow{p} \operatorname{Avar}\big(\hat{\theta}(\hat{\boldsymbol{W}})\big).$$

Proof. Homework. ■

3.2.3 Efficient GMM

If the model is just identified, the sample moment conditions can be exactly solved, and this should yield the same estimator for any weighting matrix. Indeed, if $l = k$, then $\hat{\boldsymbol{Q}}$ is square and hence (with high probability) invertible, and we can write

$$\begin{aligned}\hat{\theta}(\hat{\boldsymbol{W}}) &= \left(\hat{\boldsymbol{Q}}'\hat{\boldsymbol{W}}\hat{\boldsymbol{Q}}\right)^{-1}\hat{\boldsymbol{Q}}'\hat{\boldsymbol{W}}\hat{\boldsymbol{Q}}_{ZY} \\ &= \hat{\boldsymbol{Q}}^{-1}\hat{\boldsymbol{W}}^{-1}\hat{\boldsymbol{Q}}'^{-1}\hat{\boldsymbol{Q}}'\hat{\boldsymbol{W}}\hat{\boldsymbol{Q}}_{ZY} \\ &= \hat{\boldsymbol{Q}}^{-1}\hat{\boldsymbol{Q}}_{ZY},\end{aligned}$$

the usual IV estimator, which you may also know as $\hat{\beta}_{IV} \equiv (Z'X)^{-1}Z'Y$ or similar (keeping in mind that some texts use Z for regressors and X for instruments). In an overidentified model, $\hat{\theta}(\hat{\boldsymbol{W}})$ will nontrivially depend on $\hat{\boldsymbol{W}}$. The obvious question is whether some $\hat{\boldsymbol{W}}$ is optimal in a well-defined sense. As you will prove, any symmetric, positive definite limit $\boldsymbol{W}$ would ensure consistency. But an intuition is to give more weight to moment conditions that are less noisy. Recalling that

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} Z_i e_i \xrightarrow{d} N\left(\mathbf{0}, \mathbb{E}\left(Z_i e_i \left(Z_i e_i\right)'\right)\right) \equiv N(\mathbf{0}, \Omega),$$

one might conjecture that an estimator of Ω^{-1} would make for a good weighting matrix. This is indeed the case.

Theorem 3.2 ***Efficient GMM***

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

$$\operatorname{Avar}\left(\hat{\theta}(\hat{\boldsymbol{W}})\right) \geq \left(\boldsymbol{Q}'\Omega^{-1}\boldsymbol{Q}\right)^{-1},$$

and this bound is attained whenever $\hat{\boldsymbol{W}}$ is a consistent estimator of Ω^{-1}.

Proof. We want to establish positive semidefiniteness of

$$\left(\boldsymbol{Q}'\boldsymbol{W}\boldsymbol{Q}\right)^{-1}\left(\boldsymbol{Q}'\boldsymbol{W}\Omega\boldsymbol{W}\boldsymbol{Q}\right)\left(\boldsymbol{Q}'\boldsymbol{W}\boldsymbol{Q}\right)^{-1} - \left(\boldsymbol{Q}'\Omega^{-1}\boldsymbol{Q}\right)^{-1}.$$

It is a known fact from matrix algebra that (assuming positive definiteness of both matrices being inverted) this is equivalent to positive semidefiniteness of

$$\begin{aligned}&\boldsymbol{Q}'\Omega^{-1}\boldsymbol{Q} - \left(\boldsymbol{Q}'\boldsymbol{W}\boldsymbol{Q}\right)\left(\boldsymbol{Q}'\boldsymbol{W}\Omega\boldsymbol{W}\boldsymbol{Q}\right)^{-1}\left(\boldsymbol{Q}'\boldsymbol{W}\boldsymbol{Q}\right) \\ = \;&\boldsymbol{Q}'\Omega^{-1}\boldsymbol{Q} - \left(\boldsymbol{Q}'\Omega^{-1/2'}\Omega^{1/2\prime}\boldsymbol{W}\boldsymbol{Q}\right)\left(\boldsymbol{Q}'\boldsymbol{W}\Omega^{1/2}\Omega^{1/2\prime}\boldsymbol{W}\boldsymbol{Q}\right)^{-1}\left(\boldsymbol{Q}'\boldsymbol{W}\Omega^{1/2}\Omega^{-1/2}\boldsymbol{Q}\right),\end{aligned}$$

where $\Omega^{1/2}$ is any invertible matrix such that $\Omega^{1/2}\Omega^{1/2\prime} = \Omega$. Defining $\boldsymbol{M} \equiv \Omega^{1/2\prime}\boldsymbol{W}\boldsymbol{Q}$, the last line can be written as

$$\cdots = \boldsymbol{Q}'\Omega^{-1/2'}\left(\boldsymbol{I}_K - \boldsymbol{M}(\boldsymbol{M}'\boldsymbol{M})^{-1}\boldsymbol{M}'\right)\Omega^{-1/2}\boldsymbol{Q}.$$

Note that $\boldsymbol{I}_K - \boldsymbol{M}(\boldsymbol{M}'\boldsymbol{M})^{-1}\boldsymbol{M}'$ is the annihilator corresponding to the column space of $\boldsymbol{M}$. From this geometric interpretation, it is obviously idempotent (alternatively, verification is an easy exercise). Finally, a symmetric idempotent matrix is positive semidefinite (proving that is again an easy exercise). Hence the claim. It is easily verified that the bound is achieved if $\boldsymbol{W} = \Omega^{-1}$. ■

This raises the question of how to estimate Ω. As before, impose:

Assumption 3.6 ***Finite Fourth Moments*** *The matrix*

$$\mathbb{E}\begin{pmatrix} (Z_{i1}X_{i1})^2 & \cdots & (Z_{i1}X_{ik})^2 \\ \vdots & \ddots & \vdots \\ (Z_{il}X_{i1})^2 & \cdots & (Z_{il}X_{ik})^2 \end{pmatrix}$$

exists and is finite.

Theorem 3.3 ***Estimator of*** Ω

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

$$\begin{aligned} \hat{\Omega} &\equiv \frac{1}{n}\sum_{i=1}^{n} \hat{e}_i^2 Z_i Z_i' \\ \hat{e}_i &\equiv Y_i - X_i'\hat{\theta} \end{aligned}$$

is a consistent estimator of Ω.

Proof. Trivial. ■

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

$$\hat{\theta}(\hat{\Omega}^{-1}) \equiv \left(\hat{\boldsymbol{Q}}'\hat{\Omega}^{-1}\hat{\boldsymbol{Q}}\right)^{-1}\hat{\boldsymbol{Q}}'\hat{\Omega}^{-1}\hat{\boldsymbol{Q}}_{ZY}$$

with asymptotic variance

$$\mathrm{Avar}\left(\hat{\theta}(\hat{\Omega}^{-1})\right) = \left(\boldsymbol{Q}'\Omega^{-1}\boldsymbol{Q}\right)^{-1}$$

that can be estimated by

$$\hat{\boldsymbol{V}}_{2SGMM} \equiv \left(\hat{\boldsymbol{Q}}'\hat{\Omega}^{-1}\hat{\boldsymbol{Q}}\right)^{-1}.$$

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

$$\begin{aligned} \hat{\theta} &\equiv \hat{\theta}(\Omega_{ZZ}^{-1}) \\ \hat{e}_i &\equiv Y_i - X_i'\hat{\theta} \\ \hat{\Omega} &\equiv \frac{1}{n}\sum_{i=1}^{n} \hat{e}_i^2 Z_i Z_i' \\ \hat{\theta}_{2SGMM} &\equiv \hat{\theta}(\hat{\Omega}^{-1}). \end{aligned}$$

Here, 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. A caveat is that in very small samples, error propagation from the two-stage procedure may be the larger effect. In any case, we will see that much of the testing/inference theory requires efficient GMM.

3.2.4 Hypothesis Testing

Theorem 3.4 ***Wald-type Hypothesis Tests***

1. ***(t-statistic)*** *Consider the null hypothesis*

$$H_0 : \theta_0^{[j]} = \theta_j.$$

Under this null,

$$t \equiv \frac{\sqrt{n}\big(\hat{\theta}_j(\hat{\boldsymbol{W}}) - \theta_j\big)}{\sqrt{\big[\hat{\boldsymbol{V}}\big]_{jj}}} = \frac{\hat{\theta}_j(\hat{\boldsymbol{W}}) - \theta_j}{SE_j^*} \xrightarrow{d} N(0,1),$$

where SE_j^* is the robust standard error $\sqrt{\frac{1}{n} \cdot \big[\hat{\boldsymbol{V}}\big]_{jj}}$.

2. ***(Wald Statistic for Linear Hypotheses)*** *Consider the null hypothesis*

$$H_0 : \boldsymbol{R}\theta_0 = \boldsymbol{r}.$$

Let $\#\boldsymbol{r}$ denote the number of restrictions and assume that $\boldsymbol{R}$ is of full row rank. Then under the null,

$$W \equiv n \cdot \big(\boldsymbol{R}\hat{\theta}(\hat{\boldsymbol{W}}) - \boldsymbol{r}\big)' \big(\boldsymbol{R}\hat{\boldsymbol{V}}\boldsymbol{R}'\big)^{-1} \big(\boldsymbol{R}\hat{\theta}(\hat{\boldsymbol{W}}) - \boldsymbol{r}\big) \xrightarrow{d} \chi^2\left(\#\boldsymbol{r}\right).$$

3. ***(Wald Statistic for Nonlinear Hypotheses)*** *Conider the null hypothesis*

$$H_0 : \boldsymbol{a}(\theta_0) = \mathbf{0},$$

where the Jacobian $\boldsymbol{A}$ of $\boldsymbol{a}$ is continuous and of full row rank at θ_0. Then under the null,

$$W \equiv n \cdot \boldsymbol{a}\big(\hat{\theta}(\hat{\boldsymbol{W}})\big)' \big(\boldsymbol{A}\big(\hat{\theta}(\hat{\boldsymbol{W}})\big)\hat{\boldsymbol{V}}\boldsymbol{A}\big(\hat{\theta}(\hat{\boldsymbol{W}})\big)'\big)^{-1} \boldsymbol{a}\big(\hat{\theta}(\hat{\boldsymbol{W}})\big) \xrightarrow{d} \chi^2\left(\#\boldsymbol{a}\right).$$

Proof.

1. Is a special case of the next one.

2. Under the null,

$$\sqrt{n}\big(\boldsymbol{R}\hat{\theta}(\hat{\boldsymbol{W}}) - \boldsymbol{r}\big) \xrightarrow{d} N\Big(\mathbf{0}, \boldsymbol{R}\operatorname{Avar}\big(\hat{\theta}(\hat{\boldsymbol{W}})\big)\boldsymbol{R}'\Big),$$

and we know that $\hat{\boldsymbol{V}}$ consistently estimates $\operatorname{Avar}\big(\hat{\theta}(\hat{\boldsymbol{W}})\big)$.

3. Applying the Delta method (see below) to the above, we find that

$$\sqrt{n}\big(\boldsymbol{a}(\hat{\theta}(\hat{\boldsymbol{W}})) - \mathbf{0}\big) \xrightarrow{d} N\left(\mathbf{0}, \boldsymbol{A}(\theta)\left(\boldsymbol{Q}'\boldsymbol{W}\boldsymbol{Q}\right)^{-1}\boldsymbol{Q}'\boldsymbol{W}\Omega\boldsymbol{W}\boldsymbol{Q}\left(\boldsymbol{Q}'\boldsymbol{W}\boldsymbol{Q}\right)^{-1}\boldsymbol{A}(\theta)'\right),$$

and furthermore $\boldsymbol{A}(\hat{\theta}(\hat{\boldsymbol{W}})) \xrightarrow{p} \boldsymbol{A}(\theta)$ by continuity of $\boldsymbol{A}$.

■

Theorem 3.5 ***Delta Method***

Let $\sqrt{n}\big(\hat{\theta} - \theta\big) \xrightarrow{d} \boldsymbol{z}$ *(where* $\boldsymbol{z}$ *is a random variable) and let the function* $\boldsymbol{f} : \mathbb{R}^L \to \mathbb{R}^M$ *have continuous Jacobian* $\boldsymbol{D}$*, then*

$$\sqrt{n}\big(\boldsymbol{f}(\hat{\theta}) - \boldsymbol{f}(\theta)\big) \xrightarrow{d} \boldsymbol{D}(\theta)\,\boldsymbol{z}.$$

Proof. The Mean Value Theorem yields

$$\boldsymbol{f}(\hat{\theta}) = \boldsymbol{f}(\theta) + \boldsymbol{D}(\tilde{\theta})\big(\hat{\theta} - \theta\big),$$

where $\tilde{\theta}$ lies componentwise between θ and $\hat{\theta}$. Also,

$$\sqrt{n}\big(\hat{\theta} - \theta\big) \xrightarrow{d} \boldsymbol{z} \Rightarrow \hat{\theta} \xrightarrow{p} \theta \Rightarrow \tilde{\theta} \xrightarrow{p} \theta \Rightarrow \boldsymbol{D}(\tilde{\theta}) \xrightarrow{p} \boldsymbol{D}(\theta).$$

Taking these together, we have

$$\sqrt{n}\big(\boldsymbol{f}(\hat{\theta}) - \boldsymbol{f}(\theta)\big) = \underbrace{\boldsymbol{D}(\tilde{\theta})}_{\xrightarrow{p}\boldsymbol{D}(\theta)}\underbrace{\sqrt{n}\big(\hat{\theta} - \theta\big)}_{\xrightarrow{d}\boldsymbol{z}},$$

hence the claim. ■

We also briefly introduce the class of *Likelihood Ratio Statistics* which will be analyzed in more detail later. Consider the same null hypothesis as for the nonlinear Wald statistic. To construct a LR test statistic for it, we define the restricted estimator:

$$\begin{aligned}\tilde{\theta}\big(\hat{\Omega}^{-1}\big) \quad &\equiv \quad \arg\min_{\theta} J\big(\theta, \hat{\Omega}^{-1}\big) \\ &\qquad \text{s.t. } \theta \text{ fulfils } H_0.\end{aligned}$$

Notice that unlike with the Wald-type statistics, we restrict attention to efficient GMM.

We then define

$$LR \equiv J\big(\tilde{\theta}(\hat{\Omega}^{-1}), \hat{\Omega}^{-1}\big) - J\big(\hat{\theta}(\hat{\Omega}^{-1}), \hat{\Omega}^{-1}\big).$$

(Is it important that $\hat{\Omega}$ be the same estimator in both summands?)

We state without proof the following:

Theorem 3.6 ***Likelihood Ratio Statistic***

Impose all Assumptions **??-??** *and consider the null hypothesis* $H_0 : \boldsymbol{a}(\theta) = \mathbf{0}$ *as above. Then*

$$LR \xrightarrow{d} \chi^2_{\#\boldsymbol{a}}$$

and

$$LR - W \xrightarrow{p} 0.$$

(Are these statements nested?)

Furthermore, if H_0 *is linear and both are based on the same estimator* $\hat{\Omega}$*, then LR and W are numerically equivalent.*

The Theorem implies that our choice of test statistic does not matter in the limit, and sometimes not even in finite samples. However, the test statistics do not otherwise agree in finite samples. Some relevant considerations for choosing a test statistic in your application are:

- Invariance: Analytically equivalent H_0 will lead to analytically equivalent LR test statistics but not necessarily W statistics. (Why?)
- W may be easier to compute as it requires only the unconstrained estimate.
- W is asymptotically chi-squared for any weighting matrix.

3.2.5 Specification Testing

Recall the intuition given earlier: If the moment conditions are true, then it should be the case that

$$\bar{g}_n(\hat{\theta})' \hat{\boldsymbol{W}} \bar{g}_n(\hat{\theta}) \xrightarrow{p} 0.$$

On the other hand, this will not happen if the moment conditions are false. It should, therefore, be possible to test the model specification by checking whether $\bar{g}_n(\hat{\theta})' \hat{\boldsymbol{W}} \bar{g}_n(\hat{\theta})$ is too large.

However, to develop a test, we need a statistic whose distribution converges to something non-degenerate under the null. The above objective is not such a statistic. This is why $J(\hat{\theta}(\hat{\boldsymbol{W}}), \hat{\boldsymbol{W}})$ has been rescaled by a factor n. This leads to:

Theorem 3.7 ***Specification Test (Hansen 1982)***

Under this chapter's maintained assumptions,

$$J\big(\hat{\theta}(\hat{\Omega}^{-1}), \hat{\Omega}^{-1}\big) \xrightarrow{d} \chi^2_{l-k}.$$

Proof. By the Mean Value Theorem, we can write

$$\sqrt{n}\bar{g}_n(\hat{\theta}) = \sqrt{n}\bar{g}_n(\theta_0) + \sqrt{n}\boldsymbol{G}_n(\tilde{\theta}) \cdot (\hat{\theta} - \theta_0),$$

where $\boldsymbol{G}_n$ is the Jacobian of $\bar{g}_n$ and $\tilde{\theta}$ lies componentwise between θ_0 and $\hat{\theta}$.

Linearity here leads to a slight simplification because

$$\begin{aligned} \bar{g}_n(\theta) &= \frac{1}{n}\sum_{i=1}^{n} \bar{g}(Y_i, X_i, Z_i, \theta) = \frac{1}{n}\sum_{i=1}^{n} Z_i(Y_i - X_i'\theta) \\ \implies \boldsymbol{G}_n(\theta) &= -\frac{1}{n}\sum_{i=1}^{n} Z_i X_i' = -\hat{\boldsymbol{Q}} \end{aligned}$$

independently of the value taken by θ. (However, linearity is not actually needed for the result.)

We will show in the homework that

$$\hat{\theta}\big(\hat{\Omega}^{-1}\big) - \theta_0 = \big(\hat{\boldsymbol{Q}}'\hat{\Omega}^{-1}\hat{\boldsymbol{Q}}\big)^{-1}\hat{\boldsymbol{Q}}'\hat{\Omega}^{-1}\bar{g}_n(\theta_0).$$

Substituting in for these findings yields

$$\begin{aligned} \sqrt{n}\bar{g}_n(\hat{\theta}) &= \sqrt{n}\bar{g}_n(\theta_0) - \sqrt{n}\hat{\boldsymbol{Q}}\big(\hat{\boldsymbol{Q}}'\hat{\Omega}^{-1}\hat{\boldsymbol{Q}}\big)^{-1}\hat{\boldsymbol{Q}}'\hat{\Omega}^{-1}\bar{g}_n(\theta_0) \\ &= \sqrt{n}\bar{g}_n(\theta_0) - \sqrt{n}\boldsymbol{Q}\left(\boldsymbol{Q}'\Omega^{-1}\boldsymbol{Q}\right)^{-1}\boldsymbol{Q}'\Omega^{-1}\bar{g}_n(\theta_0) + o_p(1), \end{aligned}$$

where the second step holds because $\hat{\boldsymbol{Q}}\big(\hat{\boldsymbol{Q}}'\hat{\Omega}^{-1}\hat{\boldsymbol{Q}}\big)^{-1}\hat{\boldsymbol{Q}}'\hat{\Omega}^{-1} \xrightarrow{p} \boldsymbol{Q}\left(\boldsymbol{Q}'\Omega^{-1}\boldsymbol{Q}\right)^{-1}\boldsymbol{Q}'\Omega^{-1}$ (by arguments very similar to before) and $\sqrt{n}\bar{g}_n(\theta_0) = O_P(1)$. Next, define

$$\boldsymbol{M} \equiv \boldsymbol{Q}\left(\boldsymbol{Q}'\Omega^{-1}\boldsymbol{Q}\right)^{-1}\boldsymbol{Q}',$$

then

$$\begin{aligned} \sqrt{n}\bar{g}_n(\hat{\theta}) &= \big(\boldsymbol{I}_K - \boldsymbol{M}\Omega^{-1}\big)\sqrt{n}\bar{g}_n(\theta_0) + o_p(1) \\ \implies \Omega^{-1/2}\sqrt{n}\bar{g}_n(\hat{\theta}) &= \Omega^{-1/2}\big(\boldsymbol{I}_K - \boldsymbol{M}\Omega^{-1}\big)\sqrt{n}\bar{g}_n(\theta_0) + o_p(1) \\ &= \big(\boldsymbol{I}_K - \Omega^{-1/2}\boldsymbol{M}\Omega^{-1/2\prime}\big)\Omega^{-1/2}\sqrt{n}\bar{g}_n(\theta_0) + o_p(1). \end{aligned}$$

Consider squaring the l.h.s. of this, then we get

$$\big(\Omega^{-1/2}\sqrt{n}\bar{g}_n(\hat{\theta})\big)'\Omega^{-1/2}\sqrt{n}\bar{g}_n(\hat{\theta}) = n\cdot\big(\bar{g}_n(\hat{\theta})'\Omega^{-1/2\prime}\Omega^{-1/2}\bar{g}_n(\hat{\theta})\big) = J\big(\hat{\theta}\big(\hat{\Omega}^{-1}\big), \hat{\Omega}^{-1}\big) + o_p(1).$$

So to establish the test statistic's distribution, we need to find out how the square of the r.h.s. will be distributed. We firstly observe that under the null,

$$\sqrt{n}\bar{g}_n(\theta_0) \xrightarrow{d} N\big(\boldsymbol{0}, \Omega\big),$$

thus

$$\Omega^{-1/2}\sqrt{n}\bar{g}_n(\theta_0) \xrightarrow{d} N\big(\boldsymbol{0}, (\Omega^{-1/2})'\Omega\Omega^{-1/2}\big) = N\big(\boldsymbol{0}, \boldsymbol{I}_l\big).$$

It remains to analyze $\big(\boldsymbol{I}_l - \Omega^{-1/2}\boldsymbol{M}\Omega^{-1/2}\big)$. Undoing the definition of $\boldsymbol{M}$, we find

$$\begin{aligned} \boldsymbol{I}_l - \Omega^{-1/2}\boldsymbol{M}(\Omega^{-1/2})' &= \boldsymbol{I}_l - \Omega^{-1/2}\boldsymbol{Q}\big(\boldsymbol{Q}'(\Omega^{-1/2})'\Omega^{-1/2}\boldsymbol{Q}\big)^{-1}\boldsymbol{Q}'(\Omega^{-1/2})' \\ &= \boldsymbol{I}_l - \tilde{\boldsymbol{M}}\big(\tilde{\boldsymbol{M}}'\tilde{\boldsymbol{M}}\big)^{-1}\tilde{\boldsymbol{M}}', \end{aligned}$$

where $\tilde{\boldsymbol{M}} \equiv \Omega^{-1/2}\boldsymbol{Q}$.

This is again a projection, hence idempotent as well as symmetric. (Aside: A matrix can be interpreted as projection iff it is idempotent and symmetric.) Thus, the square of the r.h.s. boils down to the quadratic form of a multivariate standard normal with a projection matrix. This is known to generate the $\chi^2(r)$ distribution, where r is the rank of the transformation matrix (and hence the dimensionality of the transformation's image).[1]

It therefore remains to find the rank of $\big(\boldsymbol{I}_l - \tilde{\boldsymbol{M}}(\tilde{\boldsymbol{M}}'\tilde{\boldsymbol{M}})^{-1}\tilde{\boldsymbol{M}}'\big)$. This requires some matrix algebraic tricks: The rank of an idempotent matrix equals its trace, and the trace operator is linear and commutative. Hence, we can write

$$\begin{aligned} r &= \operatorname{tr}\big(\boldsymbol{I}_l - \tilde{\boldsymbol{M}}(\tilde{\boldsymbol{M}}'\tilde{\boldsymbol{M}})^{-1}\tilde{\boldsymbol{M}}'\big) \\ &= \operatorname{tr}\left(\boldsymbol{I}_l\right) - \operatorname{tr}\big(\tilde{\boldsymbol{M}}(\tilde{\boldsymbol{M}}'\tilde{\boldsymbol{M}})^{-1}\tilde{\boldsymbol{M}}'\big) \\ &= \operatorname{tr}\left(\boldsymbol{I}_l\right) - \operatorname{tr}\big(\tilde{\boldsymbol{M}}'\tilde{\boldsymbol{M}}(\tilde{\boldsymbol{M}}'\tilde{\boldsymbol{M}})^{-1}\big) \\ &= \operatorname{tr}\left(\boldsymbol{I}_l\right) - \operatorname{tr}\left(\boldsymbol{I}_k\right) \\ &= l - k. \end{aligned}$$

■

We finally mention the possibility of testing a subset of moment conditions. Thus, suppose we want to maintain the first $l_1 \leq l$ conditions and test the last $(l - l_1)$ ones. Partition Z_i as follows:

$$\underset{[l\times 1]}{Z_i} = \begin{pmatrix} \underset{[l_1\times 1]}{Z_{i1}} \\ \\ \underset{[(l-l_1)\times 1]}{Z_{i2}} \end{pmatrix}.$$

Then we want to maintain that $\mathbb{E}(Z_{i1}e_i) = \mathbf{0}$ but test whether $\mathbb{E}(Z_{i2}e_i) = \mathbf{0}$.

The idea behind the test is to see whether including the moment conditions $\mathbb{E}(Z_{i2}e_i) = \mathbf{0}$ increases $J\big(\hat{\theta}_0(\hat{\boldsymbol{W}}), \hat{\boldsymbol{W}}\big)$ by "too much," i.e. by more than would be expected through sampling variation.

The efficient GMM estimator based on the maintained moment conditions is

$$\tilde{\theta}\big(\hat{\Omega}_{11}^{-1}\big) \equiv \big(\hat{\boldsymbol{Q}}'_{Z_1X}\hat{\Omega}_{11}^{-1}\hat{\boldsymbol{Q}}_{Z_1X}\big)^{-1}\hat{\boldsymbol{Q}}'_{Z_1X}\hat{\Omega}_{11}^{-1}\hat{\boldsymbol{Q}}_{Z_1Y},$$

which minimizes

$$\tilde{J}(\theta, \hat{\Omega}_{11}^{-1}) \equiv n \cdot \bar{g}_{1n}(\theta)'\hat{\Omega}_{11}^{-1}\bar{g}_{1n}(\theta),$$

where the additional subscript 1 indicates that an object has been computed from the first l_1 moment conditions only.

[1] This is a Theorem. More specifically, the quadratic form of a Multivariate Normal Z_i with expectation $\boldsymbol{\mu}$ and variance $\boldsymbol{I}_K$ with any projection matrix $\boldsymbol{P}$ is a noncentral χ^2 with degrees of freedom equal to the rank of $\boldsymbol{P}$ and noncentrality parameter $\boldsymbol{\mu}\boldsymbol{P}\boldsymbol{\mu}'$.

Assume that $\mathbb{E}(Z_{i1}X_i')$ is of full column rank, then $\tilde{\theta}\big(\hat{\Omega}_{11}^{-1}\big)$ is covered by our previous results. Our test statistic will be the increase in J that is due to adding moment conditions, i.e.

$$C \equiv J\big(\hat{\theta}(\hat{\Omega}^{-1}), \hat{\Omega}^{-1}\big) - \tilde{J}\big(\tilde{\theta}(\hat{\Omega}_{11}^{-1}), \hat{\Omega}_{11}^{-1}\big).$$

We state without proof the asymptotic distribution of this, which should not come as a surprise.

Theorem 3.8 ***Specification Test (Newey 1985,Eichenbaum/Hansen/Singleton 1985)***

$$C \xrightarrow{d} \chi^2_{l-l_1}.$$

3.2.6 Imposing Homoskedasticity

Up to this point, we did not impose homoskedasticity. That made sense because this case is very important and allowing for it right away comes at minimal cognitive cost in a modern framework. However, an assumption of homeskedasticty is of course sometimes made and/or appropriate, and it leads to interestng simplifications that furthermore recover estimators which you may have seen before. Thus, consider imposing:

Assumption 3.7 ***Conditional Homoskedasticity***

$$\mathbb{E}(e_i^2|Z_i) = \sigma^2.$$

It immediately follows that

$$\Omega = \sigma^2\mathbb{E}(Z_iZ_i') \equiv \sigma^2\boldsymbol{Q}_{ZZ}.$$

Hence, we do not need Assumption **??** (i.e., finite fourth moments) any more. Let $\hat{\sigma}^2$ be a consistent estimator of σ^2, then it is easy to see that

$$\hat{\Omega} \equiv \hat{\sigma}^2\frac{1}{n}\sum_{i=1}^{n} Z_iZ_i' \equiv \hat{\sigma}^2\hat{\boldsymbol{Q}}_{ZZ} \xrightarrow{p} \Omega.$$

The efficient GMM estimator becomes

$$\begin{aligned}
\hat{\theta}\big(\hat{\Omega}^{-1}\big) &\equiv \big(\hat{\boldsymbol{Q}}'(\hat{\sigma}^2\hat{\boldsymbol{Q}}_{ZZ})^{-1}\hat{\boldsymbol{Q}}\big)^{-1}\hat{\boldsymbol{Q}}'(\hat{\sigma}^2\hat{\boldsymbol{Q}}_{ZZ})^{-1}\hat{\boldsymbol{Q}}_{ZY} \\
&= \left(\hat{\boldsymbol{Q}}'\hat{\boldsymbol{Q}}_{ZZ}^{-1}\hat{\boldsymbol{Q}}\right)^{-1}\hat{\boldsymbol{Q}}'\hat{\boldsymbol{Q}}_{ZZ}^{-1}\hat{\boldsymbol{Q}}_{ZY} \\
&= \hat{\theta}\left(\hat{\boldsymbol{Q}}_{ZZ}^{-1}\right) \\
&\equiv \hat{\theta}_{2SLS}.
\end{aligned}$$

In short, the first step of the two-step estimation process that generated efficient GMM is redundant. We call this estimator $\hat{\theta}_{2SLS}$ 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 $\mathbb{E}(X_i X_i')$ exists and is finite, then

$$\hat{\sigma}^2 \equiv \frac{1}{n}\sum_{i=1}^{n}(Y_i - X_i'\hat{\theta}_{2SLS})^2$$

is a consistent estimator. (We prove a very similar statement in the homework.)

We now find the following simplifications:

$$\operatorname{Avar}\big(\hat{\theta}_{2SLS}\big) = \sigma^2 \cdot \big(\boldsymbol{Q}'\boldsymbol{Q}_{ZZ}^{-1}\boldsymbol{Q}\big)^{-1},$$

which can be estimated by

$$\hat{\boldsymbol{V}} = \hat{\sigma}^2 \cdot \Big(\hat{\boldsymbol{Q}}'\hat{\boldsymbol{Q}}_{ZZ}^{-1}\hat{\boldsymbol{Q}}\Big)^{-1}.$$

The t-statistic becomes

$$t_j = \frac{[\hat{\theta}_{2SLS}]_j - \theta_j}{\Big(\frac{\hat{\sigma}^2}{n}\big[(\hat{\boldsymbol{Q}}'\hat{\boldsymbol{Q}}_{ZZ}^{-1}\hat{\boldsymbol{Q}})^{-1}\big]_{jj}\Big)^{1/2}}.$$

The nonlinear Wald statistic turns into

$$W = n \cdot \frac{\boldsymbol{a}(\hat{\theta}_{2SLS})'\big(\boldsymbol{A}(\hat{\theta}_{2SLS})(\hat{\boldsymbol{Q}}'\hat{\boldsymbol{Q}}_{ZZ}^{-1}\hat{\boldsymbol{Q}})^{-1}\boldsymbol{A}(\hat{\theta}_{2SLS})'\big)^{-1}\boldsymbol{a}(\hat{\theta}_{2SLS})}{\hat{\sigma}^2}.$$

The next two statistics turn out to predate GMM. The likelihood ratio statistic turns into

$$LR = n \cdot \frac{(\hat{\boldsymbol{Q}}_{ZY} - \hat{\boldsymbol{Q}}\tilde{\theta})'\hat{\boldsymbol{Q}}_{ZZ}^{-1}(\hat{\boldsymbol{Q}}_{ZY} - \hat{\boldsymbol{Q}}\tilde{\theta}) - (\hat{\boldsymbol{Q}}_{ZY} - \hat{\boldsymbol{Q}}\hat{\theta}_{2SLS})'\hat{\boldsymbol{Q}}_{ZZ}^{-1}(\hat{\boldsymbol{Q}}_{ZY} - \hat{\boldsymbol{Q}}\hat{\theta}_{2SLS})}{\hat{\sigma}^2},$$

where $\tilde{\theta}$ is the restricted GMM estimator. This statistic is originally due to Gallant and Jorgenson 1979.

Finally, the J-statistic turns into

$$J\big(\hat{\theta}_{2SLS}, (\hat{\sigma}^2\hat{\boldsymbol{Q}}_{ZZ}^{-1})\big) = n \cdot \frac{(\hat{\boldsymbol{Q}}_{ZY} - \hat{\boldsymbol{Q}}\hat{\theta}_{2SLS})'\hat{\boldsymbol{Q}}_{ZZ}^{-1}(\hat{\boldsymbol{Q}}_{ZY} - \hat{\boldsymbol{Q}}\hat{\theta}_{2SLS})}{\hat{\sigma}^2}.$$

This looks much like a Wald-type ("deviation squared over variance") statistic. And indeed, it had previously been derived as such (in a simultaneous equations context) and was known as Sargan's statistic.

3.3 Miscellanea

We wrap up our treatment of linear single equation GMM with some "bits and pieces."

3.3.1 Why is it Called Two Stage Least Squares?

We will henceforth think of 2SLS as linear homeoskedastic overidentified GMM and therefore as a one-stage estimator. The original development was quite different and also explains the estimator's name. In particular, 2SLS can be alternatively conceived as:

- First regressing the endogenous regressor X_i on the instrument Z_i and then regressing Y_i on the fitted values $\hat{X}_i$ (Theil, 1953);
- Same, but in the second stage regress Y_i on X_i by just-identified IV using $\hat{X}_i$ as instrument (Basmann, 1957).

Both implementations further illustrate the intuition that gives instrumental variables their name, namely that an instrument Z_i allows us to extract some movement in the regressor X_i that can be considered exogenous for the regression of Y_i on X_i. To verify that the second one indeed recovers the same 2SLS estimator that we derived, use data matrix notation write (the argument for the first implementation is similar)

$$\begin{aligned}
\hat{\theta}_{Basmann} &= (\hat{\boldsymbol{Z}}'\boldsymbol{Z})^{-1}\hat{\boldsymbol{Z}}'\boldsymbol{Y} \\
&= \left(\left(\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Z}\right)'\boldsymbol{Z}\right)^{-1}\left(\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Z}\right)'\boldsymbol{Y} \\
&= \left((\boldsymbol{X}'\boldsymbol{Z})'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Z}\right)^{-1}(\boldsymbol{X}'\boldsymbol{Z})'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y} \\
&= \left(\hat{\boldsymbol{Q}}'\hat{\boldsymbol{Q}}_{ZZ}^{-1}\hat{\boldsymbol{Q}}\right)^{-1}\hat{\boldsymbol{Q}}'\hat{\boldsymbol{Q}}_{ZZ}^{-1}\hat{\boldsymbol{Q}}_{ZY} \\
&= \hat{\theta}_{2SLS}.
\end{aligned}$$

3.3.2 The Cost of Adding 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. Even if an instrument does not in fact mater, the two-step procedure will pick that up, no? It is intuitively clear that this reasoning cannot be quite right. Two concerns are:

- The above asymptotic analysis is pointwise; in particular, it holds parameter values fixed and then send $n \to \infty$. The result is not uniform: For any n, I can find parameter values for which the asymptotic approximation is misleading. In particular, imagine the extreme case where Z_i and X_i are unrelated, then $\boldsymbol{Q}$ will not have full rank, and we will lose identification. This is excluded by assumption, but our approximation breaks down as this case is approached. This happens with weak instruments, which are the subject of an extensive literature. We do not go into that literature here but point out that layering on instruments increases the risk of adding weak instruments, in which case the asymptotic analysis presented here may be misleading.
- Validity of instruments must be assumed and is not in general testable. So in practice, by adding instruments, we always add assumptions, and our conclusions hinge on all these assumptions being correct (and our ability to convince our audience of this).

Therefore, in practice, there can be compelling reasons to not use all vaguely plausible instruments.

4 Multiple-Equation GMM

4.1 The General Case

We next extend GMM to the simultaneous estimation of multiple equations. This requires to extend our notation and adapt our assumptions. In the process, some but not all assumptions become meaningfully stronger. This is the one aspect we will discuss somewhat carefully. With notation and assumptions in place, asymptotic analysis of Multiple Equation GMM becomes a corollary of Theorem **??**.

A word on notation: In this chapter only, we drop the 0 subscript for the true parameter value so as to avoid double subscripts.

Assumption 4.1 ***Linear Model***

$$Y_{im} = X_{im}'\theta_m + e_{im}, m = 1, \ldots, M.$$

Notice that, while we will make an i.i.d. assumption across observational units, we do not restrict the joint distribution of $(e_{i1}, \ldots, e_{iM})$. For example, if we think of a panel, with i denoting observations and m waves, then this means that we allow for within-person correlation of errors. We also impose no cross-equation restrictions on coefficients, i.e. something like $\theta_1 = \theta_2$. Such restrictions will, of course, be testable. They are natural in many contexts, e.g. repeated measurements or panel data (if θ measures fixed personal characteristics). We will also consider the restriction to common coefficients as a special case.

Example 4.1 ***Two Different (Seemingly Unrelated?) Equations***

$$\begin{aligned} LW69_i &= \alpha_1 + \beta_1 schooling69_i + \gamma_1 IQ_i + \delta_1 experience69_i + \varepsilon_{i1} \\ KWW_i &= \alpha_2 + \beta_2 schooling69_i + \gamma_2 IQ_i + \varepsilon_{i2}, \end{aligned}$$

where LW is log wage and KWW is the knowledge of the world test score. Griliches (1976) estimates this specification.

Example 4.2 ***Panel Data***

The NLSY data used for Hayashi's (chapter 4) empirical exercise contain data of 1969 and 1980 and would therefore allow for estimating a two-period panel:

$$\begin{aligned} LW69_i &= \alpha_1 + \beta_1 \text{schooling69}_i + \gamma_1 \text{IQ}_i + \delta_1 \text{experience69}_i + \varepsilon_{i1} \\ LW80_i &= \alpha_2 + \beta_2 \text{schooling80}_i + \gamma_2 \text{IQ}_i + \delta_2 \text{experience80}_i + \varepsilon_{i2}. \end{aligned}$$

Let the m-th equation have instrument vector Z_{im} with cardinality $l_m \equiv \#Z_{im}$.

Assumption 4.2 ***IID***

$W_i = (Y_{i1}, \ldots, Y_{iM}, X_{i1}, \ldots, X_{iM}, Z_{i1}, \ldots Z_{iM})$ *is i.i.d.*

This assumption does not imply that, for example, Y_{i1} and Y_{i2} are mutually independent. However, the assumption is stronger than imposing that

$$W_{im} = (Y_{im}, X_{im}, Z_{im})$$

is i.i.d. for every m. For a trivial illustration, say $\{X_{i1}\}$ is i.i.d. and $X_{i2} = X_{i-1,1}$.

Assumption 4.3 ***Moment Conditions***

$$\mathbb{E}\left(Z_{im}(Y_{im} - X_{im}'\theta_m)\right) = \mathbf{0}, \forall m.$$

Equivalently,

$$\mathbb{E}g_i(W_i, \theta) = \mathbf{0},$$

where

$$g_i(W_i, \theta) \equiv \begin{bmatrix} Z_{i1}(Y_{i1} - X_{i1}'\theta_1) \\ \vdots \\ Z_{iM}(Y_{iM} - X_{iM}'\theta_M) \end{bmatrix},$$

where

$$\underset{(\sum_{m=1}^{M} k_m \times 1)}{\theta} \equiv \begin{bmatrix} \theta_1 \\ \vdots \\ \theta_M \end{bmatrix}.$$

Note that $\mathbb{E}\left(Z_{i1}(Y_{i2} - X_{i2}'\theta_2)\right) = \mathbf{0}$ is *not* imposed, although it is true that no common component of Z_{i1} and Z_{i2} can be correlated with either e_{i1} or e_{i2}. Hence, unlike Assumption **??**, this assumption is no strengthening of its single-equation counterpart.

We now consider identification. Write

$$\begin{aligned}
\mathbb{E}g_i(W_i, \theta) &= \mathbb{E}\begin{bmatrix} Z_{i1}(Y_{i1} - X_{i1}'\theta_1) \\ \vdots \\ Z_{iM}(Y_{iM} - X_{iM}'\theta_M) \end{bmatrix} \\
&= \mathbb{E}\begin{bmatrix} Z_{i1}Y_{i1} \\ \vdots \\ Z_{iM}Y_{iM} \end{bmatrix} - \mathbb{E}\begin{bmatrix} Z_{i1}X_{i1}'\theta_1 \\ \vdots \\ Z_{iM}X_{iM}'\theta_M \end{bmatrix} \\
[*] &= \mathbb{E}\begin{bmatrix} Z_{i1}Y_{i1} \\ \vdots \\ Z_{iM}Y_{iM} \end{bmatrix} - \mathbb{E}\begin{bmatrix} Z_{i1}X_{i1}' & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \ddots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & Z_{iM}X_{iM}' \end{bmatrix}\theta \\
&\equiv \boldsymbol{Q}_{ZY} - \boldsymbol{Q}_{ZX}\theta,
\end{aligned}$$

where the last equation defines $\boldsymbol{Q}_{ZY}$ and $\boldsymbol{Q}_{ZX}$.

Thus, $\mathbb{E}g_i(W_i, \theta) = \mathbf{0}$ iff $\boldsymbol{Q}_{ZX}\theta = \boldsymbol{Q}_{ZY}$. Just as with single equations, uniqueness of θ obtains iff $\boldsymbol{Q}_{ZX}$ has full column rank. This is the case iff every block of $\boldsymbol{Q}_{ZX}$ has full column rank. Alternatively, inspection of $[*]$ above directly reveals that θ is identified iff every equation individually meets the single-equation identification condition. We thus impose:

Assumption 4.4 ***Rank Condition***

$\boldsymbol{Q}_{ZX}$ is of full column rank.

(Equivalently, $\mathbb{E}\left(Z_{im}X_{im}'\right)$ is of full column rank for each m.)

Assumption 4.5 ***Regularity Condition***

$$\Omega \equiv \mathbb{E}\left(g_i g_i'\right) \text{ is nonsingular.}$$

To find the estimator, we can again write out the objective function and minimize it. In exact analogy to the manipulations of $\mathbb{E}g_i(W_i, \theta)$, we can write

$$\begin{aligned}
\bar{g}_n(\theta) &\equiv \begin{bmatrix} \frac{1}{n}\sum_{i=1}^n Z_{i1}(Y_{i1} - X_{i1}'\theta_1) \\ \vdots \\ \frac{1}{n}\sum_{i=1}^n Z_{iM}(Y_{iM} - X_{iM}'\theta_M) \end{bmatrix} \\
&= \begin{bmatrix} \frac{1}{n}\sum_{i=1}^n Z_{i1}Y_{i1} \\ \vdots \\ \frac{1}{n}\sum_{i=1}^n Z_{iM}Y_{iM} \end{bmatrix} - \begin{bmatrix} \frac{1}{n}\sum_{i=1}^n Z_{i1}X_{i1}'\theta_1 \\ \vdots \\ \frac{1}{n}\sum_{i=1}^n Z_{iM}X_{iM}'\theta_M \end{bmatrix} \\
&= \begin{bmatrix} \frac{1}{n}\sum_{i=1}^n Z_{i1}Y_{i1} \\ \vdots \\ \frac{1}{n}\sum_{i=1}^n Z_{iM}Y_{iM} \end{bmatrix} - \begin{bmatrix} \frac{1}{n}\sum_{i=1}^n Z_{i1}X_{i1}' & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \ddots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \frac{1}{n}\sum_{i=1}^n Z_{iM}X_{iM}' \end{bmatrix}\theta \\
&\equiv \hat{\boldsymbol{Q}}_{ZY} - \hat{\boldsymbol{Q}}_{ZX}\theta,
\end{aligned}$$

where the last equation defines $\hat{\boldsymbol{Q}}_{ZY}$ and $\hat{\boldsymbol{Q}}_{ZX}$.

With this notational trick, the single-equation algebra exactly replicates, and we get the GMM estimator

$$\hat{\theta}(\hat{\boldsymbol{W}}) = \left(\hat{\boldsymbol{Q}}_{ZX}'\hat{\boldsymbol{W}}\hat{\boldsymbol{Q}}_{ZX}\right)^{-1}\hat{\boldsymbol{Q}}_{ZX}'\hat{\boldsymbol{W}}\hat{\boldsymbol{Q}}_{ZY}.$$

We should, however, remember that here, $\hat{\boldsymbol{Q}}_{ZX}$ is block diagonal. Does $\hat{\boldsymbol{W}}$ have to be block diagonal, too? No! In fact, if it were, we would end up with stacked single-equation GMM estimators, which is not the same as a multiple-equation GMM. Inference from one equation toward the other is the whole point of multiple-equation GMM.

To further discuss some aspects of this estimator, we write it out less compactly for $M = 2$.

$$
\begin{aligned}
\hat{\theta}(\hat{\boldsymbol{W}}) &= \begin{bmatrix} \hat{\theta}_1(\hat{\boldsymbol{W}}) \\ \hat{\theta}_2(\hat{\boldsymbol{W}}) \end{bmatrix} \\
&= \left(\begin{bmatrix} \frac{1}{n}\sum_{i=1}^n X_{i1}Z_{i1}' & \mathbf{0} \\ \mathbf{0} & \frac{1}{n}\sum_{i=1}^n X_{i2}Z_{i2}' \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{W}}_{11} & \hat{\boldsymbol{W}}_{12} \\ \hat{\boldsymbol{W}}_{21} & \hat{\boldsymbol{W}}_{22} \end{bmatrix} \begin{bmatrix} \frac{1}{n}\sum_{i=1}^n Z_{i1}X_{i1}' & \mathbf{0} \\ \mathbf{0} & \frac{1}{n}\sum_{i=1}^n Z_{i2}X_{i2}' \end{bmatrix} \right)^{-1} \\
&\quad \cdot \begin{bmatrix} \frac{1}{n}\sum_{i=1}^n X_{i1}Z_{i1}' & \mathbf{0} \\ \mathbf{0} & \frac{1}{n}\sum_{i=1}^n X_{i2}Z_{i2}' \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{W}}_{11} & \hat{\boldsymbol{W}}_{12} \\ \hat{\boldsymbol{W}}_{21} & \hat{\boldsymbol{W}}_{22} \end{bmatrix} \begin{bmatrix} \frac{1}{n}\sum_{i=1}^n Z_{i1}Y_{i1} \\ \frac{1}{n}\sum_{i=1}^n Z_{i2}Y_{i2} \end{bmatrix} \\
&= \left(\begin{bmatrix} \left(\frac{1}{n}\sum_{i=1}^n X_{i1}Z_{i1}'\right)\hat{\boldsymbol{W}}_{11} & \left(\frac{1}{n}\sum_{i=1}^n X_{i1}Z_{i1}'\right)\hat{\boldsymbol{W}}_{12} \\ \left(\frac{1}{n}\sum_{i=1}^n X_{i2}Z_{i2}'\right)\hat{\boldsymbol{W}}_{21} & \left(\frac{1}{n}\sum_{i=1}^n X_{i2}Z_{i2}'\right)\hat{\boldsymbol{W}}_{22} \end{bmatrix} \begin{bmatrix} \frac{1}{n}\sum_{i=1}^n Z_{i1}X_{i1}' & \mathbf{0} \\ \mathbf{0} & \frac{1}{n}\sum_{i=1}^n Z_{i2}X_{i2}\text{`} \end{bmatrix} \right)^{-1} \\
&\quad \cdot \begin{bmatrix} \left(\frac{1}{n}\sum_{i=1}^n X_{i1}Z_{i1}'\right)\hat{\boldsymbol{W}}_{11} & \left(\frac{1}{n}\sum_{i=1}^n X_{i1}Z_{i1}'\right)\hat{\boldsymbol{W}}_{12} \\ \left(\frac{1}{n}\sum_{i=1}^n X_{i2}Z_{i2}'\right)\hat{\boldsymbol{W}}_{21} & \left(\frac{1}{n}\sum_{i=1}^n X_{i2}Z_{i2}'\right)\hat{\boldsymbol{W}}_{22} \end{bmatrix} \begin{bmatrix} \frac{1}{n}\sum_{i=1}^n Z_{i1}Y_{i1} \\ \frac{1}{n}\sum_{i=1}^n Z_{i2}Y_{i2} \end{bmatrix} \\
&= \begin{bmatrix} \left(\frac{1}{n}\sum_{i=1}^n X_{i1}Z_{i1}'\right)\hat{\boldsymbol{W}}_{11}\left(\frac{1}{n}\sum_{i=1}^n Z_{i1}X_{i1}'\right) & \left(\frac{1}{n}\sum_{i=1}^n X_{i1}Z_{i1}'\right)\hat{\boldsymbol{W}}_{12}\left(\frac{1}{n}\sum_{i=1}^n Z_{i2}X_{i2}'\right) \\ \left(\frac{1}{n}\sum_{i=1}^n X_{i2}Z_{i2}'\right)\hat{\boldsymbol{W}}_{21}\left(\frac{1}{n}\sum_{i=1}^n Z_{i1}X_{i1}'\right) & \left(\frac{1}{n}\sum_{i=1}^n X_{i2}Z_{i2}'\right)\hat{\boldsymbol{W}}_{22}\left(\frac{1}{n}\sum_{i=1}^n Z_{i2}X_{i2}'\right) \end{bmatrix}^{-1} \\
&\quad \cdot \begin{bmatrix} \left(\frac{1}{n}\sum_{i=1}^n X_{i1}Z_{i1}'\right)\hat{\boldsymbol{W}}_{11}\left(\frac{1}{n}\sum_{i=1}^n Z_{i1}Y_{i1}\right) + \left(\frac{1}{n}\sum_{i=1}^n X_{i1}Z_{i1}'\right)\hat{\boldsymbol{W}}_{12}\left(\frac{1}{n}\sum_{i=1}^n Z_{i2}Y_{i2}\right) \\ \left(\frac{1}{n}\sum_{i=1}^n X_{i2}Z_{i2}'\right)\hat{\boldsymbol{W}}_{21}\left(\frac{1}{n}\sum_{i=1}^n Z_{i1}Y_{i1}\right) + \left(\frac{1}{n}\sum_{i=1}^n X_{i2}Z_{i2}'\right)\hat{\boldsymbol{W}}_{22}\left(\frac{1}{n}\sum_{i=1}^n Z_{i2}Y_{i2}\right) \end{bmatrix}.
\end{aligned}
$$

If $\hat{\boldsymbol{W}}_{12} = \hat{\boldsymbol{W}}_{21} = \mathbf{0}$, then this expression reduces to

$$
\begin{aligned}
\hat{\theta}(\hat{\boldsymbol{W}}) &= \begin{bmatrix} \hat{\theta}_1(\hat{\boldsymbol{W}}) \\ \hat{\theta}_2(\hat{\boldsymbol{W}}) \end{bmatrix} \\
&= \begin{bmatrix} \left(\frac{1}{n}\sum_{i=1}^n X_{i1}Z_{i1}'\right)\hat{\boldsymbol{W}}_{11}\left(\frac{1}{n}\sum_{i=1}^n Z_{i1}X_{i1}'\right) & \mathbf{0} \\ \mathbf{0} & \left(\frac{1}{n}\sum_{i=1}^n X_{i2}Z_{i2}'\right)\hat{\boldsymbol{W}}_{22}\left(\frac{1}{n}\sum_{i=1}^n Z_{i2}X_{i2}'\right) \end{bmatrix}^{-1} \\
&\quad \cdot \begin{bmatrix} \left(\frac{1}{n}\sum_{i=1}^n X_{i1}Z_{i1}'\right)\hat{\boldsymbol{W}}_{11}\left(\frac{1}{n}\sum_{i=1}^n Z_{i1}Y_{i1}\right) \\ \left(\frac{1}{n}\sum_{i=1}^n X_{i2}Z_{i2}'\right)\hat{\boldsymbol{W}}_{22}\left(\frac{1}{n}\sum_{i=1}^n Z_{i2}Y_{i2}\right) \end{bmatrix} \\
&= \begin{bmatrix} \left(\left(\frac{1}{n}\sum_{i=1}^n X_{i1}Z_{i1}'\right)\hat{\boldsymbol{W}}_{11}\left(\frac{1}{n}\sum_{i=1}^n Z_{i1}X_{i1}'\right)\right)^{-1} & \mathbf{0} \\ \mathbf{0} & \left(\left(\frac{1}{n}\sum_{i=1}^n X_{i2}Z_{i2}'\right)\hat{\boldsymbol{W}}_{22}\left(\frac{1}{n}\sum_{i=1}^n Z_{i2}X_{i2}'\right)\right)^{-1} \end{bmatrix} \\
&\quad \cdot \begin{bmatrix} \left(\frac{1}{n}\sum_{i=1}^n X_{i1}Z_{i1}'\right)\hat{\boldsymbol{W}}_{11}\left(\frac{1}{n}\sum_{i=1}^n Z_{i1}Y_{i1}\right) \\ \left(\frac{1}{n}\sum_{i=1}^n X_{i2}Z_{i2}'\right)\hat{\boldsymbol{W}}_{22}\left(\frac{1}{n}\sum_{i=1}^n Z_{i2}Y_{i2}\right) \end{bmatrix} \\
&= \begin{bmatrix} \left(\left(\sum_{i=1}^n X_{i1}Z_{i1}'\right)\hat{\boldsymbol{W}}_{11}\left(\sum_{i=1}^n Z_{i1}X_{i1}'\right)\right)^{-1}\left(\sum_{i=1}^n X_{i1}Z_{i1}'\right)\hat{\boldsymbol{W}}_{11}\left(\sum_{i=1}^n Z_{i1}Y_{i1}\right) \\ \left(\left(\sum_{i=1}^n X_{i2}Z_{i2}'\right)\hat{\boldsymbol{W}}_{22}\left(\sum_{i=1}^n Z_{i2}X_{i2}'\right)\right)^{-1}\left(\sum_{i=1}^n X_{i2}Z_{i2}'\right)\hat{\boldsymbol{W}}_{22}\left(\sum_{i=1}^n Z_{i2}Y_{i2}\right) \end{bmatrix},
\end{aligned}
$$

i.e. the multiple-equation GMM estimator would just stack single-equation GMM estimators.

4.1.1 Efficient GMM

Optimal weighting works just as before, i.e. by choosing $\hat{\boldsymbol{W}}$ to estimate Ω^{-1}. To do this by an analog estimator, we need to again impose finite fourth moments.

Assumption 4.6 ***Finite Fourth Moments***

Let Z_{imk} denote the k'th element of Z_{im} and X_{ihj} denote the j'th element of X_{ih}. Then

$$\mathbb{E}\left((Z_{imk}X_{ihj})^2\right)$$

exists and is finite for all (k, m, j, h).

Theorem 4.1 *Let $\hat{\theta}$ be consistent for θ, maintain Assumptions* **??-??**, *and assume that $\mathbb{E}(X_{im}X_{ih}')$ as well as $\mathbb{E}(e_{im}e_{ih})$ exist and are finite for all (m, h). Then*

$$\hat{\Omega} \equiv \begin{bmatrix} \frac{1}{n}\sum_{i=1}^n \hat{e}_{i1}\hat{e}_{i1}Z_{i1}Z_{i1}' & \cdots & \frac{1}{n}\sum_{i=1}^n \hat{e}_{i1}\hat{e}_{iM}Z_{i1}Z_{iM}' \\ \vdots & \ddots & \vdots \\ \frac{1}{n}\sum_{i=1}^n \hat{e}_{iM}\hat{e}_{i1}Z_{iM}Z_{i1}' & \cdots & \frac{1}{n}\sum_{i=1}^n \hat{e}_{iM}\hat{e}_{iM}Z_{iM}Z_{iM}' \end{bmatrix},$$

where

$$\hat{e}_{im} \equiv Y_{im} - X_{im}'\hat{\theta}_m,$$

is consistent for Ω.

4.1.2 Hypothesis Testing

Hypothesis testing works just as before. A salient example is estimation of cross-equation restrictions. For example, consider the example:

$$\begin{aligned} LW69_i &= \alpha_1 + \beta_1 \text{schooling69}_i + \gamma_1 \text{IQ}_i + \delta_1 \text{experience69}_i + \varepsilon_{i1} \\ LW80_i &= \alpha_2 + \beta_2 \text{schooling80}_i + \gamma_2 \text{IQ}_i + \delta_2 \text{experience80}_i + \varepsilon_{i2}. \end{aligned}$$

Here, it would be natural to ask whether the returns to schooling and experience remained the same, i.e. $\beta_1 = \beta_2$ and $\delta_1 = \delta_2$. This could be tested by a linear Wald statistic for the null hypothesis that

$$\begin{pmatrix} 0 & 1 & 0 & 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & -1 \end{pmatrix} (\alpha_1, \beta_1, \gamma_1, \delta_1, \alpha_2, \beta_2, \gamma_2, \delta_2)' = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

4.1.3 When to Estimate Equations Jointly?

Assume we have a number of equations and wonder whether to estimate them jointly or separately. We have seen that separate GMM estimation is just the special case of joint estimation with a weighting

matrix of the block diagonal form

$$\hat{\boldsymbol{W}} = \begin{bmatrix} \hat{\boldsymbol{W}}_{11} & \mathbf{0} \\ \mathbf{0} & \hat{\boldsymbol{W}}_{22} \end{bmatrix}.$$

Asymptotically, this choice of matrix is going to be efficient if either

- the system of equations is just identified (in which case the weighting matrix does not matter) or
- Ω^{-1} is block diagonal as well, i.e.

$$\begin{aligned} \Omega &= \begin{bmatrix} \Omega_{11} & \mathbf{0} \\ \mathbf{0} & \Omega_{22} \end{bmatrix} \\ \Longrightarrow \Omega^{-1} &= \begin{bmatrix} \Omega_{11}^{-1} & \mathbf{0} \\ \mathbf{0} & \Omega_{22}^{-1} \end{bmatrix}. \end{aligned}$$

But Ω has this block diagonal structure iff $\mathbb{E}(e_{im}e_{ih}Z_{im}Z_{ih}) = 0$ for all $m \neq h$, that is, when the equations are unrelated in the sense that the error process from one is uninformative regarding the error process from the other.

Whenever Ω^{-1} is block diagonal, $\hat{\Omega}^{-1}$ and therefore efficient GMM will pick this up in the limit. Thus, one might be tempted to "estimate everything together" in order to extract maximal information. But this sounds a bit like using all instruments one can think of, and it is subject to the same limitations. In addition, if we know that Ω is block diagonal, then imposing this information, and hence estimating the equations separately, will in practice lead to more efficient estimators as joint estimation then just adds (vanishing) noise.

With regard to the risk of misspecification, note in particular that, if $\boldsymbol{W}$ is not block diagonal, misspecification of one equation will generically contaminate the entire estimator. To see this, assume that the M'th moment condition fails:

$$\mathbb{E}(Z_{iM}e_{iM}) \neq 0.$$

Recalling that

$$\hat{\theta}(\hat{\boldsymbol{W}}) - \theta = \left(\hat{\boldsymbol{Q}}_{ZX}'\hat{\boldsymbol{W}}\hat{\boldsymbol{Q}}_{ZX}\right)^{-1}\hat{\boldsymbol{Q}}_{ZX}'\hat{\boldsymbol{W}}\bar{g}_n(\theta)$$

and using standard limit arguments, we find that

$$\hat{\theta}(\hat{\boldsymbol{W}}) - \theta \xrightarrow{p} \left(\boldsymbol{Q}_{ZX}'\boldsymbol{W}\boldsymbol{Q}\right)^{-1}\boldsymbol{Q}_{ZX}'\boldsymbol{W}\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \\ \vdots \\ \mathbb{E}(Z_{iM}e_{iM}) \end{bmatrix}.$$

Unless $\boldsymbol{W}$ is block diagonal, that vector may be nonzero in *all* components.

Nonetheless, it is worth keeping in mind that if equations are informationally related, joint estimation will be more efficient. Importantly, this can be a reason to estimate equations that we are not substantively interested in.

4.2 Special Cases

We next discuss some special cases of Multiple Equation GMM. These are generated by successively imposing stronger and stronger assumptions, thereby recovering numerous estimators that were originally developed outside the GMM framework. (Memorizing the details of this is not essential.)

4.2.1 Conditional Homoskedasticity

We impose:

Assumption 4.7 ***Conditional Homoskedasticity***

$$\mathbb{E}(e_{im}e_{ih}|Z_{im}, Z_{ih}) = \sigma_{mh}.$$

It then follows that

$$\begin{aligned}
\mathbb{E}(e_{im}e_{ih}Z_{im}Z_{ih}') &= \mathbb{E}\left(\mathbb{E}(e_{im}\varepsilon_{ih}Z_{im}Z_{ih}'|Z_{im}, Z_{ih})\right) \\
&= \mathbb{E}\left(\mathbb{E}(e_{im}e_{ih}|Z_{im}, Z_{ih})\, Z_{im}Z_{ih}'\right) \\
&= \mathbb{E}\left(\sigma_{mh}Z_{im}Z_{ih}'\right) \\
&= \sigma_{mh}\mathbb{E}\left(Z_{im}Z_{ih}'\right).
\end{aligned}$$

Hence

$$\Omega = \begin{bmatrix} \sigma_{11}\mathbb{E}\left(Z_{i1}Z_{i1}'\right) & \cdots & \sigma_{1M}\mathbb{E}\left(Z_{i1}Z_{iM}'\right) \\ \vdots & \ddots & \vdots \\ \sigma_{M1}\mathbb{E}\left(Z_{iM}Z_{i1}'\right) & \cdots & \sigma_{MM}\mathbb{E}\left(Z_{iM}Z_{iM}'\right) \end{bmatrix},$$

which can be estimated by

$$\begin{aligned}
\hat{\Omega} &= \begin{bmatrix} \hat{\sigma}_{11}\left(\frac{1}{n}\sum_{i=1}^n Z_{i1}Z_{i1}'\right) & \cdots & \hat{\sigma}_{1M}\left(\frac{1}{n}\sum_{i=1}^n Z_{i1}Z_{iM}'\right) \\ \vdots & \ddots & \vdots \\ \hat{\sigma}_{M1}\left(\frac{1}{n}\sum_{i=1}^n Z_{iM}Z_{i1}'\right) & \cdots & \hat{\sigma}_{MM}\left(\frac{1}{n}\sum_{i=1}^n Z_{iM}Z_{iM}'\right) \end{bmatrix} \\
\hat{\sigma}_{mh} &\equiv \frac{1}{n}\sum_{i=1}^n (Y_{im} - X_{im}'\hat{\theta}_m)(Y_{ih} - X_{ih}'\hat{\theta}_h),
\end{aligned}$$

provided that all $\mathbb{E}(X_{im}X_{ih})$, which are implicitly estimated by $\hat{\sigma}_{mh}$, are finite. The $\hat{\theta}_m$ must come from a first-stage regression that utilizes a consistent estimator; usually, this will be the 2SLS estimator.

Thus we have an efficient GMM estimator $\hat{\theta}\big(\hat{\Omega}^{-1}\big)$. This estimator was previously known as FIVE (full-information instrumental variable efficient) estimator (Brundy and Jorgensen 1971). It can be thought of as the multiple-equation version of 2SLS; equivalently, think of 2SLS as equation-by-equation FIVE. The Sargan(-Hansen) J-statistic will have $\sum_m (l_m - k_m)$ degrees of freedom.

4.2.2 3SLS

On top of the above, assume now that all equations have the same set of instruments:

Assumption 4.8 ***Same Instruments***

$$Z_{im} = Z_i.$$

Then the above expressions have some more specific structure, and the 3SLS (three-stage least squares, Zellner/Theil 1962) estimator emerges.

Define

$$e_i \equiv \begin{bmatrix} e_{i1} \\ \vdots \\ e_{iM} \end{bmatrix}, \Sigma \equiv \mathbb{E}(e_i e_i') \equiv \begin{bmatrix} \sigma_{11} & \cdots & \sigma_{1M} \\ \vdots & \ddots & \vdots \\ \sigma_{M1} & \cdots & \sigma_{MM} \end{bmatrix},$$

then Σ can be consistently estimated by

$$\hat{\Sigma} = \begin{bmatrix} \hat{\sigma}_{11} & \cdots & \hat{\sigma}_{1M} \\ \vdots & \ddots & \vdots \\ \hat{\sigma}_{M1} & \cdots & \hat{\sigma}_{MM} \end{bmatrix} = \frac{1}{n} \sum_{i=1}^{n} \hat{e}_i \hat{e}_i',$$

where the $\hat{e}_i$ come from a first-stage regression with a consistent estimator of θ. Recall that this estimator is usually $\hat{\theta}_{2SLS}$ – this explains why the procedure is called three-stage least squares.

We will now introduce some new notation. Since $Z_{im} = Z_i$ for all m, we have

$$g_i = \begin{bmatrix} Z_{i1} e_{i1} \\ \vdots \\ Z_{iM} e_{iM} \end{bmatrix} = \begin{bmatrix} Z_i e_{i1} \\ \vdots \\ Z_i e_{iM} \end{bmatrix} \equiv e_i \otimes Z_i,$$

where the last equation defines the *Kronecker product* of vectors.

(In general, $\boldsymbol{a} \otimes \boldsymbol{b}$ is the stacked vector $\begin{bmatrix} a_1 \boldsymbol{b} \\ \vdots \\ a_{\#\boldsymbol{a}} \boldsymbol{b} \end{bmatrix}$.)

Similarly, we can write

$$\begin{aligned}\Omega &= \begin{bmatrix} \sigma_{11}\mathbb{E}\left(Z_{i1}Z_{i1}'\right) & \cdots & \sigma_{1M}\mathbb{E}\left(Z_{i1}Z_{iM}'\right) \\ \vdots & \ddots & \vdots \\ \sigma_{M1}\mathbb{E}\left(Z_{iM}Z_{i1}'\right) & \cdots & \sigma_{MM}\mathbb{E}\left(Z_{iM}Z_{iM}'\right) \end{bmatrix} \\ &= \begin{bmatrix} \sigma_{11}\mathbb{E}\left(Z_{i}Z_{i}'\right) & \cdots & \sigma_{1M}\mathbb{E}\left(Z_{i}Z_{i}'\right) \\ \vdots & \ddots & \vdots \\ \sigma_{M1}\mathbb{E}\left(Z_{i}Z_{i}'\right) & \cdots & \sigma_{MM}\mathbb{E}\left(Z_{i}Z_{i}'\right) \end{bmatrix} = \Sigma \otimes \mathbb{E}\left(Z_{i}Z_{i}'\right),\end{aligned}$$

where the last equation defines the Kronecker product of matrices.

(In general,

$$\underset{[M\times N]}{\boldsymbol{A}} \otimes \boldsymbol{B} \equiv \begin{bmatrix} A_{11}\boldsymbol{B} & \cdots & A_{1N}\boldsymbol{B} \\ \vdots & \ddots & \vdots \\ A_{M1}\boldsymbol{B} & \cdots & A_{MN}\boldsymbol{B} \end{bmatrix}.)$$

A useful fact about Kronecker products is that $(\boldsymbol{A}\otimes\boldsymbol{B})^{-1} = \boldsymbol{A}^{-1}\otimes\boldsymbol{B}^{-1}$, hence

$$\Omega^{-1} = \Sigma^{-1}\otimes\left(\mathbb{E}Z_iZ_i'\right)^{-1},$$

and this can be estimated by

$$\hat{\boldsymbol{W}} \equiv \hat{\Omega}^{-1} = \hat{\Sigma}^{-1}\otimes\Big(\frac{1}{n}\sum_{i=1}^{n} Z_iZ_i'\Big)^{-1}.$$

To keep things manageable from here, we again set $M = 2$. Recall also that $\hat{\boldsymbol{W}}_{mh} = \hat{\omega}_{mh}\times\left(\frac{1}{n}\sum_{i=1}^{n} Z_iZ_i'\right)^{-1}$. (Here, $\hat{\omega}_{mh}$ is the (m,h)-cell of $\hat{\Sigma}^{-1}$, which does not in general equal $\hat{\sigma}_{mh}^{-1}$.) Thus

$$\begin{aligned}\hat{\theta}(\hat{\boldsymbol{W}}) &= \begin{bmatrix} \left(\frac{1}{n}\sum_{i=1}^{n} X_{i1}Z_{i1}'\right)\hat{\boldsymbol{W}}_{11}\left(\frac{1}{n}\sum_{i=1}^{n} Z_{i1}X_{i1}'\right) & \left(\frac{1}{n}\sum_{i=1}^{n} X_{i1}Z_{i1}'\right)\hat{\boldsymbol{W}}_{12}\left(\frac{1}{n}\sum_{i=1}^{n} Z_{i2}X_{i2}'\right) \\ \left(\frac{1}{n}\sum_{i=1}^{n} X_{i2}Z_{i2}'\right)\hat{\boldsymbol{W}}_{21}\left(\frac{1}{n}\sum_{i=1}^{n} Z_{i1}X_{i1}'\right) & \left(\frac{1}{n}\sum_{i=1}^{n} X_{i2}Z_{i2}'\right)\hat{\boldsymbol{W}}_{22}\left(\frac{1}{n}\sum_{i=1}^{n} Z_{i2}X_{i2}'\right) \end{bmatrix}^{-1} \\ &\quad \cdot \begin{bmatrix} \left(\frac{1}{n}\sum_{i=1}^{n} X_{i1}Z_{i1}'\right)\hat{\boldsymbol{W}}_{11}\left(\frac{1}{n}\sum_{i=1}^{n} Z_{i1}Y_{i1}\right) + \left(\frac{1}{n}\sum_{i=1}^{n} X_{i1}Z_{i1}'\right)\hat{\boldsymbol{W}}_{12}\left(\frac{1}{n}\sum_{i=1}^{n} Z_{i2}Y_{i2}\right) \\ \left(\frac{1}{n}\sum_{i=1}^{n} X_{i2}Z_{i2}'\right)\hat{\boldsymbol{W}}_{21}\left(\frac{1}{n}\sum_{i=1}^{n} Z_{i1}Y_{i1}\right) + \left(\frac{1}{n}\sum_{i=1}^{n} X_{i2}Z_{i2}'\right)\hat{\boldsymbol{W}}_{22}\left(\frac{1}{n}\sum_{i=1}^{n} Z_{i2}Y_{i2}\right) \end{bmatrix} \\ &= \begin{bmatrix} \hat{\omega}_{11}\hat{\boldsymbol{A}}_{11} & \hat{\omega}_{12}\hat{\boldsymbol{A}}_{12} \\ \hat{\omega}_{21}\hat{\boldsymbol{A}}_{21} & \hat{\omega}_{22}\hat{\boldsymbol{A}}_{22} \end{bmatrix}^{-1} \begin{bmatrix} \hat{\omega}_{11}\hat{\boldsymbol{c}}_{11} + \hat{\omega}_{12}\hat{\boldsymbol{c}}_{12} \\ \hat{\omega}_{21}\hat{\boldsymbol{c}}_{21} + \hat{\omega}_{22}\hat{\boldsymbol{c}}_{22} \end{bmatrix},\end{aligned}$$

where

$$\begin{aligned}\hat{\boldsymbol{A}}_{mh} &\equiv \Big(\frac{1}{n}\sum_{i=1}^{n} X_{im}Z_i'\Big)\Big(\frac{1}{n}\sum_{i=1}^{n} Z_iZ_i'\Big)^{-1}\Big(\frac{1}{n}\sum_{i=1}^{n} Z_iX_{ih}'\Big) \\ \hat{\boldsymbol{c}}_{mh} &\equiv \Big(\frac{1}{n}\sum_{i=1}^{n} X_{im}Z_i'\Big)\Big(\frac{1}{n}\sum_{i=1}^{n} Z_iZ_i'\Big)^{-1}\Big(\frac{1}{n}\sum_{i=1}^{n} Z_iY_{ih}\Big).\end{aligned}$$

As usual, the asymptotic variance can be estimated by the "denominator":

$$\hat{\boldsymbol{V}} \equiv \begin{bmatrix} \hat{\omega}_{11}\hat{\boldsymbol{A}}_{11} & \hat{\omega}_{12}\hat{\boldsymbol{A}}_{12} \\ \hat{\omega}_{21}\hat{\boldsymbol{A}}_{21} & \hat{\omega}_{22}\hat{\boldsymbol{A}}_{22} \end{bmatrix}^{-1},$$

which here estimates

$$\begin{aligned} \operatorname{Avar}\left(\hat{\theta}(\hat{\boldsymbol{W}})\right) &= \begin{bmatrix} \omega_{11}\boldsymbol{A}_{11} & \omega_{12}\boldsymbol{A}_{12} \\ \omega_{21}\boldsymbol{A}_{21} & \omega_{22}\boldsymbol{A}_{22} \end{bmatrix}^{-1} \\ \boldsymbol{A}_{mh} &\equiv \mathbb{E}\left(X_{im}Z_i'\right)\left(\mathbb{E}\left(Z_iZ_i'\right)\right)^{-1}\mathbb{E}\left(Z_iX_{ih}'\right). \end{aligned}$$

The J-statistic will have $(Ml - \sum_m k_m)$ degrees of freedom.

4.2.3 SUR

In a third step, tighten Assumption **??** to:

Assumption 4.9 ***Cross-Equation Predetermined Regressors***

$$Z_i = \cup\{X_{i1}, \ldots, X_{iM}\}.$$

That is, the instruments are the union of the regressors. Intuitively, this means that the endogeneity problem which usually motivates instrumental variables is not present at all. Quite to the contrary, regressors are not only predetermined within their equations but also across equations, i.e. we also presume

$$\mathbb{E}(X_{im}e_{ih}) = \mathbf{0}$$

for $m \neq h$.

We thus have more exclusion restrictions than usual OLS! These restrictions can, of course, be exploited – they imply that we gain efficiency from joint estimation of the equations. This technique is called "seemingly unrelated regressions" (Zellner 1962). We here see that SUR can be interpreted as IV technique, with cross-equation restrictions turning regressors from equation m into overidentifying instruments in equation h.

Let's reconsider the example of a wage equation with auxiliary, seemingly unrelated equation as inspired by Griliches. To repeat, we have:

$$\begin{aligned} LW69_i &= \alpha_1 + \beta_1 \text{schooling69}_i + \gamma_1 \text{IQ}_i + \delta_1 \text{experience69}_i + \varepsilon_{i1} \\ KWW_i &= \alpha_2 + \beta_2 \text{schooling69}_i + \gamma_2 \text{IQ}_i + \varepsilon_{i2}. \end{aligned}$$

For an OLS specification, one would assume that

$$\mathbb{E}\begin{bmatrix} 1 \\ \text{schooling69}_i \\ \text{IQ}_i \\ \text{experience69}_i \end{bmatrix} \varepsilon_{i1} = \mathbf{0}, \mathbb{E}\begin{bmatrix} 1 \\ \text{schooling69}_i \\ \text{IQ}_i \end{bmatrix} \varepsilon_{i2} = \mathbf{0}.$$

For a SUR specification, the assumption is that

$$\mathbb{E}\begin{bmatrix} 1 \\ \text{schooling69}_i \\ \text{IQ}_i \\ \text{experience69}_i \end{bmatrix} \varepsilon_{i1} = \mathbb{E}\begin{bmatrix} 1 \\ \text{schooling69}_i \\ \text{IQ}_i \\ \text{experience69}_i \end{bmatrix} \varepsilon_{i2} = \mathbf{0},$$

which is obviously stronger.

With regard to closed-form expressions, the SUR assumptions allow for drastic simplification. Fix any m and suppose that Z_i is ordered s.t. X_{im} equals its first k_m elements. Let $\boldsymbol{D}$ denote the first k_m columns of $\boldsymbol{I}_l$, then $X_{im} = \boldsymbol{D}'Z_i$. Substituting for this yields

$$\begin{aligned} \mathbb{E}(X_{im}Z_i') &= \boldsymbol{D}'\mathbb{E}(Z_iZ_i') \\ \boldsymbol{D}'\mathbb{E}(Z_iX_{ih}') &= \mathbb{E}(X_{im}X_{ih}'). \end{aligned}$$

Now we can write

$$\begin{aligned} \boldsymbol{A}_{mh} &\equiv \mathbb{E}\left(X_{im}Z_i'\right)\left(\mathbb{E}\left(Z_iZ_i'\right)\right)^{-1}\mathbb{E}\left(Z_iX_{ih}'\right) \\ &= \boldsymbol{D}'\mathbb{E}(Z_iZ_i')\left(\mathbb{E}\left(Z_iZ_i'\right)\right)^{-1}\mathbb{E}\left(Z_iX_{ih}'\right) \\ &= \boldsymbol{D}'\mathbb{E}\left(Z_iX_{ih}'\right) \\ &= \mathbb{E}(X_{im}X_{ih}') \end{aligned}$$

and similarly

$$\begin{aligned} \hat{\boldsymbol{A}}_{mh} &= \frac{1}{n}\sum_{i=1}^{n} X_{im}X_{ih}' \\ \hat{\boldsymbol{c}}_{mh} &= \frac{1}{n}\sum_{i=1}^{n} X_{im}Y_{ih}. \end{aligned}$$

As a result, the SUR estimator (for $M = 2$) is characterized as follows:

$$\begin{aligned}
\hat{\theta}_{SUR} &= \begin{bmatrix} \hat{\omega}_{11}\sum_{i=1}^n X_{i1}X_{i1}' & \hat{\omega}_{12}\sum_{i=1}^n X_{i1}X_{i2}' \\ \hat{\omega}_{21}\sum_{i=1}^n X_{i2}X_{i1}' & \hat{\omega}_{22}\sum_{i=1}^n X_{i2}X_{i2}' \end{bmatrix}^{-1} \begin{bmatrix} \hat{\omega}_{11}\sum_{i=1}^n X_{i1}Y_{i1} + \hat{\omega}_{12}\sum_{i=1}^n X_{i1}Y_{i2} \\ \hat{\omega}_{21}\sum_{i=1}^n X_{i2}Y_{i1} + \hat{\omega}_{22}\sum_{i=1}^n X_{i2}Y_{i2} \end{bmatrix} \\
\text{Avar}\left(\hat{\theta}_{SUR}\right) &= \begin{bmatrix} \omega_{11}\mathbb{E}(X_{i1}X_{i1}') & \omega_{12}\mathbb{E}(X_{i1}X_{i2}') \\ \omega_{21}\mathbb{E}(X_{i2}X_{i1}') & \omega_{22}\mathbb{E}(X_{i2}X_{i2}') \end{bmatrix}^{-1} \\
\hat{\boldsymbol{V}} &= \begin{bmatrix} \hat{\omega}_{11}\frac{1}{n}\sum_{i=1}^n X_{i1}X_{i1}' & \hat{\omega}_{12}\frac{1}{n}\sum_{i=1}^n X_{i1}X_{i2}' \\ \hat{\omega}_{21}\frac{1}{n}\sum_{i=1}^n X_{i2}X_{i1}' & \hat{\omega}_{22}\frac{1}{n}\sum_{i=1}^n X_{i2}X_{i2}' \end{bmatrix}^{-1}.
\end{aligned}$$

The J-statistic will have $(Ml - \sum_m k_m)$ degrees of freedom.

4.2.4 SUR vs. OLS

Textbook discussions of SUR vs. OLS (e.g. p. 508-9 in Davidson and MacKinnon's book) typically resemble the remarks on multiple-equation vs. single-equation GMM given above. We now see why this is the case: Multiple-equation GMM turns into SUR under roughly the same assumptions that turn single-equation GMM into OLS. Hence, the two comparisons are analog and the remarks translate immediately – compare, discussion in a non-GMM framework. Due to their importance, let's repeat the remarks.

First, SUR and OLS will coincide whenever all equations have the same regressors. This special case of SUR is also called *multivariate regression*; it simply means to simultaneously regress different outcomes on the same regressors by separate OLS.

On the other hand, if at least one equation is overidentified, then SUR extracts more information iff error terms are correlated across equations. If they aren't, then SUR will pick this up and no harm is done in the limit, but as before, one should not naively take this fact as guide for dealing with finite samples. In practice, the overidentification would have to be traded off, on a case-by-case basis, against increased noise and increased dangers of misspecification

4.3 Common Coefficients

While the preceding specializations were close analogs to the specializations of single-equation GMM, we will now look at a specialization that makes sense only for multiple-equation GMM: We impose that the coefficients are the same across equations, i.e. $\theta_m = \theta$. In particular, we restate the linearity assumption as:

Assumption 4.10 ***Linearity with Common Coefficients***

$$Y_{im} = X_{im}'\theta + e_{im}.$$

The only expression that is much affected by this change is $\boldsymbol{Q}_{ZX}$. To see this, note that

$$\begin{aligned} g(W_i,\theta) &\equiv \begin{bmatrix} Z_{i1}\cdot(Y_{i1}-X_{i1}'\theta) \\ \vdots \\ Z_{iM}\cdot(Y_{iM}-X_{iM}'\theta) \end{bmatrix} \\ \Longrightarrow \mathbb{E}g(W_i,\theta) &= \mathbb{E}\begin{bmatrix} Z_{i1}Y_{i1} \\ \vdots \\ Z_{iM}Y_{iM} \end{bmatrix} - \mathbb{E}\begin{bmatrix} Z_{i1}X_{i1}' \\ \vdots \\ Z_{iM}X_{iM}' \end{bmatrix}\theta \\ &\equiv \boldsymbol{Q}_{ZY}-\boldsymbol{Q}_{ZX}\theta, \end{aligned}$$

where the last equation defines terms as before. The difference is that $\boldsymbol{Q}_{ZX}$ is now stacked rather than block diagonal.

This affects our rank condition, i.e. Assumption **??**. Recall it stated that

> $\boldsymbol{Q}_{ZX}$ is of full column rank.
>
> (Equivalently, $\mathbb{E}\left(Z_{im}X_{im}'\right)$ is of full column rank for each m.)

We continue to impose the first of these statements (of course, it is really a different statement now because $\boldsymbol{Q}_{ZX}$ has changed). But this statement is *not* any more equivalent to the second one! The second statement is stronger. In fact, it is easy to see that the first statement is implied whenever $\mathbb{E}\left(Z_{im}X_{im}'\right)$ has full column rank for *some* m. Then θ_m is identified, but $\theta=\theta_m$ by assumption 1". But our assumption is even weaker than that because it is also possible for identification to arise from the combination of equations, without any one of them being individually identified.

By identifying Assumption **??** with the first statement, we therefore substantially weaken it. Of course, there is no free lunch: This merely reflects that Assumption **??** became much stronger.

The sample analogs of $\boldsymbol{Q}_{ZY}$ respectively $\boldsymbol{Q}_{ZX}$ are

$$\hat{\boldsymbol{Q}}_{ZY} \equiv \begin{bmatrix} \frac{1}{n}\sum_{i=1}^n Z_{i1}Y_{i1} \\ \vdots \\ \frac{1}{n}\sum_{i=1}^n Z_{iM}Y_{iM} \end{bmatrix}, \hat{\boldsymbol{Q}}_{ZX} \equiv \begin{bmatrix} \frac{1}{n}\sum_{i=1}^n Z_{i1}X_{i1}' \\ \vdots \\ \frac{1}{n}\sum_{i=1}^n Z_{iM}X_{iM}' \end{bmatrix},$$

where again, the difference is that $\hat{\boldsymbol{Q}}_{ZX}$ is now stacked. To get the estimator and its estimated variance, we again just plug these definitions into previous results. As before, it is instructive to write

the estimator out in detail. Set $M = 2$ and write:

$$\begin{aligned}
&\hat{\theta}(\hat{\boldsymbol{W}}) = \left(\hat{\boldsymbol{Q}}_{ZX}'\hat{\boldsymbol{W}}\hat{\boldsymbol{Q}}_{ZX}\right)^{-1}\hat{\boldsymbol{Q}}_{ZX}'\hat{\boldsymbol{W}}\hat{\boldsymbol{Q}}_{ZY} \\
= &\left(\begin{bmatrix} \frac{1}{n}\sum_{i=1}^n Z_{i1}X_{i1}' \\ \frac{1}{n}\sum_{i=1}^n Z_{i2}X_{i2}' \end{bmatrix}' \begin{bmatrix} \hat{\boldsymbol{W}}_{11} & \hat{\boldsymbol{W}}_{12} \\ \hat{\boldsymbol{W}}_{21} & \hat{\boldsymbol{W}}_{22} \end{bmatrix} \begin{bmatrix} \frac{1}{n}\sum_{i=1}^n Z_{i1}X_{i1}' \\ \frac{1}{n}\sum_{i=1}^n Z_{i2}X_{i2}' \end{bmatrix}\right)^{-1} \\
&\cdot \begin{bmatrix} \frac{1}{n}\sum_{i=1}^n Z_{i1}X_{i1}' \\ \frac{1}{n}\sum_{i=1}^n Z_{i2}X_{i2}' \end{bmatrix}' \begin{bmatrix} \hat{\boldsymbol{W}}_{11} & \hat{\boldsymbol{W}}_{12} \\ \hat{\boldsymbol{W}}_{21} & \hat{\boldsymbol{W}}_{22} \end{bmatrix} \begin{bmatrix} \frac{1}{n}\sum_{i=1}^n Z_{i1}Y_{i1} \\ \frac{1}{n}\sum_{i=1}^n Z_{i2}Y_{i2} \end{bmatrix} \\
= &\left(\sum_{m=1}^2\sum_{h=1}^2 \Big(\frac{1}{n}\sum_{i=1}^n X_{im}Z_{im}'\Big)\hat{\boldsymbol{W}}_{mh}\Big(\frac{1}{n}\sum_{i=1}^n Z_{ih}X_{ih}'\Big)\right)^{-1} \cdot \sum_{m=1}^2\sum_{h=1}^2 \Big(\frac{1}{n}\sum_{i=1}^n X_{im}Z_{im}'\Big)\hat{\boldsymbol{W}}_{mh}\Big(\frac{1}{n}\sum_{i=1}^n Z_{ih}Y_{ih}\Big).
\end{aligned}$$

We can now go through the same specializations as before. We will do this in extremely abbreviated form. Imposing conditional homoskedasticity leads to a common coefficients version of $\hat{\theta}_{FIVE}$. Next, assume in addition that all equations have the same instruments, i.e. $Z_{im} = Z_i$. Then the efficient weighting matrix is again a Kronecker product:

$$\hat{\boldsymbol{W}} = \hat{\Omega}^{-1} \equiv \hat{\Sigma}^{-1} \otimes \Big(\frac{1}{n}\sum_{i=1}^n Z_i Z_i'\Big)^{-1}.$$

The efficient estimator becomes

$$\begin{aligned}
\hat{\theta}(\hat{\Omega}^{-1}) = &\left(\sum_{m=1}^2\sum_{h=1}^2 \hat{\omega}_{mh}\Big(\frac{1}{n}\sum_{i=1}^n X_{im}Z_i'\Big)\Big(\frac{1}{n}\sum_{i=1}^n Z_iZ_i'\Big)^{-1}\Big(\frac{1}{n}\sum_{i=1}^n Z_iX_{ih}'\Big)\right)^{-1} \\
&\times \sum_{m=1}^2\sum_{h=1}^2 \hat{\omega}_{mh}\Big(\frac{1}{n}\sum_{i=1}^n X_{im}Z_i'\Big)\Big(\frac{1}{n}\sum_{i=1}^n Z_iZ_i'\Big)^{-1}\Big(\frac{1}{n}\sum_{i=1}^n Z_iY_{ih}\Big).
\end{aligned}$$

This is the common coefficients version of the 3SLS estimator. Here, $\hat{\omega}_{mh}$ is the relevant cell of the matrix $\hat{\Sigma}^{-1}$; note this does not in general equal $\hat{\sigma}_{mh}^{-1}$.

What we're after is the next step: Impose the SUR assumption, namely that $Z_i = \cup\{X_{i1}, \ldots, X_{iM}\}$, implying that $\mathbb{E}(X_{im}e_{ih}) = \mathbf{0}$ for $m \neq h$. As before, this means that the instruments vanish from the expressions, and we get

$$\hat{\theta}_{RE} \equiv \left(\sum_{m=1}^2\sum_{h=1}^2 \hat{\omega}_{mh}\Big(\frac{1}{n}\sum_{i=1}^n X_{im}X_{ih}'\Big)\right)^{-1} \sum_{m=1}^2\sum_{h=1}^2 \hat{\omega}_{mh}\Big(\frac{1}{n}\sum_{i=1}^n X_{im}Y_{ih}\Big)$$

with asymptotic variance

$$\mathrm{Avar}(\hat{\theta}_{RE}) = \Big(\sum_{m=1}^2\sum_{h=1}^2 \omega_{mh}\mathbb{E}\left(X_{im}X_{ih}'\right)\Big)^{-1}$$

estimated by

$$\hat{\boldsymbol{V}} \equiv \left(\sum_{m=1}^2\sum_{h=1}^2 \hat{\omega}_{mh}\Big(\frac{1}{n}\sum_{i=1}^n X_{im}X_{ih}'\Big)\right)^{-1},$$

where ω_{mh} is the population analog of $\hat{\omega}_{mh}$.

This is once again a historically well-known estimator, namely the *random effects estimator* for panel data. We will revisit $\hat{\theta}_{RE}$ in that context. The J-statistic will have $(Ml-k)$ degrees of freedom.

We finally derive the *Pooled OLS* estimator by choosing the weighting matrix

$$\boldsymbol{I}_M \otimes \Big(\frac{1}{n}\sum_{i=1}^{n} Z_i Z_i'\Big)^{-1}.$$

Substituting for that matrix in previous expressions, we find that $\hat{\theta}_{RE}$ turns into

$$\begin{aligned}\hat{\theta}_{pooled} &\equiv \Big(\sum_{m=1}^{2}\sum_{h=1}^{2}\mathbf{1}\{m=h\}\Big(\frac{1}{n}\sum_{i=1}^{n}X_{im}X_{ih}'\Big)\Big)^{-1}\sum_{m=1}^{2}\sum_{h=1}^{2}\mathbf{1}\{m=h\}\Big(\frac{1}{n}\sum_{i=1}^{n}X_{im}Y_{ih}\Big)\\ &= \Big(\sum_{m=1}^{2}\sum_{i=1}^{n}X_{im}X_{im}'\Big)^{-1}\sum_{m=1}^{2}\sum_{i=1}^{n}X_{im}Y_{im}.\end{aligned}$$

This expression makes clear where the estimator's name comes from: $\hat{\theta}_{pooled}$ just operates OLS on all observations simultaneously. (It may be more recognizable as $(\boldsymbol{X'X})^{-1}\boldsymbol{X'Y}$.)

Note that this estimator exploits within-equation orthogonalities but not cross-equation ones. This renders it robust to failures of the latter, but also inefficient if they hold. Also, implementing the estimator by hitting the OLS button in canned packages would give incorrect standard errors because cross-equation correlation of X_i will be ignored. Correct standard errors are easily obtained by using a pooled OLS command if it exists or by directly computing

$$\hat{\boldsymbol{V}} = \Big(\sum_{m=1}^{2}\Big(\frac{1}{n}\sum_{i=1}^{n}X_{im}X_{im}'\Big)\Big)^{-1}\Big(\sum_{m=1}^{2}\sum_{h=1}^{2}\frac{\hat{\sigma}_{mh}^2}{n}\sum_{i=1}^{n}X_{im}X_{ih}'\Big)\Big(\sum_{m=1}^{2}\Big(\frac{1}{n}\sum_{i=1}^{n}X_{im}X_{im}'\Big)\Big)^{-1}.$$

5 Panel Data

From the common coefficients model, it is only a small step to panel data – at least to so-called "short panels," where the number of equations M (usually the number of time periods or panel waves T) is small relative to the number of observations in each group (time period). Short panels are analyzed by asymptotics where one lets $n \to \infty$ while M is constant, i.e. the asymptotics we considered so far.

Long panels, where M is large and n is small in comparison (e.g. a few industry or country time series), should not be analyzed in this framework. They are best seen as multiple time series, and we will not cover them in this lecture. Asymptotics where both M and n go to infinity are an active field of research.

We will now briefly cast some classic panel data estimators as GMM. The treatment will be very abridged and the matrix algebra is for your reference only.

5.1 The Random Effects Estimator

Begin by imposing common coefficients as well as the SUR assumptions:

Assumption 5.1 ***Linearity***

$$\boldsymbol{Y}_i = \boldsymbol{X}_i\theta_0 + \boldsymbol{e}_i.$$

Assumption 5.2 ***IID***

$$(\boldsymbol{Y}_i, \boldsymbol{X}_i) \text{ is i.i.d.}$$

Assumption 5.3 ***Moment Conditions***

$$\begin{aligned} \mathbb{E}(X_{im}e_{ih}) &= \mathbf{0}, \forall m, h \\ \iff \mathbb{E}(\boldsymbol{e}_i \otimes Z_i) &= \mathbf{0} \end{aligned}$$

Assumption 5.4 ***Identification***

$$\mathbb{E}(\boldsymbol{X}_i \otimes Z_i) \text{ is of full column rank.}$$

Assumption 5.5 ***Conditional Homoskedasticity***

$$\mathbb{E}(\boldsymbol{e}_i\boldsymbol{e}_i'|Z_i) = \mathbb{E}(\boldsymbol{e}_i\boldsymbol{e}_i') \equiv \Sigma.$$

Assumption 5.6 ***Nonsingularity***

$$\mathbb{E}(g_i g_i') \equiv \mathbb{E}((\boldsymbol{e}_i \otimes Z_i)(\boldsymbol{e}_i \otimes Z_i)') \text{ is nonsingular.}$$

We thus start with a rather restricted model, which happens to be the precise model on which the last section ended, i.e. SUR with common coefficients. Thus, the efficient GMM estimator is

$$\hat{\theta}_{RE} \equiv \Big(\sum_{m=1}^{M} \sum_{h=1}^{M} \hat{\omega}_{mh} \Big(\frac{1}{n} \sum_{i=1}^{n} X_{im} X_{ih}' \Big) \Big)^{-1} \sum_{m=1}^{M} \sum_{h=1}^{M} \hat{\omega}_{mh} \Big(\frac{1}{n} \sum_{i=1}^{n} X_{im} Y_{ih} \Big),$$

the *Random Effects Estimator*.

However, the orthogonality restrictions exploited by this estimator appear very strong in a panel context. To see this, consider the usual panel equation:

$$Y_{im} = X_{im}' \theta + \alpha_i + \eta_{im}.$$

In this equation, $X_{im}' \theta$ is to be interpreted as before, η_i is the error term, and α_i is called *individual effect* or *fixed effect*. We will normally stack equations for one observational unit and write

$$\boldsymbol{Y}_i = \boldsymbol{X}_i \theta + \mathbf{1}_M \cdot \alpha_i + \boldsymbol{\eta}_i.$$

What do we need to impose to generate the above model?

Consider the assumption that α_i is uncorrelated with X_i:

$$\mathbb{E}(X_{im} \alpha_i) = \mathbf{0}.$$

If we additionally assume the usual predeterminedness,

$$\mathbb{E}(X_{im} \eta_{ih}) = \mathbf{0}, \forall m, h,$$

then we can define

$$e_{im} \equiv \alpha_i + \eta_{im}$$

and conclude that

$$\mathbb{E}(X_{im} e_{ih}) = \mathbf{0}, \forall m, h,$$

i.e. that Assumption **??** holds.

We will not worry too much about exogeneity of X_{im} here, but we notice that $\mathbb{E}(X_{im} \alpha_i) = \mathbf{0}$ is a very strong assumption. It means that the covariates of observational unit i cannot be correlated with unobserved latent properties (traits, abilities, endowments...) of that unit. To further discuss this, consider a variation on the wage equations example:

$$\begin{aligned} LW69_i &= \alpha_1 + \beta \text{schooling69}_i + \gamma \text{IQ}_i + \alpha_i + \eta_{i1} \\ LW80_i &= \alpha_2 + \beta \text{schooling80}_i + \gamma \text{IQ}_i + \alpha_i + \eta_{i2} \\ LW82_i &= \alpha_3 + \beta \text{schooling82}_i + \gamma \text{IQ}_i + \alpha_i + \eta_{i3}. \end{aligned}$$

Here, we have another year and imposed that the returns to schooling and IQ are time invariant. This can be rewritten to feature common coefficients, so that it technically fits into our model.

$$\begin{bmatrix} LW69_i \\ LW80_i \\ LW82_i \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & \text{schooling69}_i & \text{IQ}_i \\ 0 & 1 & 0 & \text{schooling80}_i & \text{IQ}_i \\ 0 & 0 & 1 & \text{schooling82}_i & \text{IQ}_i \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \beta \\ \gamma \end{bmatrix} + \begin{bmatrix} \alpha_i + \eta_{i1} \\ \alpha_i + \eta_{i2} \\ \alpha_i + \eta_{i3} \end{bmatrix}.$$

So $\hat{\theta}_{RE}$ can be computed. Furthermore, the assumption that $\boldsymbol{\eta}_i$ is orthogonal to regressors may be plausible (more so than in a cross-sectional context, in fact, because of the presence of α_i). But consider α_i, the individual fixed effect on wages. What does this capture? Would we want to assume that it is uncorrelated with school achievement and measured IQ?

5.2 The Fixed Effects Estimator

Fixed-effect estimation does not rely on $\mathbb{E}(X_{im}\alpha_i) = \mathbf{0}$. The idea is to bring all equations into deviation form, i.e. replace Y_{im} by $Y_{im} - \overline{Y}_i$ etc., where $\overline{Y}_i \equiv \frac{1}{M}\sum_{m=1}^{M} Y_{im}$. Intuitively, consistency is achieved because α_i vanishes under this transformation. But there is a price: Any within–group invariant (e.g. time invariant) regressor will vanish from the equation as well. Thus, if our original estimating equation was

$$Y_{im} = \alpha_i + \beta_i Z_i + \gamma_i X_{im} + \eta_{im},$$

where the notation indicates that Z_i is constant within the observational unit, then we will not recover an estimate of β_i.

(Indeed, without assuming that $\mathbb{E}(Z_i\alpha_i) = \mathbf{0}$, we have an identification problem. Why?)

We will not algebraically develop this estimator in class, but for your reference, here is a derivation that follows Hayashi. Define

$$\begin{aligned} \boldsymbol{Q} &\equiv \boldsymbol{I}_M - \mathbf{1}_M(\mathbf{1}_M'\mathbf{1}_M)^{-1}\mathbf{1}_M' \\ &= \boldsymbol{I}_M - \frac{\mathbf{1}_M\mathbf{1}_M'}{M} \\ &= \boldsymbol{I}_M - \begin{bmatrix} 1/M & \cdots & 1/M \\ \vdots & \ddots & \vdots \\ 1/M & \cdots & 1/M \end{bmatrix}. \end{aligned}$$

This matrix is the annihilator associated with a vector of ones, thus it extracts residuals from regression

on a constant, or in other words: deviations from group means. For example,

$$\tilde{\boldsymbol{Y}}_i \equiv \boldsymbol{Q}\boldsymbol{Y}_i = \begin{bmatrix} Y_{i1} - \overline{y}_i \\ \vdots \\ Y_{iM} - \overline{y}_i \end{bmatrix} = \boldsymbol{Y}_i - \mathbf{1}_M \overline{Y}_i.$$

To see the identification problem, partition the regressor matrix into

$$\boldsymbol{X}_i = \begin{bmatrix} \boldsymbol{F}_i & \mathbf{1}_M \boldsymbol{b}_i' \end{bmatrix},$$

where $\boldsymbol{b}_i$ are the common (within-group invariant) estimators, and

$$\theta = \begin{bmatrix} \beta \\ \gamma \end{bmatrix}$$

in such a manner that the model can be written as

$$\boldsymbol{Y}_i = \boldsymbol{F}_i \beta + \mathbf{1}_M \boldsymbol{b}_i' \gamma + \mathbf{1}_M \alpha_i + \boldsymbol{\eta}_i,$$

then it will turn out that we cannot estimate γ. Worse still, we must also be careful about $\boldsymbol{F}_i$. In our above example, one might at first write

$$\boldsymbol{F}_i = \begin{bmatrix} 1 & 0 & 0 & \text{schooling69}_i \\ 0 & 1 & 0 & \text{schooling80}_i \\ 0 & 0 & 1 & \text{schooling82}_i \end{bmatrix}, \boldsymbol{b}_i = \begin{bmatrix} \text{IQ}_i \\ \text{IQ}_i \\ \text{IQ}_i \end{bmatrix}.$$

Here, $(\beta_1, \beta_2, \beta_3)$ will not be jointly identified. (Why? A very similar problem exists in the simplest linear regressions with dummies.)

If we specify $\boldsymbol{F}_i$ so that the intercept is not linearly dependent on it, e.g.

$$\boldsymbol{F}_i^* = \begin{bmatrix} 1 & 0 & \text{schooling69}_i \\ 0 & 1 & \text{schooling80}_i \\ 0 & 0 & \text{schooling82}_i \end{bmatrix},$$

then β is identified and can be estimated.

Technically, the identification condition is:

Assumption 5.7 ***Identification for Fixed Effect***

$$\mathbb{E}(\boldsymbol{Q}\boldsymbol{F}_i \otimes Z_i) \text{ is of full column rank,}$$

i.e. $\boldsymbol{F}_i$ *is of full rank after the demeaning operation which removes one degree of freedom.*

Replacing Assumption **??** with **??**, we have a model in which $\hat{\theta}_{RE}$ is inconsistent but in which consistent estimation is possible. Specifically, we can transform the model as follows:

$$\boldsymbol{QY}_i = \boldsymbol{QF}_i\beta + \boldsymbol{Q\eta}_i,$$

which by defining terms we can also write as

$$\tilde{\boldsymbol{Y}}_i = \tilde{\boldsymbol{F}}_i\beta + \tilde{\boldsymbol{\eta}}_i.$$

On top of stacking all equations for the same observation, which is already reflected in the above notation, we will now also stack observations. In other words, we write the entire data as one vector and one matrix

$$(\tilde{\boldsymbol{Y}}, \tilde{\boldsymbol{F}}) = \left(\begin{bmatrix} \tilde{\boldsymbol{Y}}_1 \\ \vdots \\ \tilde{\boldsymbol{Y}}_M \end{bmatrix}, \begin{bmatrix} \tilde{\boldsymbol{F}}_1 \\ \vdots \\ \tilde{\boldsymbol{F}}_M \end{bmatrix} \right).$$

The fixed-effects estimator just performs OLS on this.

$$\begin{aligned}\hat{\beta}_{FE} \equiv (\tilde{\boldsymbol{F}}'\tilde{\boldsymbol{F}})^{-1}\tilde{\boldsymbol{F}}'\tilde{\boldsymbol{Y}} &= \Big(\frac{1}{n}\sum_{i=1}^{n}\tilde{\boldsymbol{F}}_i'\tilde{\boldsymbol{F}}_i\Big)^{-1}\frac{1}{n}\sum_{i=1}^{n}\tilde{\boldsymbol{F}}_i'\tilde{\boldsymbol{Y}}_i = \Big(\frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{QF}_i)'\boldsymbol{QF}_i\Big)^{-1}\frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{QF}_i)'\boldsymbol{QY}_i \\ &= \Big(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{F}_i'\boldsymbol{QQF}_i\Big)^{-1}\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{F}_i'\boldsymbol{QQy}_i = \Big(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{F}_i'\boldsymbol{QF}_i\Big)^{-1}\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{F}_i'\boldsymbol{QY}_i.\end{aligned}$$

In short, $\hat{\beta}_{FE}$ is pooled OLS applied to the demeaned data.

Notice, however, that we have not explicitly estimated the moment conditions' variance in a first step, so we are implying some data-independent weighting matrix here. Furthermore, cranking out the estimator's asymptotic distribution yields

$$\text{Avar}(\hat{\beta}_{FE}) = \big(\mathbb{E}(\tilde{\boldsymbol{F}}_i'\tilde{\boldsymbol{F}}_i)\big)^{-1}\mathbb{E}(\tilde{\boldsymbol{F}}_i'\mathbb{E}(\tilde{\boldsymbol{\eta}}_i\tilde{\boldsymbol{\eta}}_i')\tilde{\boldsymbol{F}}_i)\big(\mathbb{E}(\tilde{\boldsymbol{F}}_i'\tilde{\boldsymbol{F}}_i)\big)^{-1}.$$

This expression looks suspicious because its sandwich form resembles the asymptotic variance of non-optimized GMM. Is $\hat{\beta}_{FE}$ inefficient? If yes, what weighting matrix does it use? Hence, what additional assumption would render it efficient?

The weighting matrix implicitly used here is the identity matrix. This is efficient iff

$$\mathbb{E}(\tilde{\boldsymbol{\eta}}_i\tilde{\boldsymbol{\eta}}_i') = \sigma_\eta^2\boldsymbol{I}_M$$

(the σ_η^2 term will cancel out). Substantively, this is to say that the components of $\tilde{\boldsymbol{\eta}}_{ii}$, $\{\tilde{\boldsymbol{\eta}}_{ii1}, \ldots, \tilde{\boldsymbol{\eta}}_{iiM}\}$, are uncorrelated and have the same variance. Errors may not be persistent, nor heteroskedastic, within an observational unit. (By assumption, they are i.i.d. across units anyway.) If this assumption is imposed, then $\hat{\beta}_{FE}$ is efficient, and if we really believe the assumption, we *should* impose it because

it will improve finite-sample performance. Else, one could easily turn this estimator into a two-step estimator, and in modern temrinologym that woud still be called fixed effects estimation.

If canned OLS procedures are used to implement $\hat{\beta}_{FE}$, the standard errors are wrong. The reason is that in demeaning the data, we lose n degrees of freedom by implicitly estimating n within-group means. For panels with moderately many waves, this loss of effectively an entire wave has a sizeable effect on test statistics. Of course, canned FE routines know to avoid the problem.

5.3 Fixed Effects Estimation vs. First Differencing

Whenever the identifying assumptions (moment conditions) of the fixed effects model apply, one could also think of taking first differences, i.e. estimating the equation

$$\Delta Y_{im} = \Delta X_{im}'\theta + \Delta\eta_{im}, m = 2, \ldots, M,$$

where $\Delta Y_{im} = Y_{im} - Y_{i,m-1}$ etc. This transformation also removes the individual fixed effect. The loss of one equation is even more transparent here than with the demeaning procedure.

Hayashi treats this as identical to fixed-effects estimation (see in particular analytical exercise 2 in chapter 5). In contrast, many other texts, e.g. Wooldridge, have separate sections on the two. What is going on here?

Both methods use the same moment conditions, and insofar as the moment conditions define our model, can be thought of as operating within the same model. Indeed, Wooldridge (p. 279) writes about first differences that "we emphasize that the model and the interpretation of $[\theta]$ are *exactly* as" for fixed effects.

However, one would historically identify RE-, FE-, and FD-estimators with one-step estimators that do not attempt to initially estimate Ω (like FE as defined above, unlike RE as defined above). As a result, all of these estimators are, in general, inefficient. Of course, they are efficient if the assumptions they implicitly make about Ω are true. The difference between FE and FD lies in those assumptions. In particular, FD is efficient if

$$\mathbb{E}(\Delta\tilde{\boldsymbol{\eta}}_i \Delta\tilde{\boldsymbol{\eta}}_i') = \sigma_\eta^2 \boldsymbol{I}_M.$$

Undoing the differencing operation, one finds that this holds if $\tilde{\boldsymbol{\eta}}_i$ is a random walk over m. This is an extreme contrast to FE, which is efficient if the error process has zero memory.

If one wants to leave Ω unrestricted, then FE and FD (both expanded by a first-stage regression to estimate Ω^{-1}) use the same moment conditions and are hence the same. This is Hayashi's perspective. By the same token, we read in Wooldridge (p. 285) that if "the variance matrix of $\boldsymbol{\eta}_i$ is unrestricted, it does not matter which transformation we use."

6 Detour: A More Formal Take on Identification

6.1 Identifiability in Different Contexts

As a prelude to extremum estimators, we take a more formal look at identification that connects identification as it is defined for GMM, ML, or indeed any other estimators, likelihood estimators. In particular, the essential requirement of identification is that the true parameter value can in principle be recovered from the data. That is, imagine that you had perfect knowledge of the population distribution of all observable r.v.'s in your model. A parameter θ is *identified* (or identifiable, or point identified in contrast to set identified) if such knowledge would imply knowledge of the true θ_0. This will specifically be true if the true parameter value can be expressed as function of population quantities, e.g. $\beta_0 = \mathbb{E}(X_i X_i')^{-1}\mathbb{E}(X_i Y_i')$ for OLS, assuming (as we did) existence of the inverse. Indeed, this makes clear why the rank condition is an identification condition, and it is immediate that identifiability fails without it because the population moment conditions are then fulfilled by all β in a linear subspace of $\mathbb{R}^k$.

Let's also briefly recall how identification is defined for models where likelihoods

$$f(W_i;\theta)$$

are fully specified. (The notation suggests existence of densities, but this is not essential.) 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 and, therefore, true distribution of the data. I put the word same in scare quotes because we must properly define it. In particular, likelihoods that are formally distinct but that agree with probability 1 do not count as different; intuitively, observations that allow to discriminate between them will not happen. More formally, we have:

Definition 1 *Consider a model that specifies $f(W_i;\theta)$. Then θ_0 is identified within Θ iff for all $\theta \in \Theta$ with $\theta \neq \theta_0$,*

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

or equivalently,

$$\Pr\big(f(W_i;\theta) \neq f(W_i;\theta_0)\big) \neq 0,$$

where the probability is with respect to sampling from the true data generating process.

6.2 A General Definition of Identifiability

We next provide a semi-formal notion of identification that ties all of this together. Specifically, generalize likelihood identification to allow that θ may not be mapped onto a single distribution $f(W_i;\theta)$ of the data, but onto a set of distributions. Thus, our model defines a mapping

$$\theta \mapsto \Gamma(\theta),$$

where $\Gamma(\theta)$ is the set of distributions $f(W_i)$ that are consistent with parameter value θ.[2] For example, in a GMM model we would have $\Gamma(\theta) \equiv \{f(W_i) : \mathbb{E}(W_i, \theta) = \mathbf{0}\}$, which with linear GMM would further specialize to $\{f(W_i) : \boldsymbol{Q}_{ZX}\theta = \boldsymbol{Q}_{ZY}\}$. We then say that θ_0 is identified if

$$\theta \neq \theta_0 \Rightarrow \Gamma(\theta) \cap \Gamma(\theta_0) = \emptyset.$$

While we strictly speaking only care about this condition with regard to the true value θ_0, we obviously do not know that value ex ante, and so we normally call a model identified if the above holds for any $\theta_0 \in \Theta$.[3] In other words, identification requires that the inverse mapping Γ^{-1} is single-valued.

This way of looking at identification allows to make some immediate connections to currently active literatures. For example, Γ^{-1} might be potentially informative in the sense of picking small subsets of Θ even though identification does no obtain. This is called *partial identification* and is the subject of much current research. Indeed, OLS with rank failure is a trivial example since it technically defines a subspace of $\mathbb{R}^k$ to be "true"; however, this identification is typically not interesting because it does not imply bounds on components of β_0. More interesting examples occur in estimation of games that may have multiple equilibria and in reduced-form analysis of missing-data or treatment effect models under weak conditions. For another example, if Γ^{-1} is a function but is ill-behaved in specific ways, e.g. by being not uniformly continuous, then we may have an *ill-posed inverse problem*. We will not further pursue these extensions.

6.3 Some Examples

Let's first relate these ideas back to OLS. Identifying θ only with the parameter of interest β, we can think of $\Gamma(\beta)$ as containg all joint distributions of (Y_i, X_i) that fulfill our regularity conditions (notably, $\mathbb{E}(X_i X_i')$ has full rank) and furthermore have that $\mathbb{E}(Y_i|X_i) = X_i'\beta$. Assume now that you learned such a true joint distribution without explicitly learning β_0. Could you back out β_0? Yes, because $\beta_0 = (\mathbb{E}(X_i X_i'))^{-1}\mathbb{E}(X_i Y_i)$.

[2] This is where the development is semi-formal because we could define Γ in terms of "deeper" quantities. See Arthur Lewbel's "Identification Zoo" survey for a fuller treatment.

[3] Models that may be strongly, weakly, or unidentified depending on the value taken by θ_0 are the subject of current research but beyond the scope of this lecture.

Note that this uses the regularity condition because the inverse is assumed to exist. This use is crucial. Suppose that $\mathbb{E}(X_i X_i')$ does not have full rank. The there might still be a "true" β_0 out there in the sense that Y_i is in fact generated as $X_i'\beta_0 + \varepsilon_i$, but there will be distinct parameter vectors $\tilde{\beta} \neq \beta_0$ s.t. $X_i'\beta_0 = X_i'\tilde{\beta}$ a.s., and learning the population distribution of (Y_i, X_i) will not help us distinguish the two.[4]

For an example that goes beyond this lecture's earlier chapters, consider *nonparametric mean regression*

$$Y_i = m(X_i) + \varepsilon_i$$

with the regularity conditions that X_i is continuous with full support on $\mathbb{R}^k$, that $m : \mathbb{R}^k \to \mathbb{R}$ is continuous, and that ε_i is (mean) independent of X_i. The parameter of interest is the true "value" m_0 of m. We will first show that it is not identified and then that it is identified upon also normalizing $\mathbb{E}\varepsilon_i = 0$. Note the similarity to discussion of these same assumptions on ε_i in OLS. Our analysis will showcase two important proof techniques:

- To show that the true parameter value θ_0 is not identified, show that some distinct parameter value $\tilde{\theta}$ is observationally equivalent, i.e. it generates the same population distribution of observables. Note that in order to do this, it is fair game to freely manipulate "nuisance parameters" like error distributions as long as the assumptions defining one's model are respected.

- To show that the true parameter is identified, show that an example as in the preceding bullet cannot exist. A common way to do this is to assume for sake of argument that the true distribution of observables (but not the underlying structure generating it!) is known and to show how one would back out the true parameter value. Indeed, we did that for OLS above.

Thus, to see the nonidentification part, let F denote the c.d.f. of ε_i and denote the true values of (m, F) by (m_0, F_0). Define

$$\begin{aligned} \tilde{F}(e) &= F_0(e-1) \\ \tilde{m}(X) &= m_0(X) - 1. \end{aligned}$$

In plain English, $m(\cdot)$ decreases by 1, but 1 is also added to ε_i. Then it is intuitively clear that (m_0, F_0)

[4] In fact, there will be infinitely many observationally equivalent parameter vectors. As discussed before, they form the linear subspace delineated by the underdetermined population moment conditions.

and $(\tilde{m}, \tilde{F})$ are observationally equivalent. Formally, show this by writing

$$\begin{aligned}
\Pr\left(Y_i \leq Y | X = X_i\right) &= \Pr\left(m_0(X_i) + \varepsilon_i \leq Y\right) \\
&= \Pr(\varepsilon_i \leq Y - m_0(X_i)) \\
&= F_0(Y - m_0(X_i)) \\
&= F_0(Y - (m_0(X_i) - 1) - 1) \\
&= \tilde{F}(Y - \tilde{m}(X_i)) \\
&= \Pr\left(\tilde{m}(X_i) + (\varepsilon_i + 1) \leq Y\right).
\end{aligned}$$

Assume now the additional condition that $\mathbb{E}(\varepsilon_i | X_i) = 0$. Is the model then identified? Yes. We will prove this carefully following the strategy in the second bullet above. Thus, consider an arbitrary but henceforth fixed $\tilde{\theta} \neq \theta_0$ and conclude that $\Pr(f(W_i; \tilde{\theta}) \neq f(W_i; \theta_0)) > 0$. The detailed argument goes as follows:

Let $\tilde{m} \neq m_0$ and also fix any random variable $\tilde{\varepsilon}_i$ that is consistent with our model assumptions, i.e. $\mathbb{E}(\tilde{\varepsilon}_i | X_i) = 0$. Then there exists X^* s.t. $\tilde{m}\left(X^*\right) \neq m_0\left(X^*\right)$, say $\tilde{m}\left(X^*\right) > m_0\left(X^*\right)$. (The adaptation of the argument to the reverse inequality will be clear.) Since attention is restricted to continuous functions, there exists $\delta > 0$ (think δ "small") s.t. $\tilde{m}\left(\tilde{X}\right) > m_0\left(\tilde{X}\right)$ for any $\tilde{X} \in B(X^*, \delta) = \{X : \|X - X^*\| \leq \delta\}$. It follows that

$$\mathbb{E}(m_0(X_i) + \varepsilon_i | X_i \in B(X^*, \delta)) < \mathbb{E}(\tilde{m}(X_i) + \varepsilon_i | X_i \in B(X^*, \delta)) = \mathbb{E}(\tilde{m}(X_i) + \tilde{\varepsilon}_i | X_i \in B(X^*, \delta)),$$

where the last equality uses $\mathbb{E}(\varepsilon_i | X_i) = \mathbb{E}(\tilde{\varepsilon}_i | X_i) = 0$. But since X_i has full support, $\Pr(X_i \in B(X^*, \delta)) > 0$, so if the two distinct parameter values induced the same distribution of (Y_i, X_i), the above conditional expectations would have to agree. Done!

Of course, the intuition is just that

$$\mathbb{E}(Y_i | X_i) = \mathbb{E}(m_0(X_i) + \varepsilon_i | X_i) = m_0(X_i) + \mathbb{E}(\varepsilon_i | X_i) = m_0(X_i),$$

so that we can back out m_0 from knowledge of the distribution of observables. This intuition is "morally true" and is how I would probably communicate identification in practice here. However, the argument makes sloppy use of probability theory and does not clarify how continuity of m and full support of X_i are used. The subtlety is that a conditional distribution uniquely defines a conditional expectation only almost surely on the support of the conditioning variable. That immediately clarifies how full support of X_i was used. Continuity is used because functions that differ from m_0 but only on a set of measure zero (and that therefore would be equally valid conditional expectations) cannot also be continuous. So together, the two conditions ensure that the trivial-looking step $\mathbb{E}(m_0(X_i) | X_i) = m_0(X_i)$ – hidden in the second equality – really goes through.

Note finally that identification of a parameter is a necessary but not sufficient condition for consistent estimation. To see necessity, observe that the probability limit of any estimator is a function of the population distribution of observables. Existence of a consistent estimator therefore implies identification through the general proof strategy just given. The converse is not true: An identified parameter value can be consistently estimated in "nice" but not in all conceivable cases.

7 Extremum Estimators

We will now look at the most general class of estimators considered in this lecture, namely *extremum estimators*. An estimator is an extremum estimator if it can be characterized as

$$\hat{\theta} \equiv \arg\max_{\theta \in \Theta} Q_n(W_1, \ldots, W_n; \theta),$$

where Θ denotes some predefined parameter space.

The definition raises the question of existence and uniqueness of $\hat{\theta}$. Existence is frequently obvious and will be handled in theorems below. Regarding uniqueness, if the above arg max is not a singleton, we can define $\hat{\theta}$ to be any measurable function of $(W_1, \ldots, W_n; \theta)$ s.t. $\hat{\theta} \in \arg\max_{\theta \in \Theta} Q_n(W_1, \ldots, W_n; \theta)$, that is, the estimator is any measurable selection from the arg max.[5]

Example 7.1 ***GMM***

The GMM estimator is

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

so it is an extremum estimator.

Example 7.2 ***Method of Simulated Moments, Indirect Inference***

The Method of Simulated Moments is a GMM-like estimation method for settings where the criterion function is not available in closed form. Thus, let the function $\boldsymbol{\pi} : \Theta \mapsto \mathbb{R}^K$ map parameter values θ onto implied population quantities $\boldsymbol{\pi}(\theta)$ of which well-behaved estimators $\hat{\boldsymbol{\pi}}$ are available. Indeed, initially suppose that the $\boldsymbol{\pi}$ are population moments of some observable variables and $\hat{\boldsymbol{\pi}}$ their sample analog. While we will have to assume that $\boldsymbol{\pi}$ is somewhat well-behaved, it need not be available in closed form; indeed, the whole point is that it may have to be evaluated by simulation. For example, different values of preference and technology parameters in a structural model may imply different time series of observed outcomes.

The estimator $\hat{\theta}$ is computed by matching predicted moments to sample moments. More formally,

$$\begin{aligned} \hat{\theta} &\equiv \arg\min_{\theta \in \Theta} J(\theta, \hat{\boldsymbol{W}}) \\ J(\theta, \hat{\boldsymbol{W}}) &\equiv n \cdot (\boldsymbol{\pi}(\theta) - \hat{\boldsymbol{\pi}})' \hat{\boldsymbol{W}} (\boldsymbol{\pi}(\theta) - \hat{\boldsymbol{\pi}}). \end{aligned}$$

Under reasonable conditions, this estimator behaves similarly to GMM and approximates it in the limit as simulation inaccuracy in $\boldsymbol{\pi}(\theta)$ vanishes. The econometrically interesting aspect is that if

[5] The estimator could even be randomized, though that is measurable only upon adding an extraneous randomization to the data.

computational constraints limit the precision with which $\boldsymbol{\pi}(\theta)$ is computed, simulation noise must be taken into account when conducting inference.[6]

Next, identifying $\boldsymbol{\pi}$ with population moments and $\hat{\boldsymbol{\pi}}$ with their sample analogs simplifies exposition but is not essential. For example, $\boldsymbol{\pi}$ could also be population projections of some observable r.v.'s on others, pseudotrue parameters of some likelihood model (we will define these terms later), and so on, with $\hat{\boldsymbol{\pi}}$ a well-behaved estimator. The corresponding generalization of MSM is called Indirect Inference.

Example 7.3 ***Maximum Likelihood***

Terminology would suggest that the Maximum Likelihod (ML) estimator is

$$\hat{\theta} \equiv \arg\max_{\theta \in \Theta} f(W_1, \ldots, W_n; \theta),$$

i.e. it is the value of θ that maximizes the data's likelihood, that is, it maximizes the population likelihood (or, in discrete cases, probability) of observing the data that were in fact observed, as a function of θ. This estimator is intuitively appealing, and under reasonable conditions it has certain optimality properties. Therefore, it is fair to say that this is the estimator we should use if (i) we are willing to specify a precise likelihood and (ii) we can compute the estimator. In practice, either condition can be binding; for example, GMM avoids the former, and "pseudo-ML" estimation of willfully misspecified (but tractable) likelihoods is a commonplace technique.

Indeed, we will never work with the above, "high-concept" definition. First, for reasons that will become apparent, we equivalently maximize the log-likelihood:

$$\hat{\theta} = \arg\max_{\theta \in \Theta} \big\{ \log(f(W_1, \ldots, W_n; \theta)) \big\}.$$

Also, we will assume data to be i.i.d., so that the problem further specializes to

$$\hat{\theta} = \arg\max_{\theta \in \Theta} \Big\{ \log \prod_{i=1}^{n} f(W_i; \theta) \Big\} = \arg\max_{\theta \in \Theta} \Big\{ \frac{1}{n} \sum_{i=1}^{n} \log f(W_i; \theta) \Big\}.$$

This is easier to compute and, through the rescaling by $1/n$, turns ML into an M-estimator (to be defined later), which is useful for proving things. Finally, letting $W_i = (Y_i, X_i)$ for simplicity (here, X_i may subsume endogenous regressors), the likelihood can be factorized as

$$f(Y_i, X_i; \theta) = f(Y_i | X_i; \theta) f(X_i; \theta),$$

and in most applications, $f(X_i; \theta)$ is constant in θ. In these cases, the problem further simplifies to maximization of the conditional likelihood

$$\hat{\theta} = \arg\max_{\theta \in \Theta} \Big\{ \frac{1}{n} \sum_{i=1}^{n} \log f(Y_i | X_i; \theta) \Big\},$$

[6] Indeed, without this twist, this is really just GMM. In that sense, many "MSM" estimators in the empirical literature are misnamed.

which is the typical statement of the ML estimator in practice. In this lecture, the ML estimator will always rely on an i.i.d. assumption and the last simplification will typically be available, though it is not used to show results.

As an instance of ML, consider the *probit model*:

$$\begin{aligned} Y_i &= 1\{Y_i^* \geq 0\} \\ Y_i^* &= X_i'\theta_0 + \varepsilon_i \\ \varepsilon_i &\sim N(0,1) \end{aligned}$$

This is a binary response model. Classic applications would be choices by workers to enter the labor force, or by firms to enter a market.

In the probit model, for any given observation, we find that

$$\Pr(Y_i = 1|X_i) = \Pr(X_i'\theta_0 + \varepsilon_i \geq 0) = \Pr(\varepsilon_i \geq -X_i'\theta_0) = \Pr(\varepsilon_i \leq X_i'\theta_0) = \Phi(X_i'\theta_0),$$

thus the likelihood function is given by

$$f(Y_i|X_i,\theta_0) = \Phi(X_i'\theta_0)^{Y_i}\left(1-\Phi(X_i'\theta_0)\right)^{1-Y_i}.$$

The ML estimator maximizes

$$\sum_{i=1}^{n}\log f(Y_i|X_i,\theta) = \sum_{i=1}^{n}\big(Y_i\log\Phi(X_i'\theta) + (1-Y_i)\log\left(1-\Phi(X_i'\theta)\right)\big).$$

Example 7.4 ***Nonlinear Least Squares***

Nonlinear Least Squares applies to models where we are willing to specify that

$$\mathbb{E}(Y_i|X_i) = \rho(X_i,\theta_0)$$

or equivalently that

$$\begin{aligned} Y_i &= \rho(X_i,\theta_0) + \varepsilon_i \\ \mathbb{E}(\varepsilon|X_i) &= 0. \end{aligned}$$

One could estimate this by minimizing the sum of squared residuals (i.e., generalizing OLS, albeit in a different way than GMM does). The NLS estimator for this model would be

$$\hat{\theta}_{NLS} \equiv \arg\min_{\theta\in\Theta}\frac{1}{n}\sum_{i=1}^{n}\left(Y_i-\rho(X_i,\theta)\right)^2,$$

so this is an extremum estimator as well.

As an instance of NLS, consider estimating a CES production function:

$$\begin{aligned} Q_t &= A_0 \cdot \left(\delta_0 K_i^{-\rho_0} + (1-\delta_0) L_i^{-\rho_0}\right)^{-1/\rho_0} + \varepsilon_i \\ \mathbb{E}(\varepsilon_i | K_i, L_i) &= 0 \end{aligned}$$

Here, the parameter vector $\theta_0 = (A_0, \delta_0, \rho_0)$ could be estimated by NLS.

Definition 2 ***M-Estimators***

It is sometimes useful to summarize ML, NLS, some instances of GMM under the rubric of m-estimators. An m-estimator is an extremum estimator whose objective function Q_n is a sample sum or average of a known function m:[7]

$$Q_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} m(W_i, \theta).$$

Because sample averages are subject to Laws of Large Numbers, one can often establish theoretical results at this level. Important examples of extremum estimators that are not m-estimators are two-step GMM and minimum distance estimation.

7.1 Consistency of Extremum Estimators

There is a plethora of results on consistency of extremum estimators under different conditions. Two sucvh sets of conditions are especially important: They combine well-separatedness of the population criterion's maximum at θ_0 with either concavity of the objective function or with compactness of parameter space and uniform approximation of the population criterion. We will develop the latter with the most formal detail. Thus, define:

Definition 3 ***Definition: Uniform Convergence***

A sequence of random functions $\{f_n(\theta)\}$ is said to (weakly) converge to some limiting function $f(\theta)$ uniformly over Θ if

$$\begin{aligned} & \operatorname{plim} \sup_{\theta \in \Theta} |f_n(\theta) - f(\theta)| = 0 \\ \Leftrightarrow \forall \epsilon > 0, \qquad & \lim_{n \to \infty} \Pr\left(\sup_{\theta \in \Theta} |f_n(\theta) - f(\theta)| > \epsilon \right) = 0. \end{aligned}$$

Note the order of quantifiers: It is crucial that the sup is inside the Pr. In fact,

$$\sup_{\theta \in \Theta} \lim_{n \to \infty} \Pr\left(|f_n(\theta) - f(\theta)| > \epsilon \right) = 0$$

is equivalent to

$$\lim_{n \to \infty} \Pr\left(|f_n(\theta) - f(\theta)| > \epsilon \right) = 0, \forall \theta,$$

i.e. pointwise convergence.

[7] This is Hayashi's terminology. Some other texts use m-estimator as synonym for extremum estimator.

Example 7.5 *Consider the (nonrandom) function $f_n(x) = x/n$. This converges to zero pointwise but not uniformly over x.*

Theorem 7.1 ***Consistency of Extremum Estimators (I)***

Assume:

1. *There exists a nonstochastic function $Q(\theta)$ s.t.*

$$Q_n(\theta) \xrightarrow{p} Q(\theta) \text{ uniformly in } \theta \in \Theta.$$

2. *$Q(\cdot)$ is continuous.*

3. *$Q(\cdot)$ is uniquely maximized at θ_0.*

4. *Θ is compact.*

5. *$\hat{\theta}$ exists. (A sufficient condition is that Q_n is continuous.)*

Then

$$\hat{\theta} \xrightarrow{p} \theta_0,$$

i.e. $\hat{\theta}$ is (weakly) consistent.

Proof. Fix any $\epsilon > 0$. Need to show: $\Pr\big(|\hat{\theta} - \theta_0| \geq \epsilon\big) \to 0$.

Consider

$$Q_\epsilon \equiv \max_{\theta \in \Theta, |\theta - \theta_0| \geq \epsilon} Q(\theta).$$

By assumptions, Q_ϵ exists and $Q_\epsilon < Q(\theta_0)$. Define $\delta \equiv \frac{1}{2}\left(Q(\theta_0) - Q_\epsilon\right)$.

Note that

$$\begin{aligned}
\max_{\theta \in \Theta, |\theta - \theta_0| \geq \epsilon} Q_n(\theta) &= \max_{\theta \in \Theta, |\theta - \theta_0| \geq \epsilon} \left(Q_n(\theta) - Q(\theta) + Q(\theta)\right) \\
&\leq \max_{\theta \in \Theta, |\theta - \theta_0| \geq \epsilon} \left(Q_n(\theta) - Q(\theta)\right) + \max_{\theta \in \Theta, |\theta - \theta_0| \geq \epsilon} Q(\theta) \\
&\leq \max_{\theta \in \Theta, |\theta - \theta_0| \geq \epsilon} \left(Q_n(\theta) - Q(\theta)\right) + Q_\epsilon
\end{aligned}$$

and that uniform convergence in probability of $Q_n(\theta)$ implies

$$\Pr\left(\max_{\theta \in \Theta, |\theta - \theta_0| \geq \epsilon} \left(Q_n(\theta) - Q(\theta)\right) > \delta\right) \to 0.$$

Combine these to find

$$\Pr\left(\max_{\theta \in \Theta, |\theta - \theta_0| \geq \epsilon} Q_n(\theta) > Q_\epsilon + \delta\right) \to 0,$$

or equivalently (using $Q(\theta_0) = Q_\epsilon + 2\delta$)

$$\Pr\left(\max_{\theta \in \Theta, |\theta - \theta_0| \geq \epsilon} Q_n(\theta) < Q(\theta_0) - \delta\right) \to 1.$$

Pointwise convergence of Q_n also yields

$$\Pr\left(Q_n(\theta_0) > Q(\theta_0) - \delta\right) \to 1.$$

Combining the two, we find

$$\begin{aligned} & \Pr\left(Q_n(\theta_0) > \max_{\theta\in\Theta, |\theta-\theta_0|\geq\epsilon} Q_n(\theta)\right) \to 1 \\ \Rightarrow \quad & \Pr\left(\arg\max_{\theta\in\Theta, |\theta-\theta_0|\geq\epsilon} Q_n(\theta) = \arg\max_{\theta\in\Theta,} Q_n(\theta)\right) \to 0 \end{aligned}$$

as required. ■

Assumptions 2-4 are only needed to establish the fact that $Q_\epsilon < Q(\theta_0)$ for all $\epsilon > 0$. We can drop these assumptions if we can otherwise establish that $\sup_{\theta\in\Theta, |\theta-\theta_0|\geq\epsilon} Q(\theta) < Q(\theta_0)$ for all $\epsilon > 0$, a property called well-separatedness of the maximum. While this implies 3, it means that 2 and 4 can potentially be substituted for.

Example 7.6 ***Tightness (Necessity for Sufficiency) of Uniform Convergence***

Let $\Theta = [0, 3]$ *and*

$$Q_n(\theta) = \begin{cases} 2n\theta, & 0 \leq \theta \leq 1/2n \\ 2(1 - n\theta), & 1/2n < \theta \leq 1/n \\ 0, & 1/n < \theta \leq 1 \\ \frac{n}{n+1}(\theta - 1), & 1 < \theta \leq 2 \\ \frac{n}{n+1}(3 - \theta), & 2 < \theta \leq 3 \end{cases}.$$

Here, $\theta_0 = 2$ *but* $\hat{\theta} = 1/2n$.

As a rule of thumb, the hardest assumption to verify may be part 3, which you should recognize as an identification assumption. Indeed, if $\arg\max_{\theta\in\Theta} Q(\theta)$ is set-valued, then there is no way to consistently pick θ_0 from that set's elements.[8]

The following, alternative theorem is frequently helpful.

Theorem 7.2 ***Consistency of Extremum Estimators (II)***

Assume:

1. *There exists a nonstochastic function* $Q(\theta)$ *s.t.* $Q_n(\theta) \xrightarrow{p} Q(\theta)$ *pointwise.*
2. $Q_n(W_1, \ldots, W_n; \theta)$ *is concave in* θ *for all* $(W_1, \ldots, W_n)$.
3. $Q(\theta)$ *is uniquely maximized at* $\theta_0 \in \int \Theta$.

[8]The idea of extremum estimation can be generalized to this *set identified* case. This is a very active topic in the literature and your instructor worksmon it, but we will cover it in thsi lecture only extremely tangentially.

4. Θ *is convex.*

Then $\hat{\theta}$ *exists with probability approaching* 1 *and*

$$\hat{\theta} \xrightarrow{p} \theta_0,$$

i.e. $\hat{\theta}$ *is (weakly) consistent.*

Proof. See Theorem 2.7 in Newey and McFadden (1994). ■

The general proof involves results from real analysis and guards against somewhat contrived examples that are consistent with the assumptions. A simplified version in which $\Theta = \mathbb{R}$ admits an elementary direct proof (homework) and captures the intuition. Indeed, the essential difference between this theorem and the previous one is that imposing concavity allows us to drop compactness of parameter space and also uniformity of convergence of Q_n.

In the following, we explicitly discuss specializations of Theorem **??**. However, concavity of Q_n could stand in for compactness of Θ throughout, and we will mention some uses of this fact.

Theorem 7.3 ***Consistency of M-Estimators***

Assume:

1. $\{W_i\}$ *is i.i.d.,*

2. $m(W_i, \theta)$ *is continuous in* θ *for all* W_i*,*

3. $\mathbb{E}(m(W_i, \theta))$ *is uniquely maximized on* Θ *by* θ_0*,*

4. Θ *is compact,*

5. $\mathbb{E}\left(\sup_{\theta \in \Theta,} |m(W_i, \theta)|\right) < \infty$.[9]

Then $\hat{\theta} \xrightarrow{p} \theta_0$.

The Theorem uses the structure of m-estinators to simplify assumptions through the following result:

Theorem 7.4 ***Uniform Law of Large Numbers***

Let $\{W_i\}$ *be i.i.d., let* Θ *be compact, let* $m(W_i, \theta)$ *be continuous in* θ *for all* W_i*, and let* $\mathbb{E}\left(\sup_{\theta \in \Theta,} |m(W_i, \theta)|\right) < \infty$*. Then*

$$\frac{1}{n} \sum_i m(W_i, \theta) \xrightarrow{p} \mathbb{E}(m(W_i, \theta))$$

uniformly over Θ*. Furthermore,* $\mathbb{E}(m(W_i, \theta))$ *is continuous in* θ*.*

[9] The sup could be replaced by a max; it is there for unification with standard statements of the Uniform Law of Large Numbers.

Theorem **??** then follows easily, setting $Q(W_i, \theta) \equiv \mathbb{E}m(W_i, \theta)$. More generally, uniform convergence is usually established by finding some integrable function that dominates m.

The ML-estimator in its usual form is an m-estimator, and so Theorem **??** implies. We can add an important observation, however: The identification condition that Q be uniquely maximized at θ_0 is equivalent to the usual identification condition in ML, i.e., that no parameter value $\theta \neq \theta_0$ is observationally equivalent with θ_0.

Theorem 7.5 ***Identification Condition for ML***

The identification condition, i.e. assumption 3 in Theorem **??**, *is equivalent to the usual identification condition for ML:*

$$\boldsymbol{\theta \neq \theta_0} \Longrightarrow [W_i \in A \Rightarrow f(W_i; \theta) \neq f(W_i; \theta_0)]\,, \text{ some event } A \text{ with } \Pr(A) > 0.$$

Proof. Write

$$\begin{aligned}
& \mathbb{E} \log f(W_i; \theta) - \mathbb{E} \log f(W_i; \theta_0) \\
= \;& \mathbb{E} \left(\log f(W_i; \theta) - \log f(W_i; \theta_0) \right) \\
= \;& \mathbb{E} \left(\log \frac{f(W_i; \theta)}{f(W_i; \theta_0)} \right) \\
\leq \;& \log \mathbb{E} \left(\frac{f(W_i; \theta)}{f(W_i; \theta_0)} \right) \\
= \;& \log \int \frac{f(W_i; \theta)}{f(W_i; \theta_0)} f(W_i; \theta_0) dW_i \\
= \;& \log \int f(W_i; \theta) dW_i \\
= \;& \log 1 = 0,
\end{aligned}$$

where the inequality is Jensen's inequality and where it is essential that $\mathbb{E}$ corresponds to integration weighted by the *true* likelihood $f(W_i, \theta_0)$. The inequality holds with equality iff $\frac{f(W_i;\theta)}{f(W_i;\theta_0)}$ is constant almost everywhere, which obtains iff $f(W_i; \theta) = f(W_i; \theta_0)$ almost everywhere because both integrate to 1. ■

The gist of the above proof is that we have established the *Kullback-Leibler information inequality.*

As an example, reconsider the linear regression model with normal errors,

$$\begin{aligned}
Y_i &= X_i'\beta + \varepsilon_i, \\
\varepsilon_i &\overset{i.i.d.}{\sim} N(0, \sigma^2).
\end{aligned}$$

The conditional log likelihood for observation i is

$$\log f(Y_i; \theta | X_i) = -\frac{1}{2} \log 2\pi - \frac{1}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \left(Y_i - X_i'\beta \right)^2;$$

here, $\theta \equiv (\beta', \sigma^2)'$.

Let $\theta_0 \equiv (\beta_0', \sigma_0^2)'$. Clearly σ_0^2 is identified – it equals $\text{var}(Y_i|X_i)$ and as such can be consistently estimated, which establishes identification (as you prove in a homework). This leaves the question of identifying β_0. Accoridng to the identification condition for ML, β_0 is not identified if there exists $\beta \neq \beta_0$ s.t. the distribution of $(Y_i|x_i)$ would a.s. be the same if β as opposed to β_0 were true. Given that the conditional expectation of Y_i equals $E(Y_i|X_i) = X_i'\beta_0$, this would require $E(YHi|X_i) = X_i'\beta$ for almost every X_i. This, in turn, would imply that $\mathbb{E}\left(X_i'\beta_0 - X_i'\beta\right)^2 = 0$. Conversely, identification is therefore ensured if we can claim $\beta \neq \beta_0 \Rightarrow \mathbb{E}\left(X_i'\beta_0 - X_i'\beta\right)^2 > 0$. Writing

$$\mathbb{E}\left(X_i'\beta_0 - X_i'\beta\right)^2 = \mathbb{E}\left(X_i'\left(\beta_0 - \beta\right)\right)^2 = \left(\beta_0 - \beta\right)'\mathbb{E}(X_iX_i')\left(\beta_0 - \beta\right),$$

the desired implication is seen to hold iff $\mathbb{E}(X_iX_i')$ is positive definite, which obtains iff $\mathbb{E}(X_iX_i')$ has full rank. Thus, identification analysis in this model is unchanged by adding the assumption of normal errors. You can verify that the ML estimator for this model is just the OLS estimator.

We conclude analysis of consistency by cosnidering nonlinear GMM.

Theorem 7.6 ***Consistency of Nonlinear GMM***

Assume that

1. $\{W_i\}$ *is i.i.d.,*
2. $g(W_i, \theta)$ *is continuous in* θ *for all* W_i*,*
3. $\mathbb{E}g(W_i, \theta) = \mathbf{0}$ *iff* $\theta = \theta_0$*,*
4. Θ *is compact,*
5. $\mathbb{E}\left(\max_{\theta\in\Theta} ||g(W_i, \theta)||\right) < \infty$.

Then, the (nonlinear) GMM estimator

$$\hat{\theta}_{GMM} \equiv \arg\min_{\theta\in\Theta}\left(\frac{1}{n}\sum_{i=1}^n g(W_i, \theta)\right)'\hat{\boldsymbol{W}}\left(\frac{1}{n}\sum_{i=1}^n g(W_i, \theta)\right)$$

with $\hat{\boldsymbol{W}}$ *as before, is consistent, i.e.* $\hat{\theta}_{GMM} \xrightarrow{p} \theta_0$.

Compactness of Θ was not invoked when proving consistency of linear GMM earlier in this lecture. In the context of this chapter's results, this can be explained as follows: Linearity of g implies convexity of $\left(\frac{1}{n}\sum_{i=1}^n g(W_i, \theta)\right)'\hat{\boldsymbol{W}}\left(\frac{1}{n}\sum_{i=1}^n g(W_i, \theta)\right)$, which can substitute for compactness of Θ through the alternative theorem. (By the same token, if g is linear, we can weaken assumption 5 above since we do not need a uniform LLN. This essentially recovers a result we already proved.)

As mentioned before, in a nonlinear application, the identification condition that $\mathbb{E}g(W_i, \theta) = \mathbf{0}$ iff $\theta = \theta_0$ can rarely be stated in terms of primitives; in fact, it can be hard to verify.

7.2 Asymptotic Normality of Extremum Estimators

We now collect conditions for extremum estimators to be asymptotically normal. The treatment will be a bit simpler than in Hayashi.

Theorem 7.7 ***Asymptotic Distribution of Extremum Estimators***

Assume that

1. $\hat{\theta} \xrightarrow{p} \theta_0$,
2. $\theta_0 \in \text{int}(\Theta)$,
3. *there exists a neighborhood* $\mathcal{N}$ *of* θ_0 *s.t. for any* $\theta \in \mathcal{N}$, $Q_n(\theta)$ *is twice continuously differentiable in* θ *for any* W_i,
4. $\sqrt{n}\frac{\partial Q_n(\theta_0)}{\partial \theta} \xrightarrow{d} N(\mathbf{0}, \Sigma)$, Σ *positive definite,*
5. $\boldsymbol{H}(\theta) \equiv \frac{\partial^2 Q(\theta)}{\partial \theta \partial \theta'}$ *is continuous at* θ_0.
6. $\sup_{\theta \in \mathcal{N}} \left\| \frac{\partial^2 Q_n(\theta)}{\partial \theta \partial \theta'} - \frac{\partial^2 Q(\theta)}{\partial \theta \partial \theta'} \right\| \xrightarrow{p} 0$.
7. $\boldsymbol{H}(\theta_0)$ *is nonsingular.*

Then

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N\big(\mathbf{0}, \text{Avar}(\hat{\theta})\big)$$
$$\text{Avar}(\hat{\theta}) = (\boldsymbol{H}(\theta_0))^{-1} \Sigma (\boldsymbol{H}(\theta_0))^{-1}.$$

Proof. The first step is a mean value theorem expansion:

$$\frac{\partial Q_n(\hat{\theta})}{\partial \theta} = \frac{\partial Q_n(\theta_0)}{\partial \theta} + \frac{\partial^2 Q_n(\overline{\theta})}{\partial \theta \partial \theta'}(\hat{\theta} - \theta_0),$$

where $\overline{\theta}$ lies componentwise between $\hat{\theta}$ and θ_0. Notice that this uses twice continuous differentiability of Q_n.

But we also know that

$$\hat{\theta} = \arg\max_{\theta \in \Theta,} Q_n(\theta).$$

Our assumptions imply continuity of Q_n. Since also $\theta_0 \in int(\Theta)$ by assumption and $\hat{\theta} \xrightarrow{p} \theta_0$, we have that $\hat{\theta} \in \text{int}(\Theta)$ with probability approaching 1 as $n \to \infty$. Hence, $\hat{\theta}$ is (w.p.a. 1) characterized by a first-order condition:

$$\frac{\partial Q_n(\hat{\theta})}{\partial \theta} = \mathbf{0}.$$

Substituting for this, we find that

$$\begin{aligned} \mathbf{0} &= \frac{\partial Q_n(\theta_0)}{\partial \theta} + \frac{\partial^2 Q_n(\bar{\theta})}{\partial \theta \partial \theta'}(\hat{\theta} - \theta_0) \\ \Rightarrow \mathbf{0} &= \sqrt{n}\frac{\partial Q_n(\theta_0)}{\partial \theta} + \frac{\partial^2 Q_n(\bar{\theta})}{\partial \theta \partial \theta'}\sqrt{n}(\hat{\theta} - \theta_0). \end{aligned}$$

To fix the big idea first, assume that $\frac{\partial^2 Q_n(\bar{\theta})}{\partial \theta \partial \theta'}$ is nonsingular, then we can solve this:

$$\sqrt{n}(\hat{\theta} - \theta_0) = -\left(\frac{\partial^2 Q_n(\bar{\theta})}{\partial \theta \partial \theta'}\right)^{-1}\sqrt{n}\frac{\partial Q_n(\theta_0)}{\partial \theta}.$$

Now, $\sqrt{n}\frac{\partial Q_n(\theta_0)}{\partial \theta} \xrightarrow{d} N(\mathbf{0}, \Sigma)$ by assumption; if also $\frac{\partial^2 Q_n(\bar{\theta})}{\partial \theta \partial \theta'} \xrightarrow{p} \boldsymbol{H}(\theta_0)$, we have

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N\left(\mathbf{0}, (\boldsymbol{H}(\theta_0))^{-1}\Sigma(\boldsymbol{H}(\theta_0))^{-1}\right)$$

as required.

We conclude by filling in some details: Note that

$$\begin{aligned} \left\| \frac{\partial^2 Q_n(\bar{\theta})}{\partial \theta \partial \theta'} - \frac{\partial^2 Q(\theta_0)}{\partial \theta \partial \theta'} \right\| &= \left\| \frac{\partial^2 Q_n(\bar{\theta})}{\partial \theta \partial \theta'} - \frac{\partial^2 Q(\bar{\theta})}{\partial \theta \partial \theta'} + \frac{\partial^2 Q(\bar{\theta})}{\partial \theta \partial \theta'} - \frac{\partial^2 Q(\theta_0)}{\partial \theta \partial \theta'} \right\| \\ &\leq \left\| \frac{\partial^2 Q_n(\bar{\theta})}{\partial \theta \partial \theta'} - \frac{\partial^2 Q(\bar{\theta})}{\partial \theta \partial \theta'} \right\| + \left\| \boldsymbol{H}(\bar{\theta}) - \boldsymbol{H}(\theta_0) \right\| \xrightarrow{p} 0. \end{aligned}$$

Here, the inequality is the triangle inequality; recalling that $\bar{\theta} \xrightarrow{p} \theta_0$ from consistency of $\hat{\theta}$, $\left\| \frac{\partial^2 Q_n(\bar{\theta})}{\partial \theta \partial \theta'} - \frac{\partial^2 Q(\bar{\theta})}{\partial \theta \partial \theta'} \right\| \xrightarrow{p} 0$ follows from assumption 6 and $\left\| \frac{\partial^2 Q(\bar{\theta})}{\partial \theta \partial \theta'} - \frac{\partial^2 Q(\theta_0)}{\partial \theta \partial \theta'} \right\| \xrightarrow{p} 0$ follows from continuity of $\boldsymbol{H}$. This immediately yields $\frac{\partial^2 Q_n(\bar{\theta})}{\partial \theta \partial \theta'} \xrightarrow{p} \boldsymbol{H}(\theta_0)$ and by openness of the set of nonsingular matrices also implies nonsingularity of $\frac{\partial^2 Q_n(\bar{\theta})}{\partial \theta \partial \theta'}$ (with probability approaching 1). ■

Assumption 6 would typically be derived as implication of a uniform law of large numbers. Standard conditions on the d.g.p. and a uniform (over $\mathcal{N}$) upper bound on $\left\| \frac{\partial^2 Q(\theta)}{\partial \theta \partial \theta'} \right\|$ would do the trick.

Specialization to ML

We now specialize the result to ML. This allows us to derive assumption 4 from primitives and discover an important fact about the asymptotic variance. We will assume that data are i.i.d. with individual likelihood $f(W_i; \theta_0)$. We begin by deriving two important facts.

Since $f(W_i; \theta_0)$ is a p.d.f., we have

$$\begin{aligned} \int f(W; \theta_0) dW &\equiv 1 \\ \Longrightarrow \int \frac{\partial f(W; \theta_0)}{\partial \theta} dW &\equiv \mathbf{0} \\ \Longrightarrow \int \frac{\partial \log f(W; \theta_0)}{\partial \theta} f(W; \theta_0) dW &\equiv \mathbf{0}. \end{aligned}$$

In particular, this implies that $\mathbb{E}\frac{\partial \log f(W_i;\theta_0)}{\partial \theta} = \mathbf{0}$. This result will be important in its own right, but we can also take derivatives once more:

$$\begin{aligned}
&\int \frac{\partial^2 \log f(W;\theta_0)}{\partial\theta\partial\theta'} f(W;\theta_0)dW + \int \frac{\partial \log f(W;\theta_0)}{\partial\theta}\frac{\partial \log f(W;\theta_0)}{\partial\theta'} f(W;\theta_0)dW \equiv \mathbf{0} \\
&\qquad \Longrightarrow \mathbb{E}\left(\frac{\partial^2 \log f(W_i;\theta_0)}{\partial\theta\partial\theta'}\right) + \mathbb{E}\left(\frac{\partial \log f(W_i;\theta_0)}{\partial\theta}\frac{\partial \log f(W_i;\theta_0)}{\partial\theta'}\right) \equiv \mathbf{0} \\
&\qquad\qquad \Longrightarrow \mathbb{E}\left(\frac{\partial^2 \log f(W_i;\theta_0)}{\partial\theta\partial\theta'}\right) \equiv -\mathbb{E}\left(\frac{\partial \log f(W_i;\theta_0)}{\partial\theta}\frac{\partial \log f(W_i;\theta_0)}{\partial\theta'}\right).
\end{aligned}$$

The last line is famous as *information matrix equality.*

Now write

$$\begin{aligned}
Q_n(\theta_0) &= \frac{1}{n}\sum_{i=1}^{n} \log f(W_i;\theta_0) \\
\Longrightarrow \frac{\partial Q_n(\theta_0)}{\partial\theta} &= \frac{1}{n}\sum_{i=1}^{n} \frac{\partial \log f(W_i;\theta_0)}{\partial\theta}.
\end{aligned}$$

But we just showed that $\mathbb{E}\frac{\partial \log f(W_i;\theta_0)}{\partial \theta} = \mathbf{0}$. We thus have

$$\sqrt{n}\frac{\partial Q_n(\theta_0)}{\partial\theta} \xrightarrow{d} N\left(\mathbf{0}, \mathbb{E}\left(\frac{\partial \log f(W_i;\theta_0)}{\partial\theta}\frac{\partial \log f(W_i;\theta_0)}{\partial\theta'}\right)\right)$$

by the CLT. This establishes assumption 4.

Substituting these findings into the theorem, we get that

$$\begin{aligned}
&\sqrt{n}(\hat{\theta}_{ML} - \theta_0) \xrightarrow{d} \\
&N\Bigg(\mathbf{0}, \underbrace{\left(\mathbb{E}\left(\frac{\partial^2 \log f(W_i;\theta_0)}{\partial\theta\partial\theta'}\right)\right)^{-1}}_{\boldsymbol{H}(\theta_0)} \underbrace{\mathbb{E}\left(\frac{\partial \log f(W_i;\theta_0)}{\partial\theta}\frac{\partial \log f(W_i;\theta_0)}{\partial\theta'}\right)}_{\Sigma} \underbrace{\left(\mathbb{E}\left(\frac{\partial^2 \log f(W_i;\theta_0)}{\partial\theta\partial\theta'}\right)\right)^{-1}}_{\boldsymbol{H}(\theta_0)}\Bigg).
\end{aligned}$$

Now for the punch line: Substituting in from the information matrix equality, we get

$$\sqrt{n}(\hat{\theta}_{ML} - \theta_0) \xrightarrow{d} N\left(\mathbf{0}, -\boldsymbol{H}(\theta_0)^{-1}\right).$$

This looks a bit like the simplification of the sandwich variance when we switch from ordinary GMM to efficient GMM. Should ML be in some sense efficient? Let's define the **(Fisher) information** inherent in a sample as $\mathbb{I}_n(\theta_0) \equiv -\mathbb{E}\frac{\partial^2 \log f(W_1,\ldots,W_n;\theta_0)}{\partial\theta\partial\theta'}$. Under this lecture's i.i.d. assumption, this simplifies to $\mathbb{I}_n(\theta_0) = n\mathbb{I}_1(\theta_0)$. For simplicity, we will also write $\mathbb{I}(\theta_0)$ for $\mathbb{I}_1(\theta_0)$. Then we can write $\sqrt{n}(\hat{\theta}_{ML} - \theta_0) \xrightarrow{d} N(\mathbf{0}, \mathbb{I}(\theta_0)^{-1})$.

We next take a little detour and relate this variance expression to a famous theorem.

Theorem 7.8 ***Cramér-Rao Bound***

Let θ be scalar and assume some regularity conditions, then

$$var(\hat{\theta}) \geq \left(\frac{d\mathbb{E}\hat{\theta}}{d\theta}\right)^2 \mathbb{I}_n(\theta_0)^{-1},$$

where $\mathbb{E}\hat{\theta} = \int \hat{\theta}(W) f(W;\theta_0) dW$ *is the sampling expectation of* $\hat{\theta}$ *as a function of the true value* θ_0. *(Here again,* W *collects all data.)*

If $\hat{\theta}$ *is unbiased, then it follows that*

$$\operatorname{var}(\hat{\theta}) \geq \mathbb{I}_n(\theta_0)^{-1}.$$

While not established below, this generalizes to vectors: If $\hat{\theta}$ *is unbiased, then*

$$\operatorname{var}(\hat{\theta}) - \mathbb{I}_n(\theta_0)^{-1}$$

is positive semidefinite.

Proof. Write

$$\begin{aligned}
\frac{d}{d\theta}\mathbb{E}\hat{\theta} &= \frac{d}{d\theta}\int \hat{\theta}(W) f(W;\theta_0) dW \\
&= \int \hat{\theta}(W) \frac{\partial f(W;\theta_0)}{\partial \theta} dW \\
&= \int \hat{\theta}(W) \frac{\partial f(W;\theta_0)}{\partial \theta} \frac{f(W;\theta_0)}{f(W;\theta_0)} dW \\
&= \mathbb{E}\left(\hat{\theta}(W) \frac{\partial \log f(W;\theta_0)}{\partial \theta} \right) \\
&= \mathbb{E}\left((\hat{\theta}(W) - \mathbb{E}\hat{\theta}) \frac{\partial \log f(W;\theta_0)}{\partial \theta} \right),
\end{aligned}$$

where W is the entire data vector and the last step used that $\mathbb{E}\big(\big(\mathbb{E}\hat{\theta}\big) \frac{\partial \log f(W;\theta_0)}{\partial \theta} \big) = \mathbb{E}\hat{\theta}\mathbb{E}\frac{\partial \log f(W;\theta_0)}{\partial \theta} = 0$ from before. Squaring both sides and using the Cauchy-Schwarz inequality: $\mathbb{E}(XY)^2 \leq \mathbb{E}(X^2)\mathbb{E}(Y^2)$, we get

$$\begin{aligned}
\left(\frac{d\mathbb{E}\hat{\theta}}{d\theta} \right)^2 &= \left(\mathbb{E}\left((\hat{\theta}(W) - \mathbb{E}\hat{\theta}) \frac{\partial \log f(W;\theta_0)}{\partial \theta} \right) \right)^2 \\
&\leq \underbrace{\mathbb{E}\big(\hat{\theta}(W) - \mathbb{E}\hat{\theta}\big)^2}_{\operatorname{var}(\hat{\theta})} \qquad \underbrace{\mathbb{E}\left(\frac{\partial \log f(W;\theta_0)}{\partial \theta} \right)^2}_{\mathbb{I}_n(\theta_0), \text{ using the information matrix equality}}
\end{aligned}$$

But unbiasedness of $\hat{\theta}$ implies $\frac{d}{d\theta}\mathbb{E}\hat{\theta} = 1$, hence the result. ■

Some remarks on this result follows.

- This is a finite sample result! No asymptotic approximation was used.
- An intuition for this is that the information matrix is a Hessian, i.e. it reflects the local degree of curvature of the likelihood function. More curvature means that the likelihood is more informative; less curvature leads to a rather flat surface around the maximum.

- The theorem does not directly apply to ML for a number of reasons. Firstly, it applies to unbiased estimators, whereas ML is biased. Hence, it does not follow that ML minimizes mean square error. ML is however, also asymptotically efficient in a sense defined via limit experiments, which we will not expand here.

- Just to test understanding, note also that $\text{Avar}\big(\hat{\theta}_{ML}\big) = \mathbb{I}(\theta_0)^{-1}$ does not imply $\text{Var}\big(\sqrt{n}\big(\hat{\theta}_{ML} - \theta_0\big)\big) \xrightarrow{p} \mathbb{I}(\theta_0)^{-1}$. (Why not?) The latter is also true in "nice" cases but not shown here.

There are several possibilities for estimating the asymptotic variance of the ML estimator. Two important estimators are the sample analog of $-\mathbb{E}\Big(\frac{\partial^2 \log f(W_i;\theta_0)}{\partial\theta\partial\theta'}\Big)$, which makes sense especially if analytic restrictions limit the Hessian's degrees of freedom, or the sample analog of $\mathbb{E}\Big(\frac{\partial \log f(W_i;\theta_0)}{\partial\theta}\frac{\partial \log f(W_i;\theta_0)}{\partial\theta'}\Big)$ (=*outer product of gradient estimator*) which estimates the same thing due to the information matrix equality and may be easier to evaluate.

Next, asymptotic normality of GMM estimators is implied by the above if we take $\bar{g}_n$ to be twice differentiable. However, we can use the structure of the GMM objective fucntion to improve on this.

Theorem 7.9 ***Asymptotic Distribution of GMM Estimators***

Assume that

1. $\hat{\theta}_{GMM} \xrightarrow{p} \theta_0$,
2. $\theta_0 \in int(\Theta)$,
3. *there exists a neighborhood* $\mathcal{N}$ *of* θ_0 *s.t. for any* $\theta \in \mathcal{N}$, $g(W_i, \theta)$ *is continuously differentiable in* θ *for any* W_i,
4. $\frac{1}{\sqrt{n}}\sum_{i=1}^n g(W_i, \theta_0) \xrightarrow{d} N(\mathbf{0}, \boldsymbol{S})$, $\boldsymbol{S}$ *positive definite,*
5. $\sup_{\theta\in\mathcal{N}}\Big\|\frac{1}{n}\sum_{i=1}^n \frac{\partial g(W_i,\theta)}{\partial\theta'} - \mathbb{E}\frac{\partial g(W_i,\theta)}{\partial\theta'}\Big\| \xrightarrow{p} 0.$
6. $\hat{\boldsymbol{W}} \xrightarrow{p} \boldsymbol{W}$, *where* $\boldsymbol{W}$ *is symmetric and positive definite,*
7. $\boldsymbol{G}(\theta_0)$ *is of full column rank.*

Then

$$\sqrt{n}(\hat{\theta}_{GMM} - \theta_0) \xrightarrow{d} N\big(\mathbf{0}, (\boldsymbol{G}'\boldsymbol{W}\boldsymbol{G})^{-1}\boldsymbol{G}'\boldsymbol{W}\boldsymbol{S}\boldsymbol{W}\boldsymbol{G}(\boldsymbol{G}'\boldsymbol{W}\boldsymbol{G})^{-1}\big).$$

Proof. As before, $\hat{\theta}_{GMM}$ is eventually interior. Defining $\bar{g}_n(\theta) \equiv \frac{1}{n}\sum g(W_i, \theta)$ with Jacobian $\boldsymbol{G}_n \equiv \frac{\partial \bar{g}_n(\theta)}{\partial\theta'} = \frac{1}{n}\sum \frac{\partial g(W_i,\theta)}{\partial\theta'}$, we have the first-order condition

$$\mathbf{0} = \frac{\partial Q_n(\hat{\theta}_{GMM})}{\partial\theta} = \frac{\partial}{\partial\theta}\big(\bar{g}_n(\hat{\theta}_{GMM})'\hat{\boldsymbol{W}}\bar{g}_n(\hat{\theta}_{GMM})\big) = 2\boldsymbol{G}_n\ (\hat{\theta}_{GMM})'\hat{\boldsymbol{W}}\bar{g}_n(\hat{\theta}_{GMM}).$$

On the other hand, we can use the Mean Value Theorem to expand $\bar{g}_n(\theta)$:

$$\bar{g}_n(\hat{\theta}_{GMM}) = \bar{g}_n(\theta_0) + \boldsymbol{G}_n(\bar{\theta})\big(\hat{\theta}_{GMM} - \theta_0\big).$$

Substituting for this, we find that

$$\begin{aligned}
\mathbf{0} &= \boldsymbol{G}_n(\hat{\theta}_{GMM})'\hat{\boldsymbol{W}}\big(\bar{g}_n(\theta_0) + \boldsymbol{G}_n(\bar{\theta})\big(\hat{\theta}_{GMM} - \theta_0\big)\big) \\
&= \boldsymbol{G}_n(\hat{\theta}_{GMM})'\hat{\boldsymbol{W}}\bar{g}_n(\theta_0) + \boldsymbol{G}_n(\hat{\theta}_{GMM})'\hat{\boldsymbol{W}}\boldsymbol{G}_n(\bar{\theta})\big(\hat{\theta}_{GMM} - \theta_0\big) \\
\Rightarrow \sqrt{n}\big(\hat{\theta}_{GMM} - \theta_0\big) &= -\big(\boldsymbol{G}_n(\hat{\theta}_{GMM})'\hat{\boldsymbol{W}}\boldsymbol{G}_n(\bar{\theta})\big)^{-1}\boldsymbol{G}_n(\hat{\theta}_{GMM})'\hat{\boldsymbol{W}}\sqrt{n}\bar{g}_n(\theta_0).
\end{aligned}$$

Observing that $\sqrt{n}\bar{g}_n(\theta_0) \xrightarrow{d} N(\mathbf{0}, \boldsymbol{S})$ and $\boldsymbol{G}_n \xrightarrow{p} \boldsymbol{G}$ in analogy to the previous theorem, the result now follows by Slutsky. ■

The main difference from the previous theorem is that we get away with $\bar{g}_n$ being once continuously differentiable. This is because Q_n is a quadratic form in $\bar{g}_n$, so one-time differentiability of Q_n follows from its structure, and only the next level of differentiability must be explicitly assumed. For the same reason, assumption 5 in the previous theorem is already implied by assumption 3 here. The last two assumptions here jointly stand in for the last assumption in the earlier theorem.

We can optimize the choice of $\hat{\boldsymbol{W}}$ just as before, and it will again lead to simplification of the sandwich matrix.

Comparing GMM and ML

Whenever we specify a complete likelihood for a model, we can perform ML but we could also use GMM. After all, if we know the likelihood, we can find many functions g s.t. $\mathbb{E}g(W_i, \theta) = \mathbf{0}$. To give one salient example, we know from the section on ML above that choosing $g(W_i, \theta) = \partial \log f(w_i; \theta)/\partial\theta$ will do.

This raises some interesting questions. Is there an optimal way to generate such functions from a likelihood? Can we match or beat the performance of ML? The questions are resolved by the following facts, stated here without proof:

$$\begin{aligned}
&\big(\boldsymbol{G}'\boldsymbol{S}^{-1}\boldsymbol{G}\big)^{-1} - \mathbb{I}(\theta_0)^{-1} \text{ is positive semidefinite.} \\
&\big(\boldsymbol{G}'\boldsymbol{S}^{-1}\boldsymbol{G}\big)^{-1} = \mathbb{I}(\theta_0)^{-1} \text{ if } g(W_i, \theta) = \frac{\partial \log f(W_i; \theta_0)}{\partial\theta}.
\end{aligned}$$

Some implications of these facts are as follows: GMM cannot (asymptotically) beat ML estimation performed using the true likelihood, whether or not we know that likelihood. If we know the likelihood, then we cannot improve on ML by smart application of GMM, but we can match the performance of ML by choosing $g(W_i, \theta) = \partial \log f(W_i; \theta)/\partial\theta$. The latter fact is clear because, if we perform GMM with those moment conditions, then we effectively solve the FOC's characterizing the ML estimator. That is, the GMM estimator will exactly recover the ML estimator.

This last observation is also interesting for other reasons: Knowing that ML algebraically coincides with GMM estimation of the score equations can be helpful in proving properties of ML and explains why we got away with only one-time differentiability of the moment functions in the preceding proof.

Regarding the hope that GMM may outperform ML: This is of course false. $\text{Avar}(\hat{\theta}_{GMM}) \geq \text{Avar}(\hat{\theta}_{ML})$, and, while GMM achieves the efficiency bound if the moments are chosen optimally, those optimal moments depend on the likelihood function and can, therefore, be known in general only if the likelihood is specified. In short, there is no way of getting ML efficiency without having likelihood information.

7.3 Hypothesis Testing

We will now establish the trinity of hypothesis tests for extremum estimators: the likelihood ratio, Wald, and Lagrange multiplier statistics. For all of these, we will look at the following environment and maintained assumptions.

We want to test

$$H_0 : \boldsymbol{a}(\theta_0) = \mathbf{0},$$

where $\boldsymbol{a}$ is continuously differentiable with Jacobian $\boldsymbol{A}(\theta) \equiv \frac{\partial \boldsymbol{a}(\theta)}{\partial \theta}$. We need $\boldsymbol{A}(\theta_0)$ to be of full row rank. This ensures that we are indeed looking at $\#\boldsymbol{a}$ restrictions.

We will operate with both the unconstrained extremum estimator

$$\hat{\theta} \equiv \arg\max_{\theta \in \Theta} Q_n(\theta)$$

and the constrained estimator

$$\tilde{\theta} \equiv \arg \max_{\theta \in \boldsymbol{\Theta}: \boldsymbol{a}(\theta) = \mathbf{0}} Q_n(\theta).$$

These estimators achieve objective values of $Q_n(\hat{\theta})$ respectively $Q_n(\tilde{\theta})$. $Q_n(\hat{\theta}) \geq Q_n(\tilde{\theta})$ because the latter stems from a (more) constrained optimization, but if H_0 holds, then the constraints should not bind at the limit, and hence the following things should happen:

- The difference between $\hat{\theta}$ and $\tilde{\theta}$ should vanish.
- The difference between $Q_n(\hat{\theta})$ and $Q_n(\tilde{\theta})$ should vanish.
- The shadow price of relaxing H_0, that is, the constrained problem's Lagrange multiplier, should vanish.

The trinity of test statistics is based on these three ideas. Two of them relate to the statistics we saw previously, one of them introduces a new one. We will consider them in the above order.

Definition 4 ***Test Statistics***

The Wald, Likelihood Ratio, and Lagrange Multiplier test statistics are

$$\begin{aligned} W &\equiv \sqrt{n}\boldsymbol{a}(\hat{\theta})'\big(\boldsymbol{A}(\hat{\theta})\hat{\Sigma}^{-1}\boldsymbol{A}'(\hat{\theta})\big)^{-1}\sqrt{n}\boldsymbol{a}(\hat{\theta}) \\ LR &\equiv 2n\big(Q_n(\hat{\theta}) - Q_n(\tilde{\theta})\big) \\ LM &\equiv n\frac{\partial Q_n(\tilde{\theta})}{\partial\theta}'\tilde{\Sigma}^{-1}\frac{\partial Q_n(\tilde{\theta})}{\partial\theta}, \end{aligned}$$

where $\hat{\Sigma}$ and $\tilde{\Sigma}$ are further specified below (though in principle, both could be any consistent estimator of Σ).

We will see that these statistics do asymptotically the same thing. They may not be equally easy to compute, however. Computation of W avoids solving the constrained estimation problem, whereas computing LM avoids solving the unconstrained problem. Depending on application, either could be computationally advantageous. For contrasting examples, think of hypotheses that effectively state irrelevance of numerous regressors, but also of intricate nonlinear hypotheses concerning parameters that are otherwise easy to estimate.

Theorem 7.10 ***Hypothesis Testing***

Suppose that:

1. *Taylor expansion: $\sqrt{n}\big(\hat{\theta} - \theta_0\big) = -\Psi^{-1}\sqrt{n}\frac{\partial Q_n(\theta_0)}{\partial\theta} + o_p(1)$, where the specification of Ψ depends on the exact estimator: It equals $\boldsymbol{H}(W_i, \theta_0)$ if we use the first normality theorem above, and it equals $-\boldsymbol{G}'\boldsymbol{W}\boldsymbol{G}$ for GMM estimators.*
2. *$\sqrt{n}\frac{\partial Q_n(\theta_0)}{\partial\theta} \xrightarrow{d} N(\mathbf{0}, \Sigma)$, Σ positive definite.*
3. *$\sqrt{n}\big(\tilde{\theta} - \theta_0\big) = O_P(1)$.*
4. *$\Sigma = -\Psi$.*

(All of these are either assumptions or implications of the two normality theorems above.)

Then all of the test statistics just defined converge in distribution to $\chi^2(\#\boldsymbol{a})$. Furthermore, though not proved below, they are asymptotically equivalent.

We next develop the lengthy proof of this.

(i) Wald Statistic

Using a Mean Value Theorem once again,

$$\begin{aligned} \boldsymbol{a}(\hat{\theta}) &= \boldsymbol{a}(\theta_0) + \boldsymbol{A}(\bar{\theta})\big(\hat{\theta} - \theta_0\big) \\ \Rightarrow \sqrt{n}\boldsymbol{a}(\hat{\theta}) &= \sqrt{n}\boldsymbol{A}(\bar{\theta})\big(\hat{\theta} - \theta_0\big) \\ &= \underbrace{\sqrt{n}\big(\boldsymbol{A}(\bar{\theta}) - \boldsymbol{A}(\theta_0)\big)\big(\hat{\theta} - \theta_0\big)}_{\xrightarrow{p}\mathbf{0}} + \sqrt{n}\boldsymbol{A}(\theta_0)\big(\hat{\theta} - \theta_0\big) \\ &= \boldsymbol{A}(\theta_0) \cdot \sqrt{n}\big(\hat{\theta} - \theta_0\big) + o_p(1), \end{aligned} \tag{1}$$

where we used the first two maintained assumptions and continuous differentiability of $\boldsymbol{a}$. Plugging into the first maintained assumption, we get

$$\sqrt{n}\boldsymbol{a}(\hat{\theta}) = -\boldsymbol{A}(\theta_0)\Psi^{-1}\sqrt{n}\frac{\partial Q_n\left(\theta_0\right)}{\partial \theta} + o_p(1),$$

and using the second maintained assumption,

$$\sqrt{n}\boldsymbol{a}(\hat{\theta}) \xrightarrow{d} N\big(\mathbf{0}, \boldsymbol{A}\left(\theta_0\right)\Psi^{-1}\Sigma\Psi^{-1}\boldsymbol{A}\left(\theta_0\right)'\big) = N\big(\mathbf{0}, \boldsymbol{A}\left(\theta_0\right)\Sigma^{-1}\boldsymbol{A}\left(\theta_0\right)'\big).$$

It follows that

$$\sqrt{n}\boldsymbol{a}(\hat{\theta})'\big(\boldsymbol{A}\left(\theta_0\right)\Sigma^{-1}\boldsymbol{A}\left(\theta_0\right)'\big)^{-1}\sqrt{n}\boldsymbol{a}(\hat{\theta}) \xrightarrow{d} \chi^2_{\#\boldsymbol{a}}.$$

If we have consistent estimators of $\boldsymbol{A}\left(\theta_0\right)$ and Σ, then limit results for continuous functions imply that

$$\sqrt{n}\boldsymbol{a}(\hat{\theta})'\big(\boldsymbol{A}(\hat{\theta})\hat{\Sigma}^{-1}\boldsymbol{A}(\hat{\theta})'\big)^{-1}\sqrt{n}\boldsymbol{a}(\hat{\theta}) \xrightarrow{d} \chi^2_{\#\boldsymbol{a}}$$

as well. This is the Wald statistic.

Under regularity conditions, $\boldsymbol{A}(\hat{\theta})$ will do for $\boldsymbol{A}(\hat{\theta}_0)$. With m-estimators, $\hat{\Sigma}$ can be estimated via outer product of gradients or (with ML) information, and in the case of GMM, the obvious estimator is $\hat{\Sigma} = \big(\frac{1}{n}\sum_{i=1}^n \frac{\partial g(W_i,\hat{\theta})}{\partial \theta}\big)'\hat{\boldsymbol{S}}^{-1}\big(\frac{1}{n}\sum_{i=1}^n \frac{\partial g(W_i,\hat{\theta})}{\partial \theta}\big)$ with $\hat{\boldsymbol{S}}$ as in our linear GMM section.

(ii) Preliminaries to the next two statistics

The next two statistics are closely related to the stochastic objects $\sqrt{n}\gamma_n$ and $\sqrt{n}(\tilde{\theta} - \theta_0)$, and so we will start by discovering asymptotic approximations for these two objects. (See (**??**) below for where we want to get.) This requires a few individually familiar but jointly intricate arguments.

First, the constrained optimization problem can be written out with a Lagrangian, in which case its first-order condition (rescaled by $\sqrt{n}$) is

$$\sqrt{n}\frac{\partial Q_n(\tilde{\theta})}{\partial \theta} + \sqrt{n}\boldsymbol{A}(\tilde{\theta})'\boldsymbol{\gamma}_n = \mathbf{0} \tag{2}$$

$$\sqrt{n}\boldsymbol{a}(\tilde{\theta}) = \mathbf{0}. \tag{3}$$

Next, the third maintained assumption yields

$$\sqrt{n}\boldsymbol{a}(\tilde{\theta}) = \boldsymbol{A}(\theta_0)\sqrt{n}\big(\tilde{\theta} - \theta_0\big) + o_p(1) \tag{4}$$

in precise analogy to (**??**).

Next, a Taylor expansion of $\frac{\partial Q_n(\theta)}{\partial \theta}$ about θ_0 yields

$$\sqrt{n}\frac{\partial Q_n(\tilde{\theta})}{\partial \theta} = \underbrace{\sqrt{n}\frac{\partial Q_n(\theta_0)}{\partial \theta}}_{\xrightarrow{d} N(\mathbf{0},\Sigma)} + \sqrt{n}\underbrace{\frac{\partial^2 Q_n(\theta_0)}{\partial \theta \partial \theta'}}_{\Psi}\big(\tilde{\theta} - \theta_0\big) + o_p(1). \tag{5}$$

The second and third maintained assumption now imply that $\sqrt{n}\frac{\partial Q_n(\tilde{\theta})}{\partial \theta}$, and hence $\sqrt{n}\boldsymbol{\gamma}_n$, are of order $O_p(1)$. This, in turn, allows us to write

$$\boldsymbol{A}(\tilde{\theta})'\sqrt{n}\boldsymbol{\gamma}_n = \boldsymbol{A}(\theta_0)'\sqrt{n}\boldsymbol{\gamma}_n + \big(\boldsymbol{A}(\tilde{\theta}) - \boldsymbol{A}(\theta_0)\big)'\sqrt{n}\boldsymbol{\gamma}_n = \boldsymbol{A}(\theta_0)'\sqrt{n}\boldsymbol{\gamma}_n + o_p(1) \tag{6}$$

by similar arguments as before.

Now comes the collecting of terms. Note right away that (**??**) and (**??**) jointly imply[10]

$$\boldsymbol{A}(\theta_0)\sqrt{n}\big(\tilde{\theta} - \theta_0\big) = o_p(1). \tag{7}$$

However, we can also combine (**??**), (**??**), and (**??**) to write

$$\sqrt{n}\boldsymbol{\Psi}\big(\tilde{\theta} - \theta_0\big) + \sqrt{n}\boldsymbol{A}(\theta_0)'\boldsymbol{\gamma}_n = -\sqrt{n}\frac{\partial Q_n(\theta_0)}{\partial \theta} + o_p(1). \tag{8}$$

Equations (**??**)-(**??**) jointly characterize the objects we are after – in particular, we have as many equations as unknowns. The problem is that the characterization is as yet implicit. We will have to solve this by mechanically resolving the vector equations. To keep expressions reasonably short, we introduce the shorthand $\boldsymbol{A} \equiv \boldsymbol{A}(\theta_0)$ and collect (**??**)-(**??**) into

$$\begin{bmatrix} \Psi & \boldsymbol{A}' \\ \boldsymbol{A} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \sqrt{n}(\tilde{\theta} - \theta_0) \\ \sqrt{n}\boldsymbol{\gamma}_n \end{bmatrix} = \begin{bmatrix} -\sqrt{n}\frac{\partial Q_n(\theta_0)}{\partial \theta} \\ \mathbf{0} \end{bmatrix} + o_p(1).$$

Now plug into the general formula for partitioned matrix inversion (no need to memorize this)

$$\begin{bmatrix} \boldsymbol{M}_{11} & \boldsymbol{M}_{12} \\ \boldsymbol{M}_{21} & \boldsymbol{M}_{22} \end{bmatrix}^{-1}$$

$$= \left[\begin{array}{c|c} \boldsymbol{M}_{11}^{-1} + \boldsymbol{M}_{11}^{-1}\boldsymbol{M}_{12}\big(\boldsymbol{M}_{22} - \boldsymbol{M}_{21}\boldsymbol{M}_{11}^{-1}\boldsymbol{M}_{12}\big)^{-1}\boldsymbol{M}_{21}\boldsymbol{M}_{11}^{-1} & -\boldsymbol{M}_{11}^{-1}\boldsymbol{M}_{12}\big(\boldsymbol{M}_{22} - \boldsymbol{M}_{21}\boldsymbol{M}_{11}^{-1}\boldsymbol{M}_{12}\big)^{-1} \\ \hline -\big(\boldsymbol{M}_{22} - \boldsymbol{M}_{21}\boldsymbol{M}_{11}^{-1}\boldsymbol{M}_{12}\big)^{-1}\boldsymbol{M}_{21}\boldsymbol{M}_{11}^{-1} & \big(\boldsymbol{M}_{22} - \boldsymbol{M}_{21}\boldsymbol{M}_{11}^{-1}\boldsymbol{M}_{12}\big)^{-1} \end{array}\right]$$

[10] Note the proper geometric interpretation of (**??**): It does not say that either $\boldsymbol{A}(\theta_0)$ or $\sqrt{n}(\tilde{\theta} - \theta_0)$ vanish – they don't – but that the rows of $\boldsymbol{A}(\theta_0)$ are asymptotically orthogonal to $\sqrt{n}(\tilde{\theta} - \theta_0)$.

and modestly rearrange to find

$$\sqrt{n}\begin{bmatrix} \tilde{\theta}-\theta_0 \\ \boldsymbol{\gamma}_n \end{bmatrix} = \begin{bmatrix} -\Psi^{-1}+\Psi^{-1}\boldsymbol{A}'\left(\boldsymbol{A}\Psi^{-1}\boldsymbol{A}'\right)^{-1}\boldsymbol{A}\Psi^{-1} \\ -\left(\boldsymbol{A}\Psi^{-1}\boldsymbol{A}'\right)^{-1}\boldsymbol{A}\Psi^{-1} \end{bmatrix}\sqrt{n}\frac{\partial Q_n(\theta_0)}{\partial\theta}+o_p(1). \tag{9}$$

We will now use these approximations to derive the next two test statistics.

(iii) LM Statistic

Use (**??**) to write

$$\begin{aligned}
\sqrt{n}\boldsymbol{\gamma}_n &= -\left(\boldsymbol{A}\Psi^{-1}\boldsymbol{A}'\right)^{-1}\boldsymbol{A}\Psi^{-1}\sqrt{n}\frac{\partial Q_n(\theta_0)}{\partial\theta}+o_p(1) \\
&\xrightarrow{d} N\left(\mathbf{0},\left(\boldsymbol{A}\Psi^{-1}\boldsymbol{A}'\right)^{-1}\boldsymbol{A}\Psi^{-1}\Sigma\Psi^{-1}\boldsymbol{A}'\left(\boldsymbol{A}\Psi^{-1}\boldsymbol{A}'\right)^{-1}\right) \\
&= N\left(\mathbf{0},\left(\boldsymbol{A}\Sigma^{-1}\boldsymbol{A}'\right)^{-1}\boldsymbol{A}\Sigma^{-1}\Sigma\Sigma^{-1}\boldsymbol{A}'\left(\boldsymbol{A}\Sigma^{-1}\boldsymbol{A}'\right)^{-1}\right) \\
&= N\left(\mathbf{0},\left(\boldsymbol{A}\Sigma^{-1}\boldsymbol{A}'\right)^{-1}\right).
\end{aligned}$$

It follows that if we square $\sqrt{n}\boldsymbol{\gamma}_n$ and divide by its variance, the resulting quantity

$$\sqrt{n}\boldsymbol{\gamma}_n'\boldsymbol{A}\Sigma^{-1}\boldsymbol{A}'\sqrt{n}\boldsymbol{\gamma}_n$$

converges to $\chi^2(\#\boldsymbol{a})$. This expression uses unknown population quantities, but we can now define a feasible statistic with the same asymptotic distribution by estimating $\boldsymbol{A}\Sigma^{-1}\boldsymbol{A}'$ with $\boldsymbol{A}(\tilde{\theta})\tilde{\Sigma}^{-1}\boldsymbol{A}(\tilde{\theta})'$, where $\tilde{\Sigma}$ is an estimator of Σ that falls out of computing the constrained estimation problem. (Any consistent estimator would do, but this one is simple because we already computed $\tilde{\theta}$.) Recalling that $\sqrt{n}\frac{\partial Q_n(\tilde{\theta})}{\partial\theta}+\sqrt{n}\boldsymbol{A}(\tilde{\theta})'\boldsymbol{\gamma}_n=\mathbf{0}$, we can finally write

$$LM \equiv \sqrt{n}\boldsymbol{\gamma}_n'\boldsymbol{A}(\tilde{\theta})\tilde{\Sigma}^{-1}\boldsymbol{A}(\tilde{\theta})'\sqrt{n}\boldsymbol{\gamma}_n = n\frac{\partial Q_n(\tilde{\theta})}{\partial\theta}'\tilde{\Sigma}^{-1}\frac{\partial Q_n(\tilde{\theta})}{\partial\theta}$$

and thus established the claim.

(iv) LR Statistic

By a Mean Value Theorem,

$$Q_n(\tilde{\theta}) = Q_n(\hat{\theta})+\frac{\partial Q_n(\hat{\theta})}{\partial\theta}\left(\tilde{\theta}-\hat{\theta}\right)+\frac{1}{2}\left(\tilde{\theta}-\hat{\theta}\right)'\frac{\partial^2 Q_n(\bar{\theta})}{\partial\theta\partial\theta'}\left(\tilde{\theta}-\hat{\theta}\right),$$

but $\frac{\partial Q_n(\hat{\theta})}{\partial\theta}=0$ and $\frac{\partial^2 Q_n(\bar{\theta})}{\partial\theta\partial\theta'}\xrightarrow{p}\boldsymbol{H}(W_i,\theta_0)$ by similar arguments to before. It follows that

$$\begin{aligned}
2n\left(Q_n(\hat{\theta})-Q_n(\tilde{\theta})\right) &= -\sqrt{n}\left(\tilde{\theta}-\hat{\theta}\right)'\left(\Psi+o_p(1)\right)\sqrt{n}\left(\tilde{\theta}-\hat{\theta}\right) \\
&= -\sqrt{n}\left(\tilde{\theta}-\hat{\theta}\right)'\Psi\sqrt{n}\left(\tilde{\theta}-\hat{\theta}\right)+o_p(1),
\end{aligned}$$

where we can move the $o_p(1)$ outside the quadratic form because $o_p(1)\times O_p(1)=o_p(1)$. On the other

hand, we know that

$$\begin{aligned}
& \sqrt{n}\big(\tilde{\theta}-\hat{\theta}\big) \\
= \;& \sqrt{n}\big(\tilde{\theta}-\theta_0\big)-\sqrt{n}\big(\hat{\theta}-\theta_0\big) \\
= \;& -\big(\Psi^{-1}-\Psi^{-1}\boldsymbol{A}'(\boldsymbol{A}\Psi^{-1}\boldsymbol{A}')^{-1}\boldsymbol{A}\Psi^{-1}\big)\sqrt{n}\frac{\partial Q_n(\theta_0)}{\partial\theta}-\left(-\Psi^{-1}\sqrt{n}\frac{\partial Q_n(\theta_0)}{\partial\theta}\right)+o_p(1) \\
= \;& \Psi^{-1}\boldsymbol{A}'\big(\boldsymbol{A}\Psi^{-1}\boldsymbol{A}'\big)^{-1}\boldsymbol{A}\Psi^{-1}\sqrt{n}\frac{\partial Q_n(\theta_0)}{\partial\theta}+o_p(1),
\end{aligned}$$

hence

$$\begin{aligned}
& 2n\big(Q_n(\hat{\theta})-Q_n(\tilde{\theta})\big) \\
= \;& -\left(\sqrt{n}\frac{\partial Q_n(\theta_0)}{\partial\theta}\right)'\big(\Psi^{-1}\boldsymbol{A}'(\boldsymbol{A}\Psi^{-1}\boldsymbol{A}')^{-1}\boldsymbol{A}\Psi^{-1}\big)'\Psi\Psi^{-1}\boldsymbol{A}'(\boldsymbol{A}\Psi^{-1}\boldsymbol{A}')^{-1}\boldsymbol{A}\Psi^{-1}\sqrt{n}\frac{\partial Q_n(\theta_0)}{\partial\theta}+o_p(1) \\
= \;& -\left(\sqrt{n}\frac{\partial Q_n(\theta_0)}{\partial\theta}\right)'\Psi^{-1}\boldsymbol{A}'(\boldsymbol{A}\Psi^{-1}\boldsymbol{A}')^{-1}\boldsymbol{A}\Psi^{-1}\sqrt{n}\frac{\partial Q_n\,(\theta_0)}{\partial\theta}+o_p(1) \\
= \;& \left(\sqrt{n}\frac{\partial Q_n(\theta_0)}{\partial\theta}\right)'\Sigma^{-1}\boldsymbol{A}'(\boldsymbol{A}\Sigma^{-1}\boldsymbol{A}')^{-1}\boldsymbol{A}\Sigma^{-1}\sqrt{n}\frac{\partial Q_n(\theta_0)}{\partial\theta}+o_p(1).
\end{aligned}$$

This is looking good: $\sqrt{n}\frac{\partial Q_n(\theta_0)}{\partial\theta}\xrightarrow{d}N(\mathbf{0},\Sigma)$ by assumption, hence

$$\boldsymbol{A}\Sigma^{-1}\sqrt{n}\frac{\partial Q_n(\theta_0)}{\partial\theta}\xrightarrow{d}N\big(\mathbf{0},\boldsymbol{A}\Sigma^{-1}\Sigma\Sigma^{-1}\boldsymbol{A}'\big)=N\big(\mathbf{0},\boldsymbol{A}\Sigma^{-1}\boldsymbol{A}'\big).$$

This asymptotic variance cancels against the central term of the quadratic form, and hence we are looking at the norm of a $\#\boldsymbol{a}$-dimensional, standard normal vector. You will by now know what that means:

$$LR\equiv 2n\big(Q_n(\hat{\theta})-Q_n(\tilde{\theta})\big)\xrightarrow{d}\chi^2_{\#\boldsymbol{a}}.$$

That's it! Again, we state without proof that the three statistics do not only converge to the same distribution, but are asymptotically equivalent, i.e. their numerical difference converges to zero. Asymptotically, they therefore identify the same realizations as outliers. However, it is well known that in finite samples, the three statistics can differ significantly!

A Geometric Interpretation

Here is an attempt to make sense of some of the algebra. By assumption, we know the asymptotic behavior of $\sqrt{n}\big(\hat{\theta}-\theta_0\big)$; we also need to understand the behavior of $\sqrt{n}\big(\tilde{\theta}-\theta_0\big)$ and $\sqrt{n}\boldsymbol{\gamma}_n$. This is done by a decomposition. For simplicity, assume that: (i) $\Theta\subset\mathbb{R}^2$, so the optimization problem can be depicted by isoquants in $\mathbb{R}^2$; (ii) Ψ is the negative identity matrix, thus around its maximum, the population objective function is approximated by a sphere; (iii) we test only one restriction, thus $\boldsymbol{a}$ is scalar valued, and the Jacobian $\boldsymbol{A}$ becomes a row vector corresponding to gradient ∇a. Recall that

because a is continuously differentiable, its graph is now a smooth curve; in particular, it has a unique tangent hyperplane (actually a straight line) at every point on its graph, and the gradient ∇a can be interpreted as orthogonal vector of this hyperplane. (More generally, the graph of $\boldsymbol{a}$ is a differentiable manifold and therefore is locally approximated by linear subspaces.)

The two equations that we derived and then collected in matrix form specialize to

$$\begin{aligned} \nabla a\left(\theta_0\right)' \sqrt{n}\left(\tilde{\theta}-\theta_0\right) &\approx 0 \\ -\sqrt{n}\left(\tilde{\theta}-\theta_0\right)+\nabla a\left(\theta_0\right) \sqrt{n} \gamma_n &\approx -\sqrt{n} \frac{\partial Q_n\left(\theta_0\right)}{\partial \theta}, \end{aligned}$$

which now are just three equations in scalars (the first equation is in scalars, the second one collects two such equations) corresponding to unknowns $(\tilde{\theta}, \gamma_n) \in \mathbb{R}^3$, thus we can solve for the unknowns. To get some more intuition, note that the first maintained assumption and some rearrangement transform the second equation above into

$$\sqrt{n}\left(\tilde{\theta}-\theta_0\right)-\nabla a\left(\theta_0\right) \sqrt{n} \gamma_n \approx \sqrt{n}\left(\hat{\theta}-\theta_0\right).$$

Thus, the vector $\sqrt{n}\left(\hat{\theta}-\theta_0\right)$ is decomposed into $\sqrt{n}\left(\tilde{\theta}-\theta_0\right)$ and $-\nabla a\left(\theta_0\right) \sqrt{n} \boldsymbol{\gamma}_n$. This decomposition is orthogonal: The vector $\nabla a\left(\theta_0\right) \sqrt{n} \gamma_n$ is just the (rescaled) gradient of $a\left(\theta_0\right)$ times a number γ_n, the Lagrange multiplier on the constraint. $\sqrt{n}\left(\tilde{\theta}-\theta_0\right)$, on the other hand, is orthogonal to said gradient by the first equation, reflecting the geometric fact that both θ_0 and $\tilde{\theta}$ are on the null, thus $\left(\tilde{\theta}-\theta_0\right)$ is orthogonal to the null's gradient. (If you have no visual intuition for this, please check the Wikipedia entry on Lagrange multipliers for two nice diagrams and remember that because of our choice of Ψ, the blue isoquants in the picture are circles.)

The decomposition is orthogonal because we set $\Psi=-\boldsymbol{I}$, thus isoquants of the objective function are circles. More generally, $\sqrt{n}\left(\tilde{\theta}-\theta_0\right)$ is the projection of $\sqrt{n}\left(\hat{\theta}-\theta_0\right)$ onto the linearized constraint set, i.e. it is the closest point to $\sqrt{n}\left(\hat{\theta}-\theta_0\right)$ on the constraint set, where closest is defined in terms of Ψ^{-1}. $\nabla a\left(\theta_0\right) \sqrt{n} \gamma_n$ is the residual from the projection.

8 Worked Examples

8.1 Estimating a Poisson Distribution

This is a somewhat easy example that we will use to work through application of the ML theorems in detail. (It also reproduces an old exam question.) Let y_i be distributed i.i.d. Poisson with true parameter value $\lambda_0 > 0$. Recall the Poisson distribution has probability mass function $\Pr(y_i = k) = \lambda^k e^{-\lambda}/k!$, and mean and variance λ.

It's intuitively obvious that we want to estimate λ by $\bar{y}$, and the Lindeberg-Lévy CLT immediately yields $\sqrt{n}(\bar{y} - \lambda_0) \xrightarrow{d} N(0, \lambda_0)$. To analyze this as an example of Maximum Likelihood, write

$$\begin{aligned}
Q_n(\lambda) &= \frac{1}{n}\sum_i \left(y_i \log\lambda - \lambda - \log\left(y_i!\right)\right) \\
\frac{\partial Q_n(\lambda)}{\partial\lambda} &= \frac{1}{n}\sum_i \left(\frac{y_i}{\lambda} - 1\right) \\
\frac{\partial^2 Q_n(\lambda)}{\partial\lambda^2} &= -\frac{1}{n}\sum_i \frac{y_i}{\lambda^2} \\
Q(\lambda) &= \mathbb{E}\left(y_i \log\lambda - \lambda - \log\left(y_i!\right)\right) \\
\frac{\partial Q(\lambda)}{\partial\lambda} &= \mathbb{E}\left(\frac{y_i}{\lambda} - 1\right) \\
\frac{\partial^2 Q(\lambda)}{\partial\lambda^2} &= -\mathbb{E}\frac{y_i}{\lambda^2}.
\end{aligned}$$

Inspection of $\frac{\partial^2 Q_n(\lambda)}{\partial\lambda^2}$ reveals strict concavity, so that $\hat{\lambda}_{ML}$ is characterized by a FOC. In particular,

$$\frac{1}{n}\sum_i \left(\frac{y_i}{\lambda} - 1\right) \stackrel{!}{=} 0$$

can be solved for $\hat{\lambda}_{ML} = \bar{y}$. Next, we can apply a consistency theorem for extremum estimators. Pointwise convergence of Q_n to Q is clear. The parameter space is not compact, but Q_n is strictly concave; because λ is a scalar, we can therefore invoke the relatively simple convergence theorem for concave objectives from a homework. (The version using compactness of Θ would work only after bounding λ from above but also from below by a strictly positive number below; else, uniform convergence is simply not true.)

We finally verify the conditions of Theorem **??**. 1. was established above, 2. holds if $\lambda > 0$, 3. is evident from the above displays. 4. holds because y_i is i.i.d. with expectation and variance λ_0, hence

$$\sqrt{n}\frac{\partial Q_n(\lambda_0)}{\partial\lambda} = \sqrt{n}\frac{1}{n}\sum_i \left(\frac{y_i}{\lambda_0} - 1\right) \xrightarrow{d} N(0, \lambda_0^{-1}).$$

5. follows by inspection of the display. To verify 6., note that $\frac{\partial^2 Q(\lambda_0)}{\partial\lambda^2} = -\frac{\lambda_0}{\lambda_0^2} = -\lambda_0^{-1}$. 7. follows because $\lambda_0 > 0$. The theorem applies, and we can substitute into its conclusion to find

$$\sqrt{n}\big(\hat{\lambda} - \lambda_0\big) \xrightarrow{d} N\big(0, (-\lambda_0^{-1})^{-1}\lambda_0^{-1}(-\lambda_0^{-1})^{-1}\big) = N(0, \lambda_0)$$

as expected.

Our results do not apply if $\lambda_0 = 0$ because interiority is then violated. Note that in this special case, we have not only a failure of proof but a failure of result: The distribution of both y_i and $\bar{y}$ will be degenerate and completely concentrated at 0. As one might expect from this observation, the above result is also not uniformly true as $\lambda_0 \to 0$; it fails along drifting parameters of the form $\lambda_n = \gamma/\sqrt{n}$. For a red flag that points at this non-uniformity, note in particular that the Hessian, which is really just a second derivative here, approaches singularity as $\lambda_0 \to 0$.

8.2 Maximum Likelihood Analysis of Linear Regression

8.2.1 Single Equation

We next reconsider the linear regression model:

$$\begin{aligned} y_i &= \boldsymbol{x}_i'\boldsymbol{\beta}_0 + \varepsilon_i \\ \varepsilon_i &\sim N(0, \sigma^2). \end{aligned}$$

We already analyzed identification. We will maximize the conditional log likelihood

$$\log f(y_i|\boldsymbol{x}_i, \boldsymbol{\theta}) = -\frac{1}{2}\log 2\pi - \frac{1}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\left(y_i - \boldsymbol{x}_i'\boldsymbol{\beta}\right)^2,$$

hence the ML estimator $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\beta}}, \hat{\sigma}^2)$ solves

$$\begin{aligned} \hat{\boldsymbol{\theta}} &= \arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} Q_n(\boldsymbol{\beta}, \sigma^2), \\ Q_n(\boldsymbol{\beta}, \sigma^2) &= \sum_{i=1}^n \Big(-\frac{1}{2}\log 2\pi - \frac{1}{2}\log\sigma^2 - \frac{1}{2\sigma^2}(y_i - \boldsymbol{x}_i'\boldsymbol{\beta})^2\Big). \end{aligned}$$

Likelihood functions frequently have separability properties that make it convenient to solve this by *concentrating out*, that is, by solving for some parameters first and then extremizing the value function over the other parameters. The optimized-out objective function is often called *concentrated* with respect to the optimized-out parameter. In the specific example, let's first find $\hat{\boldsymbol{\beta}}$, which must solve

$$\hat{\boldsymbol{\beta}} = \arg\max_\beta \Big\{-\sum_{i=1}^n (y_i - \boldsymbol{x}_i'\boldsymbol{\beta})^2\Big\} = \arg\min_\beta \sum_{i=1}^n (y_i - \boldsymbol{x}_i'\boldsymbol{\beta})^2,$$

so it is the least squares estimator. In a second step, we optimize the concentrated objective function to find that

$$\begin{aligned} \hat{\sigma}^2 &= \arg\max_{\sigma^2} \sum_{i=1}^n \Big(-\frac{1}{2}\log 2\pi - \frac{1}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\big(y_i - \boldsymbol{x}_i'\hat{\boldsymbol{\beta}}\big)^2\Big) \\ &= \arg\min_{\sigma^2} \Big\{n\log\sigma^2 + \frac{1}{\sigma^2}\sum_{i=1}^n \big(y_i - \boldsymbol{x}_i'\hat{\boldsymbol{\beta}}\big)^2\Big\}. \end{aligned}$$

This problem has FOC

$$\begin{aligned} &\frac{n}{\sigma^2} - \frac{1}{(\sigma^2)^2}\sum_{i=1}^n \big(y_i - \boldsymbol{x}_i'\hat{\boldsymbol{\beta}}\big)^2 = 0 \\ \Rightarrow\quad &\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n \big(y_i - \boldsymbol{x}_i'\hat{\boldsymbol{\beta}}\big)^2. \end{aligned}$$

The slight surprise here, if you haven't seen it before, is that the ML estimator $\hat{\sigma}^2$ does not feature the degrees-of-freedom-adjustment, i.e. it is not $\hat{\sigma}^2_{OLS} = \frac{1}{n-K}\sum_{i=1}^n \Big(y_i - \boldsymbol{x}_i'\hat{\boldsymbol{\beta}}\Big)^2$. On the one hand, this shows that $\hat{\sigma}^2_{OLS}$ is not as obviously optimal as you might have thought. On the other hand, it is

a reminder that ML estimators are, in general, biased (because it remains true that under assumptions maintained here, $\hat{\sigma}^2_{OLS}$ is unbiased).

For completeness, we observe that the estimator's asymptotic distribution can also be derived from our development for Maximum Likelihood estimators. In particular, defining $\boldsymbol{\theta} \equiv (\boldsymbol{\beta}', \sigma^2)'$, we have

$$\begin{aligned}
Q(\boldsymbol{\theta}) &= \mathbb{E}\Big(-\frac{1}{2}\log 2\pi - \frac{1}{2}\log\sigma^2 - \frac{1}{2\sigma^2}(y_i - \boldsymbol{x}_i'\boldsymbol{\beta})^2\Big) \\
\frac{\partial Q(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} &= \begin{bmatrix} \frac{1}{\sigma^2}\mathbb{E}\big(\boldsymbol{x}_i(y_i - \boldsymbol{x}_i'\boldsymbol{\beta})\big) \\ -\frac{1}{2\sigma^2} + \frac{1}{2\sigma^4}\mathbb{E}(y_i - \boldsymbol{x}_i'\boldsymbol{\beta})^2 \end{bmatrix} \\
\frac{\partial^2 Q(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}'} &= \left[\begin{array}{c|c} -\frac{1}{\sigma^2}\mathbb{E}(\boldsymbol{x}_i\boldsymbol{x}_i') & -\frac{1}{\sigma^4}\mathbb{E}\big(\boldsymbol{x}_i(y_i - \boldsymbol{x}_i'\boldsymbol{\beta})\big) \\ \hline -\frac{1}{\sigma^4}\mathbb{E}\big((y_i - \boldsymbol{x}_i'\boldsymbol{\beta})\boldsymbol{x}_i'\big) & \frac{1}{2\sigma^4} - \frac{1}{\sigma^6}\mathbb{E}(y_i - \boldsymbol{x}_i'\boldsymbol{\beta})^2 \end{array}\right].
\end{aligned}$$

As a small exercise to check my algebra, you should be able to convince yourself that: (i) Assuming identification, $\frac{\partial Q(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$ is the zero vector iff evaluated at the true values of $\boldsymbol{\beta}$ and σ^2; (ii) without identification, it is the zero vector on a linear subspace, i.e. the likelihood has a ridge.

Fortunately, to compute the asymptotic variance, we need to invert the last matrix only if evaluated at the true parameter values, where it much simplifies:

$$\begin{aligned}
\frac{\partial^2 Q(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}'} &= \left[\begin{array}{c|c} -\frac{1}{\sigma_0^2}\mathbb{E}(\boldsymbol{x}_i\boldsymbol{x}_i') & -\frac{1}{\sigma_0^4}\mathbb{E}\big(\boldsymbol{x}_i(y_i - \boldsymbol{x}_i'\boldsymbol{\beta}_0)\big) \\ \hline -\frac{1}{\sigma_0^4}\mathbb{E}\big((y_i - \boldsymbol{x}_i'\boldsymbol{\beta}_0)\boldsymbol{x}_i'\big) & \frac{1}{2\sigma_0^4} - \frac{1}{\sigma_0^6}\mathbb{E}(y_i - \boldsymbol{x}_i'\boldsymbol{\beta}_0)^2 \end{array}\right] \\
&= \left[\begin{array}{c|c} -\frac{1}{\sigma_0^2}\mathbb{E}(\boldsymbol{x}_i\boldsymbol{x}_i') & \mathbf{0} \\ \hline \mathbf{0} & -\frac{1}{2\sigma_0^4} \end{array}\right] \\
\Longrightarrow -\left(\frac{\partial^2 Q(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}'}\right)^{-1} &= \left[\begin{array}{c|c} \sigma_0^2\big(\mathbb{E}\boldsymbol{x}_i\boldsymbol{x}_i'\big)^{-1} & \mathbf{0} \\ \hline \mathbf{0} & 2\sigma_0^4 \end{array}\right].
\end{aligned}$$

We see that $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$ are asymptotically independent.[11] While we know that the asymptotic variance expression for $\hat{\boldsymbol{\beta}}$ holds under much weaker conditions than we here imposed, the one for $\hat{\sigma}^2$ does not (and in general exists only if ε_i has a finite fourth moment).

We conclude by verifying the information matrix equality. The top left submatrix equals

$$\sigma^{-4}\mathbb{E}\big(\boldsymbol{x}_i(y_i - \boldsymbol{x}_i'\boldsymbol{\beta})^2\boldsymbol{x}_i'\big) \underset{\boldsymbol{\theta}=\boldsymbol{\theta}_0}{=} \sigma_0^{-2}\mathbb{E}\big(\boldsymbol{x}_i\boldsymbol{x}_i'\big)$$

using the Law of Iterated Expectations and $\mathbb{E}\big((y_i - \boldsymbol{x}_i'\boldsymbol{\beta}_0)^2|\boldsymbol{x}_i\big) = \sigma_0^2$. The bottom right entry equals

$$\frac{1}{4\sigma^4} + \frac{1}{4\sigma^8}\mathbb{E}(y_i - \boldsymbol{x}_i'\boldsymbol{\beta})^4 - \frac{1}{2\sigma^6}\mathbb{E}(y_i - \boldsymbol{x}_i'\boldsymbol{\beta})^2 \underset{\boldsymbol{\theta}=\boldsymbol{\theta}_0}{=} \frac{1}{4\sigma_0^4} + \frac{3}{4\sigma_0^4} + \frac{1}{\sigma_0^4} = \frac{1}{2\sigma_0^4},$$

using that the fourth central moment of the Normal equals $3\sigma^4$. For the off-diagonal elements, we have

$$-\frac{1}{2\sigma^4}\mathbb{E}\big(\boldsymbol{x}_i(y_i - \boldsymbol{x}_i'\boldsymbol{\beta})\big) + \frac{1}{2\sigma^6}\mathbb{E}\big(\boldsymbol{x}_i(y_i - \boldsymbol{x}_i'\boldsymbol{\beta})^3\big) \underset{\boldsymbol{\theta}=\boldsymbol{\theta}_0}{=} \mathbf{0}.$$

[11] In fact, under the normality assumption, they are finite sample independent, which is essential in establishing the exact distributions of t- and F-statistic.

8.2.2 Multiple Equations

We next derive the estimator with many equations. Write

$$\boldsymbol{y}_i = \boldsymbol{\Pi}_0' \boldsymbol{x}_i + \boldsymbol{\nu}_i,$$

where

$$\boldsymbol{y}_i = \begin{bmatrix} y_{i1} \\ y_{i2} \\ \vdots \\ y_{iM} \end{bmatrix}, \boldsymbol{\Pi}_0' = \begin{bmatrix} \boldsymbol{\pi}_{01}' \\ \boldsymbol{\pi}_{02}' \\ \vdots \\ \boldsymbol{\pi}_{0M}' \end{bmatrix}, \boldsymbol{\nu}_i = \begin{bmatrix} \nu_{i1} \\ \nu_{i2} \\ \vdots \\ \nu_{iM} \end{bmatrix}.$$

To recall, this stacks M equations, each of which regresses a different outcome y_{im} on the same regressors. Recall our assumptions:

- $\{y_i, \boldsymbol{x}_i\}$ is i.i.d.,
- $\mathbb{E}(\boldsymbol{\nu}_i \otimes \boldsymbol{x}_i) = \mathbf{0}$,
- $\mathbb{E}(\boldsymbol{x}_i \boldsymbol{x}_i')$ is nonsingular,
- $\mathbb{E}(\boldsymbol{\nu}_i \boldsymbol{\nu}_i' | \boldsymbol{x}_i) = \boldsymbol{\Omega}_0$ positive definite.

These assumptions suffice to estimate the model by GMM, leading to $\hat{\boldsymbol{\Pi}}_{OLS} = \left(\sum_i \boldsymbol{x}_i \boldsymbol{x}_i'\right)^{-1} \sum_i \boldsymbol{x}_i \boldsymbol{y}_i'$. To estimate the model by ML, we also assume that $\boldsymbol{\nu}_i | \boldsymbol{x}_i \sim N(\mathbf{0}, \boldsymbol{\Omega}_0)$.

Using the closed-form expression for multivariate normal densities, we find that the conditional log likelihood function is

$$\log f(\boldsymbol{y}_i | \boldsymbol{x}_i; \boldsymbol{\Pi}, \boldsymbol{\Omega}) = -\frac{M}{2} \log 2\pi + \frac{1}{2} \log \left| \boldsymbol{\Omega}^{-1} \right| - \frac{1}{2} \left(\boldsymbol{y}_i - \boldsymbol{\Pi}' \boldsymbol{x}_i \right)' \boldsymbol{\Omega}^{-1} \left(\boldsymbol{y}_i - \boldsymbol{\Pi}' \boldsymbol{x}_i \right),$$

hence our objective function is

$$\begin{aligned}
Q_n(\boldsymbol{\Pi}, \boldsymbol{\Omega}) &= -\frac{M}{2} \log 2\pi + \frac{1}{2} \log \left| \boldsymbol{\Omega}^{-1} \right| - \frac{1}{2n} \sum_{i=1}^n \left(\boldsymbol{y}_i - \boldsymbol{\Pi}' \boldsymbol{x}_i \right)' \boldsymbol{\Omega}^{-1} \left(\boldsymbol{y}_i - \boldsymbol{\Pi}' \boldsymbol{x}_i \right) \\
&= -\frac{M}{2} \log 2\pi + \frac{1}{2} \log \left| \boldsymbol{\Omega}^{-1} \right| - \frac{1}{2n} \sum_{i=1}^n \operatorname{tr}\left(\left(\boldsymbol{y}_i - \boldsymbol{\Pi}' \boldsymbol{x}_i \right)' \boldsymbol{\Omega}^{-1} \left(\boldsymbol{y}_i - \boldsymbol{\Pi}' \boldsymbol{x}_i \right) \right) \\
&= -\frac{M}{2} \log 2\pi + \frac{1}{2} \log \left| \boldsymbol{\Omega}^{-1} \right| - \frac{1}{2n} \sum_{i=1}^n \operatorname{tr}\left(\boldsymbol{\Omega}^{-1} (\boldsymbol{y}_i - \boldsymbol{\Pi}' \boldsymbol{x}_i)(\boldsymbol{y}_i - \boldsymbol{\Pi}' \boldsymbol{x}_i)' \right) \\
&= -\frac{M}{2} \log 2\pi + \frac{1}{2} \log \left| \boldsymbol{\Omega}^{-1} \right| - \frac{1}{2} \operatorname{tr}\left(\frac{\boldsymbol{\Omega}^{-1}}{n} \sum_{i=1}^n (\boldsymbol{y}_i - \boldsymbol{\Pi}' \boldsymbol{x}_i)(\boldsymbol{y}_i - \boldsymbol{\Pi}' \boldsymbol{x}_i)' \right) \\
&\equiv -\frac{M}{2} \log 2\pi + \frac{1}{2} \log \left| \boldsymbol{\Omega}^{-1} \right| - \frac{1}{2} \operatorname{tr}\left(\boldsymbol{\Omega}^{-1} \hat{\boldsymbol{\Omega}}(\boldsymbol{\Pi}) \right),
\end{aligned}$$

where the last line defines $\hat{\boldsymbol{\Omega}}(\boldsymbol{\Pi})$ in analogy to $\hat{\sigma}^2$ in our warm-up example. (Hayashi's notation here anticipates that this guy is going to be our estimator, which we strictly speaking don't know yet.

Notice also the algebraic trick: A scalar, interpreted as $[1 \times 1]$-matrix, is its own trace, but then rules for manipulating trace operators apply.)

We first optimize with respect to $\boldsymbol{\Omega}$. A fact that you need not memorize is that if $\boldsymbol{A}$ and $\boldsymbol{B}$ are symmetric, positive definite, and conformable, then

$$\arg\max_{\boldsymbol{A}} \{\log|\boldsymbol{A}| - \mathrm{tr}(\boldsymbol{A}\boldsymbol{B})\} = \boldsymbol{B}^{-1}.$$

Thus, we find that $\boldsymbol{\Omega}^* = \hat{\boldsymbol{\Omega}}(\boldsymbol{\Pi})$.

We can now concentrate the objective function by plugging in the estimator. Then

$$\begin{aligned}
\hat{\boldsymbol{\Pi}}_{ML} &= \arg\max_{\boldsymbol{\Pi}} \Big\{ -\frac{M}{2}\log 2\pi + \frac{1}{2}\log\big|\hat{\boldsymbol{\Omega}}(\boldsymbol{\Pi})^{-1}\big| - \frac{1}{2}\mathrm{tr}\left(\boldsymbol{I}_M\right) \Big\} \\
&= \arg\max_{\boldsymbol{\Pi}} \Big\{ -\frac{M}{2}\log 2\pi - \frac{1}{2}\log\big|\hat{\boldsymbol{\Omega}}(\boldsymbol{\Pi})\big| - \frac{M}{2} \Big\} \\
&= \arg\min_{\boldsymbol{\Pi}} \ \big|\hat{\boldsymbol{\Omega}}(\boldsymbol{\Pi})\big| \\
&= \arg\min_{\boldsymbol{\Pi}} \ \left| \frac{1}{n}\sum_{i=1}^{n} \left(\boldsymbol{y}_i - \boldsymbol{\Pi}'\boldsymbol{x}_i\right)\left(\boldsymbol{y}_i - \boldsymbol{\Pi}'\boldsymbol{x}_i\right)' \right|.
\end{aligned}$$

To see that this is solved by the OLS estimator, write

$$\begin{aligned}
&\sum_{i=1}^{n} \left(\boldsymbol{y}_i - \boldsymbol{\Pi}'\boldsymbol{x}_i\right)\left(\boldsymbol{y}_i - \boldsymbol{\Pi}'\boldsymbol{x}_i\right)' \\
= \ &\sum_{i=1}^{n} \big(\boldsymbol{y}_i - \hat{\boldsymbol{\Pi}}'_{OLS}\boldsymbol{x}_i + \hat{\boldsymbol{\Pi}}'_{OLS}\boldsymbol{x}_i - \boldsymbol{\Pi}'\boldsymbol{x}_i\big)\big(\boldsymbol{y}_i - \hat{\boldsymbol{\Pi}}'_{OLS}\boldsymbol{x}_i + \hat{\boldsymbol{\Pi}}'_{OLS}\boldsymbol{x}_i - \boldsymbol{\Pi}'\boldsymbol{x}_i\big)' \\
= \ &\sum_{i=1}^{n} \big(\hat{\boldsymbol{\nu}}_i + (\hat{\boldsymbol{\Pi}}'_{OLS} - \boldsymbol{\Pi}')\boldsymbol{x}_i\big)\big(\hat{\boldsymbol{\nu}}_i + (\hat{\boldsymbol{\Pi}}'_{OLS} - \boldsymbol{\Pi}')\boldsymbol{x}_i\big)' \\
= \ &\sum_{i=1}^{n} \big(\hat{\boldsymbol{\nu}}_i\hat{\boldsymbol{\nu}}'_i + \hat{\boldsymbol{\nu}}_i\boldsymbol{x}'_i(\hat{\boldsymbol{\Pi}}_{OLS} - \boldsymbol{\Pi}) + (\hat{\boldsymbol{\Pi}}'_{OLS} - \boldsymbol{\Pi}')\boldsymbol{x}_i\hat{\boldsymbol{\nu}}'_i + (\hat{\boldsymbol{\Pi}}'_{OLS} - \boldsymbol{\Pi}')\boldsymbol{x}_i\boldsymbol{x}'_i(\hat{\boldsymbol{\Pi}}_{OLS} - \boldsymbol{\Pi})\big) \\
= \ &\sum_{i=1}^{n} \hat{\boldsymbol{\nu}}_i\hat{\boldsymbol{\nu}}'_i + \sum_{i=1}^{n} \big(\hat{\boldsymbol{\Pi}}'_{OLS} - \boldsymbol{\Pi}'\big)\boldsymbol{x}_i\boldsymbol{x}'_i\big(\hat{\boldsymbol{\Pi}}_{OLS} - \boldsymbol{\Pi}\big),
\end{aligned}$$

where the last step uses that from the geometry of OLS, $\sum_{i=1}^{n} \boldsymbol{x}_i\hat{\boldsymbol{\nu}}'_i = \boldsymbol{0}$. Now,

$$\sum_{i=1}^{n} \big(\hat{\boldsymbol{\Pi}}'_{OLS} - \boldsymbol{\Pi}'\big)\boldsymbol{x}_i\boldsymbol{x}'_i\big(\hat{\boldsymbol{\Pi}}_{OLS} - \boldsymbol{\Pi}\big)$$

is a quadratic form, and it is true that if $\boldsymbol{A}$ and $\boldsymbol{B}$ are positive semidefinite, then $|\boldsymbol{A} + \boldsymbol{B}| \geq |\boldsymbol{A}|$. The objective is therefore minimized by $\hat{\boldsymbol{\Pi}}_{OLS}$, which achieves value $\sum_{i=1}^{n} \hat{\boldsymbol{\nu}}_i\hat{\boldsymbol{\nu}}'_i$. Substituting into $\boldsymbol{\Omega}^* = \hat{\boldsymbol{\Omega}}(\boldsymbol{\Pi})$, it now follows that $\hat{\boldsymbol{\Omega}}_{ML} = \hat{\boldsymbol{\Omega}}(\hat{\boldsymbol{\Pi}}_{OLS})$.

We will not go through consistency and asymptotic normality of this. Our analysis of GMM implies that the assumptions used to construct this estimator, i.e. independent draws and normality, are not really needed for its consistency and asymptotic normality. In cases where these assumptions do not hold, we could therefore think of the OLS estimator through its ML justification as a pseudo-ML estimator, illustrating that pseudo-ML can be "successful."

8.3 Binary Response

Binary response models can generally be expressed in the form

$$y_i = \mathbf{1}\left\{\phi(\boldsymbol{x}_i, \varepsilon_i; \boldsymbol{\theta}_0) \geq 0\right\}$$

which is frequently specialized to

$$\begin{aligned} y_i &= \mathbf{1}\left\{\boldsymbol{x}_i'\boldsymbol{\beta}_0 - \varepsilon_i \geq 0\right\} \\ \iff \Pr(y_i = 1|\boldsymbol{x}_i) &= \Pr(\varepsilon_i \leq \boldsymbol{x}_i'\boldsymbol{\beta}_0) = F_\varepsilon\left(\boldsymbol{x}_i'\boldsymbol{\beta}_0\right). \end{aligned}$$

Assume that $\mathbb{E}\boldsymbol{x}_i\boldsymbol{x}_i'$ is nonsingular and that F_ε is strictly increasing, then the model is identified up to a scale normalization and (if $\boldsymbol{x}_i$ has a constant component) a location normalization. Other than that, different assumptions about F_ε lead to different models. Thus, assume that ε_i is logistically distributed, then we have the *logit model*

$$\Pr(y_i = 1|\boldsymbol{x}_i) = \frac{\exp(\boldsymbol{x}_i'\boldsymbol{\beta}_0)}{1 + \exp(\boldsymbol{x}_i'\boldsymbol{\beta}_0)}.$$

We then have log likelihood

$$\log f(y_i|\boldsymbol{x}_i; \boldsymbol{\theta}) = y_i \log \frac{\exp(\boldsymbol{x}_i'\boldsymbol{\beta})}{1 + \exp(\boldsymbol{x}_i'\boldsymbol{\beta})} + (1 - y_i) \log \frac{1}{1 + \exp(\boldsymbol{x}_i'\boldsymbol{\beta})},$$

and the maximum likelihood estimator is characterized as maximizer of

$$\begin{aligned} Q_n(\boldsymbol{\beta}) &= \frac{1}{n}\sum_i \left(y_i \log \frac{\exp(\boldsymbol{x}_i'\boldsymbol{\beta})}{1 + \exp(\boldsymbol{x}_i'\boldsymbol{\beta})} + (1 - y_i) \log \frac{1}{1 + \exp(\boldsymbol{x}_i'\boldsymbol{\beta})} \right) \\ &= \frac{1}{n}\sum_i \big(y_i \boldsymbol{x}_i'\boldsymbol{\beta} - y_i \log\left(1 + \exp(\boldsymbol{x}_i'\boldsymbol{\beta})\right) - (1 - y_i) \log\left(1 + \exp(\boldsymbol{x}_i'\boldsymbol{\beta})\right) \big) \\ &= \frac{1}{n}\sum_i \big(y_i \boldsymbol{x}_i'\boldsymbol{\beta} - \log\left(1 + \exp(\boldsymbol{x}_i'\boldsymbol{\beta})\right) \big). \end{aligned}$$

We can thus write

$$\frac{\partial Q_n(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \frac{1}{n}\sum_i \left(y_i \boldsymbol{x}_i - \frac{\exp(\boldsymbol{x}_i'\boldsymbol{\beta})}{1 + \exp(\boldsymbol{x}_i'\boldsymbol{\beta})} \boldsymbol{x}_i \right) = \frac{1}{n}\sum_i \big(y_i - F_\varepsilon\left(\boldsymbol{x}_i'\boldsymbol{\beta}\right) \big) \boldsymbol{x}_i$$

and

$$\frac{\partial^2 Q_n(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} = -\frac{1}{n}\sum_i F_\varepsilon\left(\boldsymbol{x}_i'\boldsymbol{\beta}\right) \left(1 - F_\varepsilon\left(\boldsymbol{x}_i'\boldsymbol{\beta}\right)\right) \boldsymbol{x}_i \boldsymbol{x}_i',$$

applying the formula $F(t) = e^t/(1 + e^t) \Rightarrow F'(t) = F(t)(1 - F(t))$ to $F_\varepsilon\left(\boldsymbol{x}_i\boldsymbol{\beta}\right) = \frac{\exp(\boldsymbol{x}_i'\boldsymbol{\beta})}{1+\exp(\boldsymbol{x}_i'\boldsymbol{\beta})}$.

We can now establish consistency and asymptotic normality as in the previous example. Note in particular that the Hessian of Q_n is negative definite and the sample criterion function therefore concave, which together with likelihood identification and pointwise consistency of Q_n for Q implies consistency even without compactness of Θ. Also, the last display above immediately gives uniform boundedness of the (sample and population) Hessian.

We close with two asides:

- The above algebra, with expectations replacing sample averages, yields

$$\begin{aligned}\frac{\partial Q(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} &= \mathbb{E}\big(\big(y_i - F_\varepsilon\left(\boldsymbol{x}_i'\boldsymbol{\beta}\right)\big)\boldsymbol{x}_i\big)\\ \frac{\partial^2 Q(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}'} &= -\mathbb{E}\big(F_\varepsilon\left(\boldsymbol{x}_i'\boldsymbol{\beta}\right)\left(1 - F_\varepsilon\left(\boldsymbol{x}_i'\boldsymbol{\beta}\right)\right)\boldsymbol{x}_i\boldsymbol{x}_i'\big).\end{aligned}$$

  At first glance, the information matrix equality (specifically, that the line marked (*) below equals minus the Hessian) may not appear obvious. However, write

$$\begin{aligned}& \mathbb{E}\big[\big(\big(y_i - F_\varepsilon\left(\boldsymbol{x}_i'\boldsymbol{\beta}_0\right)\big)\boldsymbol{x}_i\big)\big(\big(y_i - F_\varepsilon\left(\boldsymbol{x}_i'\boldsymbol{\beta}_0\right)\big)\boldsymbol{x}_i\big)'\big]\\ =\ & \mathbb{E}\big[\big(y_i - F_\varepsilon\left(\boldsymbol{x}_i'\boldsymbol{\beta}_0\right)\big)^2\boldsymbol{x}_i\boldsymbol{x}_i'\big] \quad (*)\\ =\ & \mathbb{E}\big[\mathbb{E}\big[\big(y_i - F_\varepsilon\left(\boldsymbol{x}_i'\boldsymbol{\beta}_0\right)\big)^2|\boldsymbol{x}_i\big]\boldsymbol{x}_i\boldsymbol{x}_i'\big]\\ =\ & \mathbb{E}\big[\mathrm{var}(y_i|\boldsymbol{x}_i)\boldsymbol{x}_i\boldsymbol{x}_i'\big]\\ =\ & \mathbb{E}\big[F_\varepsilon\left(\boldsymbol{x}_i'\boldsymbol{\beta}_0\right)\left(1 - F_\varepsilon\left(\boldsymbol{x}_i'\boldsymbol{\beta}_0\right)\right)\boldsymbol{x}_i\boldsymbol{x}_i'\big]\\ =\ & -\frac{\partial^2 Q(\boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}'},\end{aligned}$$

  where we used the Law of Iterated Expectations followed by our knowledge that, conditionally on $\boldsymbol{x}_i$, y_i is distributed Bernoulli with parameter $F_\varepsilon\left(\boldsymbol{x}_i'\boldsymbol{\beta}_0\right)$ and therefore has mean $F_\varepsilon\left(\boldsymbol{x}_i'\boldsymbol{\beta}_0\right)$ and variance $F_\varepsilon\left(\boldsymbol{x}_i'\boldsymbol{\beta}_0\right)\left(1 - F_\varepsilon\left(\boldsymbol{x}_i'\boldsymbol{\beta}_0\right)\right)$.

- This model can also be estimated by GMM. (It is just identified, so really we're just doing plain method of moments.) To do so, we need a function $\boldsymbol{g}$ of the dimensionality of $\boldsymbol{x}_i$ s.t. $\mathbb{E}\boldsymbol{g}(y_i, \boldsymbol{x}_i; \boldsymbol{\beta}_0) = 0$. As a general rule, if conditional expectations can be written out, they immediately give rise to such functions. In the present example, the Law of Iterated Expectations yields

$$\mathbb{E}\left(\boldsymbol{x}_i\left(y_i - F_\varepsilon\left(\boldsymbol{x}_i'\boldsymbol{\beta}_0\right)\right)\right) = \mathbb{E}\left(\boldsymbol{x}_i\mathbb{E}\left(y_i - F_\varepsilon\left(\boldsymbol{x}_i'\boldsymbol{\beta}_0\right)|\boldsymbol{x}_i\right)\right) = \mathbf{0}.$$

  so our method of moments estimator is defined by the sample analog of this,

$$\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i\big(y_i - F_\varepsilon(\boldsymbol{x}_i'\hat{\boldsymbol{\beta}})\big) = \mathbf{0}.$$

  We see that the estimators algebraically coincide. That is, the natural GMM estimator uses the score equations as moment conditions and therefore is exactly the ML estimator. Hence, this estimator is efficient in the strong sense of replicating the asymptotic variance of ML.

8.4 Tobit Type II and Heckman Two-Step

We next analyze the Type II Tobit and the Heckman Two-Step ("Heckit") strategy of adjusting for nonignorable censoring. This development can be generalized in many ways and stands at the beginning of a huge literature in applied econometrics.

8.4.1 The Model

The following model is known as Type II Tobit. (To economize on subscripts, I drop the "0" subscript for true parameter value.)

$$\begin{aligned} y_{1i}^* &= \boldsymbol{x}_{1i}'\boldsymbol{\beta}_1 + \varepsilon_{1i} \\ y_{2i}^* &= \boldsymbol{x}_{2i}'\boldsymbol{\beta}_2 + \varepsilon_{2i} \\ y_{1i} &= \mathbf{1}\left\{y_{1i}^* \geq 0\right\} \\ y_{2i} &= y_{2i}^* \times \mathbf{1}\left\{y_{1i}^* \geq 0\right\} \\ \begin{pmatrix} \varepsilon_{1i} \\ \varepsilon_{2i} \end{pmatrix} &\sim N\left(\mathbf{0}, \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}\right). \end{aligned}$$

The original motivating examples for the model are that the decision to enter the labor force may depend on some covariates (like having children) that do not determine the wage conditional on entering the labor force, and that a government's decision to extend development aid to some country may be driven by some considerations (like human rights) that do not influence the amount of aid conditional on there being any. (The causes need not be separate – the model generalizes Type I Tobit, in which selection and outcome equation coincide.) The first equation is also called *selection equation*, the second one is the *outcome equation*. Typically, the parameter of substantive interest is $\boldsymbol{\beta}_2$. Note that, under assumptions maintained here, $\boldsymbol{\beta}_2$ could be estimated by regressing y_{2i}^* on $\boldsymbol{x}_{2i}$ if the former were observable. However, unless $\rho = 0$, we cannot estimate $\boldsymbol{\beta}_2$ by regressing y_{2i} on $\boldsymbol{x}_{2i}$ in the subsample where $y_{1i} = 1$. This is because by conditioning on $y_{1i} = 1$, we select for high ε_{1i} and therefore for high [low] ε_{2i} if ρ is positive [negative]. That is, the subsample on which y_{2i}^* is observed is selective.

The selection model embodied in the first and third equations gives rise to likelihood

$$\Pr(y_{1i} = 0|\boldsymbol{x}_{1i}, \boldsymbol{x}_{2i}, \boldsymbol{\theta}) = \Phi\left(\frac{-\boldsymbol{x}_{1i}'\boldsymbol{\beta}_1}{\sigma_1}\right).$$

The likelihood of y_{2i} conditionally on $\boldsymbol{x}_{1i}$, $\boldsymbol{x}_{2i}$, and $y_{1i} = 1$ (i.e. y_{2i} being observed) equals

$$\begin{aligned}
f(y_{2i}|\boldsymbol{x}_{1i}, \boldsymbol{x}_{2i}, y_{1i} = 1; \boldsymbol{\theta}) &= \Pr(y_{1i} = 1|y_{2i}, \boldsymbol{x}_{1i}, \boldsymbol{x}_{2i}; \boldsymbol{\theta}) \times f(y_{2i}|\boldsymbol{x}_{1i}, \boldsymbol{x}_{2i}; \boldsymbol{\theta}) / \Phi\left(\frac{\boldsymbol{x}_{1i}'\boldsymbol{\beta}_1}{\sigma_1}\right) \\
&= \Pr(\varepsilon_{1i} > -\boldsymbol{x}_{1i}\boldsymbol{\beta}_1|y_{2i}, \boldsymbol{x}_{1i}, \boldsymbol{x}_{2i}; \boldsymbol{\theta}) \times \frac{1}{\sigma_2}\phi\left(\frac{y_{2i} - \boldsymbol{x}_{2i}'\boldsymbol{\beta}_2}{\sigma_2}\right) / \Phi\left(\frac{\boldsymbol{x}_{1i}'\boldsymbol{\beta}_1}{\sigma_1}\right) \\
&= \frac{1}{\sigma_2}\left(1 - \Phi\left(\frac{-\boldsymbol{x}_{1i}'\boldsymbol{\beta}_1 - \rho\frac{\sigma_1}{\sigma_2}(y_{2i} - \boldsymbol{x}_{2i}'\boldsymbol{\beta}_2)}{\sigma_1\sqrt{1-\rho^2}}\right)\right)\phi\left(\frac{y_{2i} - \boldsymbol{x}_{2i}'\boldsymbol{\beta}_2}{\sigma_2}\right) / \Phi\left(\frac{\boldsymbol{x}_{1i}'\boldsymbol{\beta}_1}{\sigma_1}\right) \\
&= \frac{1}{\sigma_2}\Phi\left(\frac{\boldsymbol{x}_{1i}'\boldsymbol{\beta}_1 + \rho\frac{\sigma_1}{\sigma_2}(y_{2i} - \boldsymbol{x}_{2i}'\boldsymbol{\beta}_2)}{\sigma_1\sqrt{1-\rho^2}}\right)\phi\left(\frac{y_{2i} - \boldsymbol{x}_{2i}'\boldsymbol{\beta}_2}{\sigma_2}\right) / \Phi\left(\frac{\boldsymbol{x}_{1i}'\boldsymbol{\beta}_1}{\sigma_1}\right),
\end{aligned}$$

where we used that the distribution of $\varepsilon_{1i}|\varepsilon_{2i}$ is $N\big(\rho\frac{\sigma_1}{\sigma_2}\varepsilon_{2i}, \sigma_1^2(1-\rho^2)\big)$. Note that all appearances of y_{2i} on the r.h.s. really are y_{2i}^*; I anticipate that the two are the same on the event we condition on.

8.4.2 Identification and ML Estimation

We will establish identification by conducting the following thought experiment: If we had perfect knowledge of the distribution of the data and therefore of the above likelihoods as "black-box functions" of $(\boldsymbol{x}_{1i}, \boldsymbol{x}_{2i}, y_{1i}, y_{2i})$, could we back out the true parameter values? Indeed, it is immediately clear that the probit part of the model identifies $\boldsymbol{\beta}_1/\sigma_1$ (as long as $\mathbb{E}\boldsymbol{x}_i\boldsymbol{x}_i'$ is invertible) but not either of these parameters in isolation. Observe also that $f(y_{2i}|\boldsymbol{x}_{1i}, \boldsymbol{x}_{2i}, y_{2i} = 1; \boldsymbol{\theta})$ can be rewritten as

$$f(y_{2i}|\boldsymbol{x}_{1i}, \boldsymbol{x}_{2i}, y_{2i} = 1; \boldsymbol{\theta}) = \frac{1}{\sigma_2}\Phi\left(\frac{x_{1i}'\frac{\boldsymbol{\beta}_1}{\sigma_1} + \frac{\rho}{\sigma_2}(y_{2i} - \boldsymbol{x}_{2i}'\boldsymbol{\beta}_2)}{\sqrt{1-\rho^2}}\right)\phi\left(\frac{y_{2i} - \boldsymbol{x}_{2i}\boldsymbol{\beta}_2}{\sigma_2}\right) / \Phi\left(\frac{\boldsymbol{x}_{1i}'\boldsymbol{\beta}_1}{\sigma_1}\right).$$

Since this expression also depends on $\boldsymbol{\beta}_1$ and σ_1 only through $\boldsymbol{\beta}_1/\sigma_1$, we can really identify only this ratio. The natural way to deal with this is to normalize $\sigma_1 = 1$ and define $\boldsymbol{\theta} \equiv (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \sigma_2, \rho)$. This leads to the following simplification:

$$\begin{aligned}
\Pr(y_{1i} = 0|\boldsymbol{x}_{1i}, \boldsymbol{x}_{2i}, \boldsymbol{\theta}) &= \Phi\left(-\boldsymbol{x}_{1i}'\boldsymbol{\beta}_1\right) \\
f(y_{2i}|\boldsymbol{x}_{1i}, \boldsymbol{x}_{2i}, y_{2i} = 1; \boldsymbol{\theta}) &= \frac{1}{\sigma_2}\Phi\left(\frac{\boldsymbol{x}_{1i}'\boldsymbol{\beta}_1 + \frac{\rho}{\sigma_2}(y_{2i} - \boldsymbol{x}_{2i}'\boldsymbol{\beta}_2)}{\sqrt{1-\rho^2}}\right)\phi\left(\frac{y_{2i} - \boldsymbol{x}_{2i}'\boldsymbol{\beta}_2}{\sigma_2}\right) / \Phi\left(\boldsymbol{x}_{1i}'\boldsymbol{\beta}_1\right).
\end{aligned}$$

We will now show that the abridged $\boldsymbol{\theta}$ *is* identified. To begin, recall that the selection equation identifies $\boldsymbol{\beta}_1$. We can therefore treat $\boldsymbol{\beta}_1$, hence $\Phi\left(\boldsymbol{x}_{1i}'\boldsymbol{\beta}_1\right)$, as known. Use this to define

$$\begin{aligned}
\tilde{f}(y_{2i}|\boldsymbol{x}_{1i}, \boldsymbol{x}_{2i}; \boldsymbol{\theta}) &= f(y_{2i}|\boldsymbol{x}_{1i}, \boldsymbol{x}_{2i}, y_{2i} = 1; \boldsymbol{\theta}) \Pr(y_{1i} = 1|\boldsymbol{x}_{1i}, \boldsymbol{x}_{2i}, \boldsymbol{\theta}) \\
&= \frac{1}{\sigma_2}\Phi\left(\frac{\boldsymbol{x}_{1i}'\boldsymbol{\beta}_1 + \frac{\rho}{\sigma_2}(y_{2i} - \boldsymbol{x}_{2i}'\boldsymbol{\beta}_2)}{\sqrt{1-\rho^2}}\right)\phi\left(\frac{y_{2i} - \boldsymbol{x}_{2i}'\boldsymbol{\beta}_2}{\sigma_2}\right).
\end{aligned}$$

Observe that

$$\begin{aligned}
\frac{\partial \tilde{f}(\cdot)}{\partial y_{2i}} &= \frac{1}{\sigma_2}\phi\left(\frac{\boldsymbol{x}_{1i}'\boldsymbol{\beta}_1 + \frac{\rho}{\sigma_2}\left(y_{2i} - \boldsymbol{x}_{2i}'\boldsymbol{\beta}_2\right)}{\sqrt{1-\rho^2}}\right)\frac{\rho}{\sigma_2\sqrt{1-\rho^2}}\phi\left(\frac{y_{2i} - \boldsymbol{x}_{2i}'\boldsymbol{\beta}_2}{\sigma_2}\right) \\
&\quad + \frac{1}{\sigma_2^2}\Phi\left(\frac{\boldsymbol{x}_{1i}'\boldsymbol{\beta}_1 + \frac{\rho}{\sigma_2}\left(y_{2i} - \boldsymbol{x}_{2i}'\boldsymbol{\beta}_2\right)}{\sqrt{1-\rho^2}}\right)\phi'\left(\frac{y_{2i} - \boldsymbol{x}_{2i}'\boldsymbol{\beta}_2}{\sigma_2}\right) \\
\frac{\partial \tilde{f}(\cdot)}{\partial \boldsymbol{x}_{2i}} &= \frac{1}{\sigma_2}\phi\left(\frac{\boldsymbol{x}_{1i}'\boldsymbol{\beta}_1 + \frac{\rho}{\sigma_2}\left(y_{2i} - \boldsymbol{x}_{2i}'\boldsymbol{\beta}_2\right)}{\sqrt{1-\rho^2}}\right)\left(-\frac{\rho\boldsymbol{\beta}_2}{\sigma_2\sqrt{1-\rho^2}}\right)\phi\left(\frac{y_{2i} - \boldsymbol{x}_{2i}'\boldsymbol{\beta}_2}{\sigma_2}\right) \\
&\quad - \frac{1}{\sigma_2^2}\Phi\left(\frac{\boldsymbol{x}_{1i}'\boldsymbol{\beta}_1 + \frac{\rho}{\sigma_2}\left(y_{2i} - \boldsymbol{x}_{2i}'\boldsymbol{\beta}_2\right)}{\sqrt{1-\rho^2}}\right)\phi'\left(\frac{y_{2i} - \boldsymbol{x}_{2i}'\boldsymbol{\beta}_2}{\sigma_2}\right)\boldsymbol{\beta}_2 \\
\frac{\partial \tilde{f}(\cdot)}{\partial \boldsymbol{x}_{1i}} &= \frac{1}{\sigma_2}\phi\left(\frac{\boldsymbol{x}_{1i}'\boldsymbol{\beta}_1 + \frac{\rho}{\sigma_2}\left(y_{2i} - \boldsymbol{x}_{2i}'\boldsymbol{\beta}_2\right)}{\sqrt{1-\rho^2}}\right)\frac{\boldsymbol{\beta}_1}{\sqrt{1-\rho^2}}\phi\left(\frac{y_{2i} - \boldsymbol{x}_{2i}'\boldsymbol{\beta}_2}{\sigma_2}\right).
\end{aligned}$$

A first observation is that

$$\boldsymbol{\beta}_2 = -\frac{\partial \tilde{f}(\cdot)/\partial \boldsymbol{x}_{2i}}{\partial \tilde{f}(\cdot)/\partial y_{2i}}.$$

(Without algebra, this is really clear from the fact that y_{2i} and $\boldsymbol{x}_{2i}$ enter the likelihood only through $(y_{2i} - \boldsymbol{x}_{2i}'\boldsymbol{\beta}_2)$.)

Having identified $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$, we can choose to evaluate $\tilde{f}$ at arguments where $\boldsymbol{x}_{1i}'\boldsymbol{\beta}_1 = y_{2i} - \boldsymbol{x}_{2i}'\boldsymbol{\beta}_2 = 0$. At any such value of the argument, we have

$$\begin{aligned}
\left.\frac{\partial \tilde{f}(\cdot)}{\partial \boldsymbol{x}_{1i}}\right\|_{\boldsymbol{x}_{1i}'\boldsymbol{\beta}_1 = y_{2i} - \boldsymbol{x}_{2i}'\boldsymbol{\beta}_{2i} = 0} &= \boldsymbol{\beta}_1 \times \frac{1}{\sigma_2\sqrt{1-\rho^2}}\big(\phi(0)\big)^2 \\
\left.\frac{\partial \tilde{f}(\cdot)}{\partial y_{2i}}\right\|_{\boldsymbol{x}_{1i}'\boldsymbol{\beta}_1 = y_{2i} - \boldsymbol{x}_{2i}'\boldsymbol{\beta}_{2i} = 0} &= \frac{\rho}{\sigma_2^2\sqrt{1-\rho^2}}\big(\phi(0)\big)^2.
\end{aligned}$$

With $\boldsymbol{\beta}_1$ known, these are basically two equations in the two remaining unknowns (σ_2, ρ).

We close with two remarks:

- This is an example of a non-constructive identification proof: We established in a thought experiment that perfect knowledge of the likelihood would allow us to back out parameter values, but our estimation strategy should not be to solve sample analogs of these equations. After all, they involve derivatives of likelihoods evaluated at specific (themselves in practice estimated) parameter values.

- The argument made use of some *support conditions*. For the purpose of our thought experiment, we can freely take derivatives of likelihoods to be known, but we may evaluate these derivatives only at values of $(\boldsymbol{x}_{1i}, \boldsymbol{x}_{2i}, y_{1i}, y_{2i})$ on the support of the true distribution. Because of the normality assumption on ε_{2i}, we actually know that $y_{2i} - \boldsymbol{x}_{2i}'\boldsymbol{\beta}_2 = 0$ occurs on the support. We do not really know that $\boldsymbol{x}_{1i}'\boldsymbol{\beta}_1 = 0$ can occur but our use of that assumption was not tight, i.e. we could work around it as long as there is some variation in $\boldsymbol{x}_{1i}$. Support assumptions on covariates are often made to enable arguments like the above.

Now, how should we estimate this model? An obvious approach is Maximum Likelihood. Substituting in for the standard normal p.d.f. and dropping constants, the objective function is

$$Q_n(\boldsymbol{\theta}) =$$
$$\frac{1}{n}\sum_{i=1}^{n}\left[(1-y_{1i})\log\Phi\left(-\boldsymbol{x}_{1i}'\boldsymbol{\beta}_1\right)+y_{1i}\left(\log\Phi\left(\frac{\boldsymbol{x}_{1i}'\boldsymbol{\beta}_1+\frac{\rho}{\sigma_2}\left(y_{2i}-\boldsymbol{x}_{2i}'\boldsymbol{\beta}_2\right)}{\sqrt{1-\rho^2}}\right)-\frac{1}{2}(y_{2i}-\boldsymbol{x}_{2i}'\boldsymbol{\beta}_2)^2-\log\sigma_2\right)\right],$$

where $\boldsymbol{\theta}\equiv(\boldsymbol{\beta}_1,\boldsymbol{\beta}_2,\sigma_2,\rho)$.

8.4.3 "Heckit"

ML is the statistically efficient way to estimate such models; many implementations exist, and there are also some shortcuts like pre-estimating $\boldsymbol{\beta}_1$ by probit to get a good initial point. However, with larger parameter vectors, the problem remains involved even by modern standards because the likelihood is multimodal. Heckman proposed a two-step method for this type of model. This method builds on the observation that

$$\begin{aligned}\mathbb{E}(y_{2i}|\boldsymbol{x}_{1i},\boldsymbol{x}_{2i},y_{1i}=1) &= \mathbb{E}(\boldsymbol{x}_{2i}'\boldsymbol{\beta}_2+\varepsilon_{2i}|\boldsymbol{x}_{1i},\boldsymbol{x}_{2i},y_{1i}=1)\\ &= \boldsymbol{x}_{2i}'\boldsymbol{\beta}_2+\mathbb{E}(\varepsilon_{2i}|\varepsilon_{1i}\geq-\boldsymbol{x}_{1i}'\boldsymbol{\beta}_1).\end{aligned}$$

To get a grip on the last expression, let z_i be distributed standard normal and recall the following, standard algebra:

$$\begin{aligned}\mathbb{E}(z_i|z_i\geq t) &= \frac{\int_{z=t}^{\infty}z\phi(z)dz}{\int_{z=t}^{\infty}\phi(z)dz}=\frac{(2\pi)^{-1/2}\int_{z=t}^{\infty}ze^{-z^2/2}dz}{\Phi(-t)}\\ &= \frac{(2\pi)^{-1/2}\left[-e^{-z^2/2}\right]_t^{\infty}}{\Phi(-t)}=\frac{(2\pi)^{-1/2}e^{-t^2/2}}{\Phi(-t)}=\frac{\phi(t)}{\Phi(-t)}\equiv\lambda(-t),\end{aligned}$$

where the last equality defines the *Inverse Mills Ratio*.

Now,

$$\begin{aligned}\mathbb{E}(\varepsilon_{2i}|\varepsilon_{1i}\geq-\boldsymbol{x}_{1i}'\boldsymbol{\beta}_1) &= \frac{\int_{-\boldsymbol{x}_{1i}'\boldsymbol{\beta}_1}^{\infty}\mathbb{E}(\varepsilon_{2i}|\varepsilon_1)\phi\left(\varepsilon_1\right)d\varepsilon_1}{\int_{-\boldsymbol{x}_{1i}'\boldsymbol{\beta}_1}^{\infty}\phi\left(\varepsilon_1\right)d\varepsilon_1}\\ &= \frac{\int_{-\boldsymbol{x}_{1i}'\boldsymbol{\beta}_1}^{\infty}\rho\sigma_2\varepsilon_1\phi\left(\varepsilon_1\right)d\varepsilon_1}{\int_{-\boldsymbol{x}_{1i}'\boldsymbol{\beta}_1}^{\infty}\phi\left(\varepsilon_1\right)d\varepsilon_1}=\rho\sigma_2\frac{\int_{-\boldsymbol{x}_{1i}'\boldsymbol{\beta}_1}^{\infty}\varepsilon_1\phi\left(\varepsilon_1\right)d\varepsilon_1}{\int_{-\boldsymbol{x}_{1i}'\boldsymbol{\beta}_1}^{\infty}\phi\left(\varepsilon_1\right)d\varepsilon_1}=\rho\sigma_2\lambda(\boldsymbol{x}_{1i}'\boldsymbol{\beta}_1),\end{aligned}$$

where we used that due to our normalization, ε_{1i} is standard normal.

Thus, for those data points where $y_{1i}=1$, we can write

$$y_{2i}=\boldsymbol{x}_{2i}'\boldsymbol{\beta}_2+\rho\sigma_2\lambda(\boldsymbol{x}_{1i}'\boldsymbol{\beta}_1)+\eta_i$$

where $\mathbb{E}(\eta_i|\boldsymbol{x}_{1i},\boldsymbol{x}_{2i},y_{1i}=1)=0$. If we knew $\boldsymbol{\beta}_1$, we could therefore just estimate $\boldsymbol{\beta}_2$ by running a OLS regression of y_{2i} on $(\boldsymbol{x}_{2i},\lambda\left(\boldsymbol{x}_{1i}'\boldsymbol{\beta}_1\right))$. (The last component of the estimator would estimate $\rho\sigma_2$.

The population variance of $(y_{2i}|\boldsymbol{x}_{1i}, \boldsymbol{x}_2 i, y_{1i} = 1)$ equals $\sigma_2^2(1-\rho^2)$, so that separate estimates of ρ and σ_2 can be backed out.) In reality, things are a bit more complicated because $\boldsymbol{\beta}_2$ is unknown. Heckman established that the following two-step procedure ("Heckit") works:

Step 1. Use probit to estimate $\boldsymbol{\beta}_1$. Call this estimator $\hat{\boldsymbol{\beta}}_1$.

Step 2. Restrict attention to observations with $y_{1i} = 1$. Use OLS to estimate the equation

$$y_{2i} = \boldsymbol{x}_{2i}'\boldsymbol{\beta}_2 + \rho\sigma_2\lambda(\boldsymbol{x}_{1i}'\hat{\boldsymbol{\beta}}_1) + \eta_i.$$

To reiterate, that this works is not completely obvious because of the estimated regressor $\lambda(\boldsymbol{x}_{1i}'\hat{\boldsymbol{\beta}}_1)$ on the r.h.s., but it is nonetheless true (and arguments of this sort have since been much generalized). Unsurprisingly, inference theory changes and OLS standard errors would not be valid.

The Heckman two-step method generalizes easily to variations on the above model. It can be seen as the first appearance of a control function approach: correcting for selectivity by introducing a function into the regression that compensates selection bias. The "Heckit" estimator is hardcoded in most canned packages. Notice, though, that it is inefficient: It exploits normality assumptions for identification (and is sensitive to their failure!) but then fails to fully use them for estimation, and in this case there is no reason to believe that the moment conditions coincide with the score equations. In particular, an ML estimator would use second-stage information also in the estimation of $\boldsymbol{\beta}_1$. The choice between ML and Heckit estimation of this model depends on how complex the likelihood is in a given application. Both estimation methods are implemented in Stata and similar packages.

8.5 Maximum of a Uniform Distribution

This is a less regular example and one in which ML and GMM may disagree. Let x_i be i.i.d. uniformly distributed on $[0, \alpha_0]$. The aim is to estimate α_0. We briefly note that $\mathbb{E}x_i = \alpha_0/2$ and so $\hat{\alpha} = 2\bar{x}$ is a GMM estimator based on moment condition $\mathbb{E}(2x_i - \alpha_0) = 0$. We understand the behavior of this estimator very well. But in this example, it is not the ML estimator and turns out to be rather inefficient indeed. (That said, it is BLUE. So the ML estimator must be either biased or nonlinear. We will see that it is both.)

Let's compute the ML estimator. The likelihood for a single observation is $f(x; \alpha) = 1/\alpha \times \mathbf{1}\{0 \leq x \leq \alpha\}$, and so the sample criterion function equals

$$Q_n(\alpha) = \frac{1}{n}\sum_{i=1}^{n} \log\left(1/\alpha \times \mathbf{1}\{x_i \in [0, \alpha]\}\right) = \begin{cases} -\log \alpha & \text{if } \max_i \{x_i\} \leq \alpha \\ -\infty & \text{otherwise} \end{cases}$$

with population analog

$$Q(\alpha) = \mathbb{E} \log f(x_i; \alpha) = \begin{cases} -\log \alpha & \text{if } \alpha \leq \alpha_0 \\ -\infty & \text{otherwise} \end{cases}$$

By inspection, these problems are solved by $\hat{\alpha} = \max_i \{x_i\}$ respectively by the true value α_0.

These objective functions are discontinuous, and therefore of course not differentiable, at their maxima. This means that none of our extremum estimator theorems are immediately applicable, and it is a massive red flag with regard to our asymptotic distribution theorem. Indeed, while consistency of the estimator obtains (it's intuitively obvious from the closed-form expression, follows from developments below, but could also be shown by recovering well-separatedness of the population maximum), neither $\sqrt{n}$-consistency nor asymptotic normality are true. Plotting the objective function can provide an intuition: The Hessian at the solution is not well-defined, but the vertical drop loosely suggests an "unbounded Hessian," which by the theorem's formula would suggest an asymptotic variance of 0 for $\sqrt{n}(\hat{\alpha} - \alpha_0)$. In the same way that an asymptotic variance of ∞ would suggest failure of $\sqrt{n}$-consistency, this observation suggests, but does of course not prove, so-called "superconsistency." Is the estimator faster than $\sqrt{n}$-consistent?

The answer is yes; in fact, the true rate of convergence is n. To see this, let's approximate the c.d.f. of $n(\hat{\alpha} - \alpha_0)$. This c.d.f. is obviously 1 for nonnegative arguments, another pointer that asymptotic normality will fail (and also implying that the estimator cannot be unbiased). For arguments $t \leq 0$,

we have

$$\begin{aligned}
& \Pr(n(\hat{\alpha}-\alpha_0)\leq t) \\
= \;& \Pr\left(\max_{i=1,...,n}\{x_i\}\leq \alpha_0+t/n\right) \\
= \;& \Pr\left(x_1\leq \alpha_0+t/n,...,x_n\leq \alpha_0+t/n\right) \\
= \;& \Pr\left(x_1/\alpha_0\leq 1+t/\left(n\alpha_0\right),...,x_n/\alpha_0\leq 1+t/\left(n\alpha_0\right)\right) \\
= \;& \left(1+t/\left(n\alpha_0\right)\right)^n \\
\to \;& e^{t/\alpha_0},
\end{aligned}$$

where the last step invokes a well-known nonstochastic limit formula. Note this is 1 at $t=0$ as expected, so we derived a coherent limit c.d.f. In fact, we showed that $-n(\hat{\alpha}-\alpha_0)$ converges to an exponential distribution whose dispersion increases with α_0. A corollary is that $n(\hat{\alpha}-\alpha_0)=O_p(1)$, i.e. the estimator is consistent at rate n.

We conclude the example with some additional observations.

- Our standard inference theory does not apply, but it is relatively easy to construct hypothesis tests for this setting. Because the model is fully parameterized, we can even do exact tests. In particular, suppose that $H_0:\alpha_0=\alpha$ holds for some fixed value α. We clearly reject this null if $\hat{\alpha}>\alpha$, so the testing problem is effectively one-sided. From the above algebra, we have (setting $\alpha=\alpha_0$ and flipping the sign on t) $\Pr(n(\alpha-\hat{\alpha})>t)=\left(1-t/\left(n\alpha\right)\right)^n$, so that the 95% critical value is $n\alpha(1-.05^{1/n})$.

  We can construct a CI by inverting this test. Notice that the critical value depends on α (the test statistic is not asymptotically *pivotal*) and so in principle we need to recompute it at each α. In this example, the condition that the test statistic be smaller than the critical value can be solved in closed form, and we get a CI of $[\hat{\alpha},20^{1/n}\hat{\alpha}]$.

  While there is no need for asymptotic CI's if we have computationally cheap finite sample ones, in principle we can compute an asymptotic counterpart as well. From the last line of algebra above, it would be $[\hat{\alpha},\hat{\alpha}\frac{n}{n+\log.05}]$. Note that this is well-defined only if $n+\log.05>0$, which is the case for $n\geq 3$. This weird feature is an artefact of using an asymptotic approximation (namely the limit formula) at very small n. The intervals become very similar for moderate n.

- The ML estimator has negative bias, but standard formulae for order statistics imply that $\mathbb{E}\hat{\alpha}=\frac{n}{n+1}\alpha_0$, so we can define a bias-corrected (in fact unbiased) and still superconsistent estimator $\tilde{\alpha}=\frac{n+1}{n}\hat{\alpha}$. For this particular estimation problem, that is the preferred estimator. (Theorems about asymptotic efficiency of ML do not apply due to the problem's irregularity.)

8.6 (Smoothed) Maximum Score

Consider the binary choice model

$$y_i = \mathbf{1}\{\boldsymbol{x}_i'\boldsymbol{\beta}_0 + \varepsilon_i \geq 0\},$$

where the researcher observes $(y_i, \boldsymbol{x}_i)$. We do not assume an exact distribution for ε_i, but we do assume that ε_i is continuous and that $\text{med}(\varepsilon_i) = 0$. For simplicity, we also assume continuous $\boldsymbol{x}_i$ with full support other than having a constant.

While the median assumption amounts to a location normalization, the assumptions are still weaker than for probit, say, and so $\boldsymbol{\beta}_0$ can only be identified up to scale. Unlike in earlier examples, it is convenient to normalize $|\boldsymbol{\beta}_0| = 1$ and impose the same restriction on estimators. Then

$$\boldsymbol{\beta}_0 \in \arg\max_{\boldsymbol{\beta}:|\boldsymbol{\beta}|=1} \mathbb{E}\left((2y_i - 1)\,\mathbf{1}\{\boldsymbol{x}_i'\boldsymbol{\beta} \geq 0\}\right)$$

because

$$\mathbb{E}\left((2y_i - 1)\,\mathbf{1}\{\boldsymbol{x}_i'\boldsymbol{\beta} \geq 0\}\right) = \mathbb{E}\big(\mathbb{E}(2y_i - 1|\boldsymbol{x}_i)\mathbf{1}\{\boldsymbol{x}_i'\boldsymbol{\beta} \geq 0\}\big)$$

and the r.h.s outer integrand is maximized pointwise by $\boldsymbol{\beta}_0$ because

$$\mathbb{E}\left(2y_i - 1|\boldsymbol{x}_i\right) \geq 0 \Leftrightarrow \boldsymbol{x}_i'\boldsymbol{\beta}_0 \geq 0.$$

If $\boldsymbol{x}_i$ has full support, then for any $\boldsymbol{\beta} \neq \boldsymbol{\beta}_0$, the set $\{\boldsymbol{x}_i : \boldsymbol{x}_i'\boldsymbol{\beta} \times \boldsymbol{x}_i'\boldsymbol{\beta}_0 < 0\}$ has positive probability and so the above arg max is unique. Without this assumption, $\boldsymbol{\beta}_0$ may be *partially identified*: Knowledge of the population distribution of observables restricts $\boldsymbol{\beta}$ to a nontrivial but also nonsingleton subset of the unit sphere.

Note that the empirical distribution of $\boldsymbol{x}_i$ has at most n mass points and so cannot have full support. Thus, the estimator

$$\hat{\boldsymbol{\beta}} = \arg\max_{\boldsymbol{\beta}:|\boldsymbol{\beta}|=1} \frac{1}{n}\sum_{i=1}^{n} (2y_i - 1)\,\mathbf{1}\{\boldsymbol{x}_i'\boldsymbol{\beta} \geq 0\}$$

is not well-defined as written because the arg max is not unique. In the following, let $\hat{\boldsymbol{\beta}}$ be a measurable selection from the arg max, e.g. the element that minimizes the first, then the second, etc. component.

This is the *Maximum Score* estimator (Manski, 1975). It has historic importance as one of the first *nonparametric* (actually semiparametric in modern terminology because of the linear index structure) estimators. It is also supremely ill-behaved. Under reasonable conditions, consistency can be established along the lines of the theorems provided in this lecture. As an aside, this shows the theorems' power because the resulting proof is much shorter than the original one. But with regard to the asymptotic distribution, red flags abound. To begin, the Hessian at the sample arg max is $\mathbf{0}$. This suggests, but does of course not prove, a slower than $\sqrt{n}$ rate of convergence. This conjecture is true:

The true rate is $n^{1/3}$ (Kim and Pollard, 1990); an estimator with $\sqrt{n}$-convergence does not exist under the assumptions (Chamberlain, 1986, who formalizes the intuition just alluded to), and the asymptotic distribution is intractable.[12]

Many of these issues are due to the extreme non-smoothness of the objective function. This raises the possibility that artifically smoothing the objective function might make for better behaved estimators. There is even a vague intuition that it might help with efficiency because it brings to bear whether a given $\boldsymbol{x}_i'\boldsymbol{\beta}$ is very close to zero or not. All these intuitions are correct. In particular, write

$$\hat{\boldsymbol{\beta}} = \arg \max_{\boldsymbol{\beta}:|\boldsymbol{\beta}|=1} \frac{1}{n} \sum_{i=1}^{n} (2y_i - 1)\, g_n(\boldsymbol{x}_i'\boldsymbol{\beta}),$$

where the function $g_n(t)$ is smooth, has $g(0) = 1/2$, asymptotes 0 [1] as $t \to -\infty$ [∞], and converges to $\mathbf{1}\{t \geq 0\}$ at a certain rate as $n \to \infty$. If you have seen kernel density estimation or kernel regression before, you'll realize that this is very similar.

This estimator is called the *Smoothed Maximum Score* estimator. It was suggested by Horowitz (1992), who also showed that it is asymptotically normal. Its rate of convergence is not $n^{1/2}$, but can be arbitrarily close to $n^{1/2}$ if the distribution of $\boldsymbol{x}_i$ is sufficiently favorable and g_n is chosen smartly.[13]

[12] A naive reaction might be to "just bootstrap" the distribution, but the simple nonparametric bootstrap is demonstrably inconsistent here (Abrevaya and Huang, 2005).

[13] Moreover, the estimator cannot only be bootstrapped; Horowitz (2002) shows that for test statistics based on a studentized estimator, the bootstrap achieves asymptotic refinement.

8.7 A (too hard) Exam Question with Solution

This is an exam question from the very first time I taught this lecture. Clearly I was miscalibrated on what's reasonable. The question is based on Honoré and Hu, Estimation of cross sectional and panel data censored regression models with endogeneity, *Journal of Econometrics* 2004. Note also that when this question was written, the lecture used the concepts of stationarity and ergodicity.

8.7.1 Questions

Consider the following model of censored panel data:

$$y_{it} = \max\left\{0, y_{i,t-1}\gamma_0 + \alpha_i + \varepsilon_{it}\right\},$$

where ε_{it} is distributed i.i.d. with expectation $\mathbb{E}\varepsilon_{it} = 0$. Assume also that the process has infinite past and that the distribution of ε_{it} is such that $\Pr(y_{it} = 0) > 0$ (e.g. ε_{it} has full support).

1 Let the index set $\{t_k\}$ collect all t s.t. $y_{i,t-2} = 0$. Consider the sequence $\{\boldsymbol{w}_{ik}\}$ defined by $\boldsymbol{w}_{ik} = (y_{it_k}, y_{i,t_k-1})$. Prove that this sequence is stationary and ergodic. (Hint: This can be done without much algebra.)

2 Define

$$e_{it}(\gamma) \equiv \max\left\{0, y_{it} - y_{i,t-1}\gamma\right\}.$$

What is $\mathbb{E}(e_{it}(\gamma)|y_{i,t-2} = 0)$? This expression simplifies at $\gamma = \gamma_0$.

3 Describe a strategy for estimating γ_0 that uses the above insights. Define an estimator and argue that it is consistent. You may invoke regularity conditions, but should state them. Do not prove theorems.

4 Does normality of the estimators follow straightforwardly from results given in the lecture? Explain.

8.7.2 Answers

1 If $y_{i,t-2} = 0$, then

$$\begin{aligned} y_{i,t-1} &= \max\left\{0, \alpha_i + \varepsilon_{i,t-1}\right\} \\ y_{it} &= \max\left\{0, \gamma_0 \max\left\{0, \alpha_i + \varepsilon_{i,t-1}\right\} + \alpha_i + \varepsilon_{it}\right\}, \end{aligned}$$

so $\boldsymbol{w}_{ik}$ depends on $(\varepsilon_{it}, \varepsilon_{i,t-1})$. Together with the distributional assumption on ε_i, this immediately implies stationarity. Ergodicity is even simpler because conditional on $y_{i,t-2} = 0$, it is clear that y_{it}

is independent of $y_{i,t-s}$ for any $s \geq 3$. Therefore, in the subsequence, we have $\boldsymbol{w}_{ik} \perp \boldsymbol{w}_{il}$ whenever $|k-l| \geq 2$.

2 Assume $y_{i,t-2} = 0$. Using expressions from 1.1, we find

$$\begin{aligned}
\mathbb{E}(e_{it}(\gamma)|y_{i,t-2}=0) &= \mathbb{E}(\max\{0, y_{it} - y_{i,t-1}\gamma\}\,|y_{i,t-2}=0) \\
&= \mathbb{E}\left(\max\left\{0, \underline{\max\{0, \gamma_0 \max\{0, \alpha_i + \varepsilon_{i,t-1}\} + \alpha_i + \varepsilon_{it}\}} - \gamma \max\{0, \alpha_i + \varepsilon_{i,t-1}\}\right\}\right) \\
&= \mathbb{E}\left(\max\{0, \gamma_0 \max\{0, \alpha_i + \varepsilon_{i,t-1}\} + \alpha_i + \varepsilon_{it} - \gamma \max\{0, \alpha_i + \varepsilon_{i,t-1}\}\}\right) \\
&= \mathbb{E}\left(\max\{0, (\gamma_0 - \gamma) \max\{0, \alpha_i + \varepsilon_{i,t-1}\} + \alpha_i + \varepsilon_{it}\}\right),
\end{aligned}$$

using that whenever 0 binds in the underlined maximization, 0 will also bind in the outer maximization. It follows that

$$\mathbb{E}(e_{it}(\gamma_0)|y_{i,t-2}=0) = \mathbb{E}\left(\max\{0, \alpha_i + \varepsilon_{it}\}\right) = \mathbb{E}\left(y_{i,t-1}|y_{i,t-2}=0\right),$$

where the r.h. equality follows from recalling that $y_{i,t-1} = \max\{0, \alpha_i + \varepsilon_{i,t-1}\}$ and ε_{it} and $\varepsilon_{i,t-1}$ are i.i.d. Furthermore, this equality is easily seen to hold only if $\gamma = \gamma_0$.

3 It is now clear that we can estimate γ_0 by GMM. We will use $\mathbb{E}(e_{it}(\gamma_0)|y_{i,t-2}=0) = \mathbb{E}\left(y_{i,t-1}|y_{i,t-2}=0\right)$ as moment condition, meaning that we set $g(\boldsymbol{w}_k;\gamma) = \max\{0, y_{it_k} - y_{i,t_k-1}\gamma\} - y_{i,t_k-1}$. This will be optimized over $\gamma \in \mathbb{R}^+$. We proved stationarity and ergodicity of $\boldsymbol{w}_k$ in 1.1. Continuity of g is clear. $\mathbb{E}\left(\max_{\boldsymbol{\gamma}} |g(\boldsymbol{w}_k;\gamma)|\right) < \infty$ holds because

$$\begin{aligned}
&\max_{\boldsymbol{\gamma}\in\mathbb{R}^+} |g(\boldsymbol{w}_k;\gamma)| = \max_{\boldsymbol{\gamma}\in\mathbb{R}^+} |\max\{0, y_{it_k} - y_{i,t_k-1}\gamma\} - y_{i,t_k-1}| \\
&= \max_{\boldsymbol{\gamma}\in\mathbb{R}^+} |\max\{-y_{i,t_k-1}, y_{it_k} - (1+\gamma)y_{i,t_k-1}\}| \leq |\max\{-y_{i,t_k-1}, y_{it_k} - y_{i,t_k-1}\}| = |y_{it_k} - y_{i,t_k-1}|
\end{aligned}$$

if $y_{it_k} - y_{i,t_k-1} > 0$; else, we continue after the inequality with

$$\ldots \leq y_{i,t_k-1}.$$

Thus we can claim consistency.

4 No! The max-operator prevents continuous differentiability of g. Normality does in fact obtain, but is quite hard to show.

8.8 O_P and o_P

This section briefly recalls O_P and o_P notation. These generalize nonstochastic concepts: For nonstochastic sequences x_n and a_n, recall that $x_n = o(a_n)$ if $x_n/a_n \to 0$ and $x_n = O(a_n)$ if no subsequence of x_n/a_n diverges; equivalently, there exist finite scalars (M, N) s.t. $x_n \leq Ma_n$ for all $n \geq N$.[14] Intuitively, in the first case x_n is eventually small compared to a_n; in the second case, they may be of the same order though x_n nay also be smaller.

Important special cases arise for $a_n = (1, 1, \dots)$: We have $x_n = o(1)$ if $x_n \to 0$ and $x_n = O(1)$ if $\limsup_{n\to\infty} x_n < \infty$. Note in these simple examples that certain rules for combining "big O, small o" terms are easy to derive. For example, if $x_n \to 0$ and $y_n \to 0$, then $x_n y_n \to 0$, hence "$o(1) \times o(1) = o(1)$." Similarly, "$O(1) + O(1) = O(1)$," "$O(1) + o(1) = O(1)$," and (importantly) "$O(1)o(1) = o(1)$." Note also that the case of $a_n = 1$ really encompasses all cases because one could define $\tilde{x}_n = x_n/a_n$.

The generalization to stochastic sequences X_n goes as follows.[15]

- $X_n = o_P(a_n)$ ("is of smaller stochastic order than") if $X_n/a_n \xrightarrow{p} 0$.
- $X_n = O_P(a_n)$ ("is of the same stochastic order as") if for each $\epsilon > 0$, there exist M_ϵ, N_ϵ s.t. $\Pr(|X_n/a_n| \geq M_\epsilon) \leq \epsilon$ for all $n \geq N_\epsilon$.

Again, the special case of $a_n = 1$ is of interest. We say that X_n *vanishes* if $X_n = o_P(1)$ and that X_n is *stochastically bounded* if $X_n = O_P(1)$. Note that, if X_n is stochastically bounded, this does not imply a deterministic upper bound on X_n, even for large n; however X_n cannot take arbitrarily high values with nonvanishing probability. Indeed, a "typical" instance of $X_n = O_P(1)$ is that $X_n \xrightarrow{d} Z$, where Z is some known random variable. If X_n is a studentized test statistic, we might have $Z \sim N(0, 1)$, so very large values of Z are possible though of course unlikely. For another important use of the term, note that an estimation error $(\hat{\theta} - \theta_0)$ is said to be of (stochastic) order $n^{-1/2}$ if

$$\hat{\theta} - \theta_0 = O_P(n^{-1/2}) \Leftrightarrow \sqrt{n}(\hat{\theta} - \theta_0) = O_P(1).$$

The estimator is then called "$\sqrt{n}$-consistent." Also, the qualifier *stochastic* is sometimes dropped if we are obviously talking about random variables as in this last example.

Finally, it is an easy homework to prove that all the above rules for "O-o-algebra" extend, e.g. "$o_P(1) + o_P(1) = o_P(1)$." This is of paramount importance to simplify otherwise extremely tedious arguments that involve multiple limit taking. In proofs, this algebra is routinely, and without much

[14] In maths texts, it is more common to define the concepts in terms of functions. That is more general but for our purposes, it is the same: Just think of x_n and a_n as values of functions defined on $\mathbb{N}$.

[15] The subscript P alludes to the probability space on which X_n lives, but the notation is routinely used without defining such. However, if such a space was defined using a different letter, that letter should be used.

elaboration, used to absorb several $o_P(1)$ terms into one; importantly, it is understood that different appearances of "$o_P(1)$" in the same proof need not refer to literally the same term, just to terms that all vanish.

9 Quantile Regression

9.1 Introduction

Just as mean regression models conditional means, quantile regression models conditional quantiles. This can be useful to assess effects that differ on different parts of a distribution, as is common in economic applications. For example, treatment effects from an educational intervention might be different dependent on unobserved ability. In addition, quantiles and quantile regression have certain robustness properties, though we will not elaborate these.[16] We will restrict our technical analysis to the median to avoid one additional parameter. The generalization is conceptually easy and relies on generalizing the below, symmetric loss function to the asymmetric "tick" loss function (hence the additional parameter).

Let the r.v.'s $\{y_i, \boldsymbol{x}_i\}$ denote outcomes and covariates. You may recall that a median solves $\min_c \mathbb{E}(|y_i - c|)$. We will actually define it as solving $\min_c \mathbb{E}(|y_i - c| - |y_i|)$; this definition does not require that $\mathbb{E}(|y_i - c|)$ and $\mathbb{E}|y_i|$ exist and is equivalent when they do. We will also assume that the median is unique. Recall this is the case if the relevant c.d.f. F is strictly increasing at $1/2$, in which case the minimization problem is uniquely solved by $= F^{-1}(.5)$. We also note that a unique median is characterized by an analog to a score equation:

$$\mathbb{E}(\text{sg}(y_i - c)) = 0,$$

where $\text{sg}(x) \in \{-1, 0, 1\}$ is the sign function. Similarly, if we postulate

$$y_i \quad = \quad \mathbf{x}_i'\boldsymbol{\beta}_0 + \varepsilon_i,$$

where $\boldsymbol{x}_i$ is a r.v. supported on $\boldsymbol{R}^K$ and ε_i has unique median 0, then we have the 'moment' condition $\mathbb{E}(\text{sg}(\varepsilon_i)|\boldsymbol{x}_i) = 0$ and the true parameter value

$$\boldsymbol{\beta}_0 = \arg\min_{\boldsymbol{\beta}} \mathbb{E}(|y_i - \boldsymbol{x}_i'\boldsymbol{\beta}| - |\varepsilon_i|)$$

with natural estimator

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \frac{1}{n}\sum_{i=1}^{n} |y_i - \boldsymbol{x}_i'\boldsymbol{\beta}|.$$

This looks somewhat like m-estimation, and the game will be to adapt m-estimator asymptotics. However, there are some important differences, most notably that the sample criterion function is not smooth.

We will impose the following assumptions:

[16] To begin, moments don't need to exist for the below development to go through. More subtly, quantile regression is robust in a technical sense (associated with Huber) that is roughly translated as insensitivity to outliers.

Assumption 9.1 *$\boldsymbol{w}_i = (y_i, \boldsymbol{x}_i)$ is i.i.d.*

Assumption 9.2 $\mathbb{E}(\|\boldsymbol{x}_i\|^2) < \infty$.

Assumption 9.3 *The unobservables ε_i have a conditional density $f(\epsilon|\boldsymbol{x}_i)$ and conditional median $q_{.5}(\varepsilon_i|\boldsymbol{x}_i) = 0$.*

Assumption 9.4 *The matrix $\boldsymbol{C} \equiv \mathbb{E}(f(0|\boldsymbol{x}_i)\boldsymbol{x}_i\boldsymbol{x}_i')$ is positive definite.*

For Assumption **??**, what is really needed is the implication that the distribution of $(y_i|\boldsymbol{x}_i)$ has a unique median at $\boldsymbol{x}_i'\boldsymbol{\beta}_0$. Assumption **??** is a local identification condition ensuring curvature of the population criterion function at the true parameter value.

9.2 Characterizing the Estimator

The estimator does not admit a closed-form expression. However, analysis of the optimization problem informs both an algorithm for finding it and asymptotic analysis later on.

The objective function $Q_n(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^n |y_i - \mathbf{x}_i'\boldsymbol{\beta}|$ is not differentiable but has well-defined directional derivatives. For any unit vector $\boldsymbol{w}$,

$$\begin{aligned}\frac{\partial}{\partial\gamma}Q_n(\boldsymbol{\beta}-\gamma\boldsymbol{w})\Big\|_{\gamma=0^+} &= \frac{1}{n}\sum_{i=1}^n \frac{\partial}{\partial\gamma}|y_i - \boldsymbol{x}_i'\boldsymbol{\beta} + \gamma\boldsymbol{x}_i'\boldsymbol{w}|\Big\|_{\gamma=0^+} \\ &= \frac{1}{n}\sum_{i=1}^n \big(\boldsymbol{x}_i'\boldsymbol{w}\,\mathrm{sg}(y_i - \boldsymbol{x}_i'\boldsymbol{\beta}) + |\boldsymbol{x}_i'\boldsymbol{w}|\mathbf{1}\{y_i = \boldsymbol{x}_i'\boldsymbol{\beta}\}\big) \qquad (10)\end{aligned}$$

and similarly

$$\frac{\partial}{\partial\gamma}Q_n(\boldsymbol{\beta}-\gamma\boldsymbol{w})\Big\|_{\gamma=0^-} = \frac{1}{n}\sum_{i=1}^n \big(\boldsymbol{x}_i'\boldsymbol{w}\,\mathrm{sg}(y_i - \boldsymbol{x}_i'\boldsymbol{\beta}) - |\boldsymbol{x}_i'\boldsymbol{w}|\mathbf{1}\{y_i = \boldsymbol{x}_i'\boldsymbol{\beta}\}\big). \qquad (11)$$

The derivatives have a simple intuition: For each $(y_i, \boldsymbol{x}_i)$, if we are not at the nondifferentiability of $|y_i - \boldsymbol{x}_i'\boldsymbol{\beta}|$, the derivative is the left-hand term; but at the nondifferentiability, the right-derivative is positive and the left-derivative is negative irrespective of the sign of $\boldsymbol{x}_i'\boldsymbol{w}$. Also, observe that γ itself does not enter the derivative, illustrating that the objective function is piecewise linear.

This last observation informs solution algorithms. In particular, the minimum is necessarily attained at some $\hat{\boldsymbol{\beta}}$ s.t. $y_i = \boldsymbol{x}_i'\hat{\boldsymbol{\beta}}$ for some i and typically $y_i = \boldsymbol{x}_i'\hat{\boldsymbol{\beta}}$ for K distinct values of i. Visually, $\hat{\boldsymbol{\beta}}$ must exactly fit some and typically will exactly fit K observations. This insight immediately gives us a finite, if large, set of candidate solutions. As in Linear Programming, where these values are called *basic solutions*, solution algorithms attempt to efficiently explore them. For example, the simplex method swaps constraints one at a time according to a certain rule.

Next, observe that

$$\begin{aligned}
&\frac{\partial}{\partial\gamma}Q_n(\hat{\boldsymbol{\beta}}-\gamma\boldsymbol{w})\bigg\|_{\gamma=0-} \le 0 \le \frac{\partial}{\partial\gamma}Q_n(\hat{\boldsymbol{\beta}}-\gamma\boldsymbol{w})\bigg\|_{\gamma=0+} \\
\implies\quad & -\frac{1}{n}\sum_{i=1}^n |\boldsymbol{x}_i'\boldsymbol{w}|\mathbf{1}\{y_i=\boldsymbol{x}_i'\hat{\boldsymbol{\beta}}\} \le \frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i'\boldsymbol{w}\,\mathrm{sg}(y_i-\boldsymbol{x}_i'\hat{\boldsymbol{\beta}}) \le \frac{1}{n}\sum_{i=1}^n |\boldsymbol{x}_i'\boldsymbol{w}|\mathbf{1}\{y_i=\boldsymbol{x}_i'\hat{\boldsymbol{\beta}}\} \\
\implies\quad & \left|\frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i'\boldsymbol{w}\,\mathrm{sg}(y_i-\boldsymbol{x}_i'\hat{\boldsymbol{\beta}})\right| \le \frac{1}{n}\sum_{i=1}^n |\boldsymbol{x}_i'\boldsymbol{w}|\mathbf{1}\{y_i=\boldsymbol{x}_i'\hat{\boldsymbol{\beta}}\} \\
&\qquad\qquad \le \frac{1}{n}\sum_{i=1}^n \|\boldsymbol{x}_i\|\mathbf{1}\{y_i=\boldsymbol{x}_i'\hat{\boldsymbol{\beta}}\} = O_P\left(\frac{\max_i\|\boldsymbol{x}_i\|}{n}\right) = o_P(n^{-1/2}),
\end{aligned}$$

where the inequality is the Cauchy-Schwarz inequality (recall that $\boldsymbol{w}$ is a unit vector), the next step uses that ε_i is distributed continuously, hence with probability 1 we have $\frac{2}{n}\sum_{i=1}^n \|\boldsymbol{x}_i\|\times\mathbf{1}\{y_i=\boldsymbol{x}_i'\hat{\boldsymbol{\beta}}\} \le \frac{2K}{n}\max_i\|\boldsymbol{x}_i\|$, and the last step invokes known probability bounds on order statistics using Assumption **??**.

The insight here is that $\hat{\boldsymbol{\beta}}$ is characterized by something that resembles a FOC up to a remainder term that vanishes faster than $\sqrt{n}$. Indeed, note that the above algebra implies the same bound of $o_P(n^{-1/2})$ on both directional derivatives themselves and therefore on all subderivaties. The way we will directly use the algebra later is to observe that it implies

$$\frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i\,\mathrm{sg}(y_i-\boldsymbol{x}_i'\hat{\boldsymbol{\beta}}) = o_P(n^{-1/2}). \tag{12}$$

Intuitively, you want to think of this as an analog to the OLS score equation (a.k.a. FOC)

$$\frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i(y_i-\boldsymbol{x}_i'\hat{\boldsymbol{\beta}}_{OLS}) = \mathbf{0}.$$

It's obviously not the same, but with some further argument it will turn out to be good enough.

9.3 Consistency

We will show consistency of $\hat{\boldsymbol{\beta}}$ by checking the assumptions of Theorem **??**. Specifically, we can easily see that Q_n is convex and we will show that its probability limit Q exists, is convex, and is uniquely minimized at the true value. Recall in particular that due to the convexity, it suffices to show $Q_n(\cdot)\xrightarrow{p}Q(\cdot)$ *pointwise*.

Define $S_n(\boldsymbol{\beta}) \equiv \frac{1}{n}\sum_{i=1}^n |y_i-\boldsymbol{x}_i'\boldsymbol{\beta}|$ and $Q_n(\boldsymbol{\beta}) \equiv S_n(\boldsymbol{\beta}) - S_n(\boldsymbol{\beta}_0)$, then $\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} Q_n(\boldsymbol{\beta})$.

Reparameterize $\boldsymbol{\delta} \equiv \boldsymbol{\beta} - \boldsymbol{\beta}_0$ and write

$$\begin{aligned} Q_n(\boldsymbol{\delta}) &= \frac{1}{n}\sum_{i=1}^{n}\left(|y_i - \boldsymbol{x}_i'\boldsymbol{\beta}| - |y_i - \boldsymbol{x}_i'\boldsymbol{\beta}_0|\right) \\ &= \frac{1}{n}\sum_{i=1}^{n}\left(|y_i - \mathbf{x}_i'\boldsymbol{\beta}_0 - \mathbf{x}_i'(\boldsymbol{\beta} - \boldsymbol{\beta}_0)| - |y_i - \mathbf{x}_i'\boldsymbol{\beta}_0|\right) \\ &= \frac{1}{n}\sum_{i=1}^{n}\left(|\varepsilon_i - \mathbf{x}_i'\boldsymbol{\delta}| - |\varepsilon_i|\right). \end{aligned}$$

By the Triangle and Cauchy-Schwarz inequalities,

$$|\varepsilon_i - \boldsymbol{x}_i'\boldsymbol{\delta}| \leq |\varepsilon_i| + |\boldsymbol{x}_i'\boldsymbol{\delta}| \Longrightarrow |\varepsilon_i - \boldsymbol{x}_i'\boldsymbol{\delta}| - |\varepsilon_i| \leq |\boldsymbol{x}_i'\boldsymbol{\delta}| \leq \|\boldsymbol{x}_i\|\|\boldsymbol{\delta}\|.$$

Because we only aim to show pointwise convergence, we can think of $\boldsymbol{\delta}$ as fixed. Assumption 2 now (more than) ensures that a LLN applies to Q_n, so we can write

$$\begin{aligned} Q_n(\boldsymbol{\delta}) \xrightarrow{p} Q(\boldsymbol{\delta}) &\equiv \mathbb{E}\left(|\varepsilon_i - \boldsymbol{x}_i'\boldsymbol{\delta}| - |\varepsilon_i|\right) \\ &= \mathbb{E}\left(\operatorname{sg}(\varepsilon_i - \boldsymbol{x}_i'\boldsymbol{\delta}) \times (\varepsilon_i - \boldsymbol{x}_i'\boldsymbol{\delta}) - \operatorname{sg}(\varepsilon_i) \times \varepsilon_i\right) \\ &= \mathbb{E}\left((\operatorname{sg}(\varepsilon_i - \boldsymbol{x}_i'\boldsymbol{\delta}) - \operatorname{sg}(\varepsilon_i)) \times (\varepsilon_i - \boldsymbol{x}_i'\boldsymbol{\delta})\right) - \mathbb{E}\left(\operatorname{sg}(\varepsilon_i) \times \boldsymbol{x}_i'\boldsymbol{\delta}\right) \\ &= \mathbb{E}\left((\operatorname{sg}(\varepsilon_i - \boldsymbol{x}_i'\boldsymbol{\delta}) - \operatorname{sg}(\varepsilon_i)) \times (\varepsilon_i - \boldsymbol{x}_i'\boldsymbol{\delta})\right) \qquad (13) \\ &\geq 0. \end{aligned}$$

Here, the third last step uses an add-and-subtract trick; the next step uses that by the Law of Iterated Expectations and Assumption **??**, the far right expectation equals zero; and the last step uses that the integrand in (**??**) is nonnegative (the terms being multiplied cannot be nonzero with opposing signs). Since $Q(\cdot) \geq 0$ and $Q(\mathbf{0}) = 0$ by inspection, $\boldsymbol{\delta} = \mathbf{0}$ minimizes Q.

We finally show that this minimum is unique and also derive the Hessian for future reference. A closer look reveals that the integrand in (**??**) is only strictly positive if either $0 < \varepsilon_i < \boldsymbol{x}_i'\boldsymbol{\delta}$ or $\boldsymbol{x}_i'\boldsymbol{\delta} < \varepsilon_i < 0$. With some book-keeping, the Law of Iterated Expectations yields (the expectation below is over $\boldsymbol{x}_i$, the integrals are conditional expectations)

$$\begin{aligned} Q(\boldsymbol{\delta}) &= \mathbb{E}\left(\int_0^{\boldsymbol{x}_i'\boldsymbol{\delta}} 2(\boldsymbol{x}_i'\boldsymbol{\delta} - \varepsilon) f(\varepsilon|\boldsymbol{x}_i) d\varepsilon \times \mathbf{1}\{\boldsymbol{x}_i'\boldsymbol{\delta} > 0\} + \int_{\boldsymbol{x}_i'\boldsymbol{\delta}}^{0} 2(\varepsilon - \boldsymbol{x}_i'\boldsymbol{\delta}) f(\varepsilon|\boldsymbol{x}_i) d\varepsilon \times \mathbf{1}\{\boldsymbol{x}_i'\boldsymbol{\delta} < 0\}\right) \\ &= \mathbb{E}\int_0^{\boldsymbol{x}_i'\boldsymbol{\delta}} 2(\boldsymbol{x}_i'\boldsymbol{\delta} - \varepsilon) f(\varepsilon|\boldsymbol{x}_i) d\varepsilon \end{aligned}$$

and therefore

$$\begin{aligned} \frac{\partial Q(\boldsymbol{\delta})}{\partial \boldsymbol{\delta}} &= 2\mathbb{E}\int_0^{\mathbf{x}_i'\boldsymbol{\delta}} \boldsymbol{x}_i' f(\varepsilon|\boldsymbol{x}_i) d\varepsilon \\ \frac{\partial^2 Q(\boldsymbol{\delta})}{\partial \boldsymbol{\delta} \partial \boldsymbol{\delta}'} &= 2\mathbb{E}\boldsymbol{x}_i\boldsymbol{x}_i' f(\boldsymbol{x}_i'\boldsymbol{\delta}|\boldsymbol{x}_i). \end{aligned}$$

Here, we take for granted that we can take derivatives inside the expectation. However, when replicating, note that the derivatives nominally contain additional terms that happen to equal zero. It is now easily verified that $\frac{\partial Q(\mathbf{0})}{\partial \boldsymbol{\delta}} = \mathbf{0}$ and $\frac{\partial^2 Q(\mathbf{0})}{\partial \boldsymbol{\delta} \partial \boldsymbol{\delta}'} = 2\mathbf{C}$; the latter is positive definite by Assumption 4.

An important take-away here is that, while Q_n is piecewise linear and therefore not differentiable, Q is smooth. In order to adapt m-estimator asymptotics, we will therefore have to expand Q and not Q_n and then argue that the approximation is good enough.

9.4 Asymptotic Distribution

Again, the main issue is that Q_n is not everywhere differentiable. Furthermore, while the kinks occur on a set of measure zero, the minimum is achieved on this set, i.e. the estimator 'seeks out the kinks' (just as solutions to LP occur at basic solutions). We will rather expand (the derivative of) Q. Of course, we will have to argue that we can pass from one to the other.

Define $\boldsymbol{g}_i(\boldsymbol{\beta}) \equiv \boldsymbol{x}_i \times \text{sg}(y_i - \boldsymbol{x}_i'\boldsymbol{\beta})$ and think of median regression as analogous to m-estimation based on the score equation

$$\mathbb{E}\boldsymbol{g}_i(\boldsymbol{\beta}_0) = \mathbf{0}. \tag{14}$$

To adapt the usual m-estimator arguments, we need three ingredients. First, letting $\bar{\boldsymbol{g}}_n \equiv \frac{1}{n}\sum_{i=1}^n \boldsymbol{g}_i$ as before, an empirical moment condition $\bar{\boldsymbol{g}}_n(\hat{\boldsymbol{\beta}}) = \mathbf{0}$ is *not* available, but we showed in (**??**) that

$$\sqrt{n}\bar{\boldsymbol{g}}_n(\hat{\boldsymbol{\beta}}) = o_P(1). \tag{15}$$

Next, it can be shown that

$$\boldsymbol{\nu}_n(\boldsymbol{\beta}) \equiv \sqrt{n}\left(\bar{\boldsymbol{g}}_n(\boldsymbol{\beta}) - \mathbb{E}\boldsymbol{g}_i(\boldsymbol{\beta})\right)$$

is stochastically equicontinuous at $\boldsymbol{\beta}_0$, i.e. $\hat{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}_0$ implies $\boldsymbol{\nu}_n(\hat{\boldsymbol{\beta}}) - \boldsymbol{\nu}_n(\boldsymbol{\beta}_0) \xrightarrow{p} \mathbf{0}$. This gives us

$$\sqrt{n}\big(\bar{\boldsymbol{g}}_n(\hat{\boldsymbol{\beta}}) - \mathbb{E}\boldsymbol{g}_i(\hat{\boldsymbol{\beta}})\big) = \sqrt{n}\big(\bar{\boldsymbol{g}}_n(\boldsymbol{\beta}_0) - \mathbb{E}\boldsymbol{g}_i(\boldsymbol{\beta}_0)\big) + o_P(1).$$

This step is not easy, i.e. we are brushing a lot of empirical process theory under the rug. In any case, substituting from (**??**) and (**??**) into the above yields

$$\sqrt{n}\mathbb{E}\boldsymbol{g}_i(\hat{\boldsymbol{\beta}}) = -\sqrt{n}\bar{\boldsymbol{g}}_n(\boldsymbol{\beta}_0) + o_P(1). \tag{16}$$

This is a very important step because it tightly links the sample 'score' at $\boldsymbol{\beta}_0$ to the corresponding population 'score' but at $\hat{\boldsymbol{\beta}}$. (The latter is random because $\hat{\boldsymbol{\beta}}$ is.) This is the "bridge" that allows us to effectively expand Q.

Third, as we already showed consistency, the Mean Value Theorem guarantees existence of $\bar{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}_0$ s.t.

$$\begin{aligned} \sqrt{n}\mathbb{E}\boldsymbol{g}_i(\hat{\boldsymbol{\beta}}) &= \sqrt{n}\mathbb{E}\boldsymbol{g}_i(\boldsymbol{\beta}_0) + \frac{\partial \mathbb{E}\boldsymbol{g}_i(\bar{\boldsymbol{\beta}})}{\partial \boldsymbol{\beta}'}\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \\ \Rightarrow \sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) &= \left[\frac{\partial \mathbb{E}\boldsymbol{g}_i(\bar{\boldsymbol{\beta}})}{\partial \boldsymbol{\beta}'}\right]^{-1} \sqrt{n}\mathbb{E}\boldsymbol{g}_i(\hat{\boldsymbol{\beta}}). \end{aligned}$$

Substituting from (**??**) and using $\bar{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}_0$, we get

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) = -\left[\frac{\partial \mathbb{E}\boldsymbol{g}_i(\boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}'}\right]^{-1} \sqrt{n}\bar{\boldsymbol{g}}_n(\boldsymbol{\beta}_0) + o_P(1).$$

Our final touch is to apply a plain vanilla CLT

$$\sqrt{n}\bar{\boldsymbol{g}}_n(\boldsymbol{\beta}_0) = \sqrt{n}\big(\bar{\boldsymbol{g}}_n(\boldsymbol{\beta}_0) - \mathbb{E}\boldsymbol{g}_i(\boldsymbol{\beta}_0)\big) \xrightarrow{d} N\big(\mathbf{0}, \mathbb{E}\left(\boldsymbol{g}_i(\boldsymbol{\beta}_0)\boldsymbol{g}_i(\boldsymbol{\beta}_0)'\right)\big)$$

and the simplifications

$$\begin{aligned} \mathbb{E}\left(\boldsymbol{g}_i(\boldsymbol{\beta}_0)\boldsymbol{g}_i(\boldsymbol{\beta}_0)'\right) &= \mathbb{E}\left(\mathbf{x}_i\mathbf{x}_i' \times \mathrm{sg}(y_i - \mathbf{x}_i'\boldsymbol{\beta}_0)^2\right) = \mathbb{E}\mathbf{x}_i\mathbf{x}_i' \\ \frac{\partial \mathbb{E}\boldsymbol{g}_i(\boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}'} &= 2\boldsymbol{C}. \end{aligned}$$

We conclude that

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \xrightarrow{d} N\big(\mathbf{0}, \boldsymbol{C}^{-1}\mathbb{E}(\boldsymbol{x}_i\boldsymbol{x}_i')\boldsymbol{C}^{-1}/4\big)$$

with further simplification to

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \xrightarrow{d} N\left(\mathbf{0}, \frac{(\mathbb{E}\boldsymbol{x}_i\boldsymbol{x}_i')^{-1}}{4f(0)^2}\right)$$

if $\epsilon_i \perp\!\!\!\perp \boldsymbol{x}_i$ (note that $\boldsymbol{C}$ factorizes in this case). As a sanity check, if we interpret the sample median as estimator in median regression of y_i on only a constant, we successfully recover its well-known asymptotic distribution.

Plug-in estimation of the above variances would involve estimation of conditional densities. Standard practice is to avoid this step and use *bootstrap* standard errors. This sagways to the next chapter. Note, however, that using bootstrap standard errors still relies on the above results because the very concept of standard error presupposes asymptotic normality. More subtly, even nonparametric percentile-t bootstrap confidence intervals (we will define these terms later) would require the above arguments for justification because (as we will discover) they rely on asymptotic normality as well.

10 Non- and Semiparametric Regression

First things first: In statistics and econometrics, the term **nonparametric** does not indicate the absence of parameters or even parameterizations. It rather means that the data generating process (d.g.p.) is specified up to an **infinite dimensional** unknown quantity. Informally, this quantity may even be referred to as parameter. Conversely, a parametric model is specified up to a **finite dimensional** parameter, at least where it matters.[17]

The difference cannot be overstated. In a parametric world, as $n \to \infty$, n necessarily becomes large relative to size of the model. Also, typically every observation is at least slightly relevant for every parameter. For broad classes of well-behaved parameters, this allows for $\sqrt{n}$-consistent estimation. In contrast, since it is obviously impossible to literally estimate an infinite dimensional model from finitely many observations, nonparametric methods typically rely on a sample size dependent coarsening of the space of d.g.p.'s under consideration. In other words, the model being fitted becomes more complex with sample size, and this typically precludes $\sqrt{n}$-consistency. This is especially easy to intuit with kernel density estimation, where the density at a given point is estimated off an effective sample that diverges but at a rate slower than n. Consistency here is at a rate corresponding to the square root of effective sample size. The rate at which models are coarsened is an important tuning parameter, and figuring out how to pick it is an important part of nonparametric estimation methodology.

Why bother? The rise of nonparametric (and semiparametric, and seminonparametric...) methods is partly due to concerns with credible identification. Parametric models typically combine substantive economic assumptions, convenience assumptions, and maybe also regularity conditions. The convenience assumptions are often distributional (e.g., logit vs probit) or assume linearity of a relationship that we truly "only" believe to be smooth monotone etc. Nonparametrics can in principle avoid many of those, leading to more credible conclusions, though at a cost that will become obvious. Other factors that gave rise to nonparametrics are the availability of larger data sets and computing resources and the (endogenous, of course) development of appropriate theory.

In these notes, we develop the baseline theory of kernel estimation in some detail. This is partly to illustrate how nonparametric asymptotics work. The treatment later becomes more breezy, but you will notice that many ideas reappear.

[17] Leaving the distribution of an error term ϵ_i unspecified when a CLT will clearly apply, as with standard OLS, does not count, and so this example is considered parametric. However, if one actually wanted to estimate the density of ϵ_i in this same model, the estimation problem would be semiparametric.

10.1 Kernel Density Estimation

10.1.1 Overview

Say we want to estimate the distribution of a random variable x_i. An obvious estimator for the c.d.f. F, namely the empirical distribution, was already discussed and used in the bootstrap. The analogous estimator of the density (assuming of course that one exists) is the empirical probability mass function (p.m.f.), which consists of at most n mass points. This estimator is technically not even a density. In contrast to the empirical c.d.f., it is also often useless: The empirical c.d.f. consistently estimates the population c.d.f., but the empirical p.m.f. consistently estimates the density almost nowhere.

One fix to this is kernel density estimation, that is, the density is estimated by a weighted average of nearby mass points of the empirical p.m.f. This can be intuited as smoothing out histograms or as estimating F' by some smoothed arc slope (discrete derivative) of F_n. The basic idea is obvious enough, but there are many ways to do the smoothing. In particular, one will have to choose a **kernel** and a **bandwidth**. In practice, choice of kernel is often not too consequential, but the bandwidth is the "coarsening parameter" mentioned above and its choice tends to matter a lot.

In the scalar case, a kernel density estimator of $f(x)$ can be written as

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} k\left(\frac{x_i - x}{h}\right),$$

where $k : \boldsymbol{R} \mapsto \boldsymbol{R}$ is the kernel function and h is the bandwidth. For a simple example, let $k(t) = 1/2 \cdot \mathbf{1}\{|t| \leq 1\}$, i.e. the **uniform kernel**, then one can write

$$\hat{f}(x) = \frac{1}{2nh} \sum_{i=1}^{n} \mathbf{1}\{x - h \leq x_i \leq x + h\} = \frac{1}{2h} P_n([x - h, x + h]),$$

where P_n is the empirical p.m.f.

This particular estimator is a step function. This can be avoided by using continuous kernels. But the uniform kernel makes some of the trade-offs in kernel estimation very obvious. In particular, we estimate the density at x by averaging the empirical p.m.f. over $[x - h, x + h]$. A larger choice of bandwidth h means that we average over more observations, and so the variance of our estimator will decrease. But it also means that we average over a set where the true density may be increasingly different from the true density at x, so the bias will increase. This is quite like with histograms: If we "oversmooth" (too large bin size), the thing becomes eventually constant, if we "undersmooth" (too small bin size), it becomes too wiggly and will eventually include unwanted zeros.

In practice, we attempt to choose h to intelligently resolve this bias/variance trade-off. For example, in an "oracle" situation where we know the density to be estimated, we can find the choice of h that minimizes mean square error

$$\mathbb{E}\big(\hat{f}(x) - f(x)\big)^2$$

at some x of interest. It is intuitive (we'll do algebra later) that the h optimizing this will (i) decrease with n (therefore, you will often see the notation h_n), (ii) depend on how smooth the true density is at x, with a smoother true density calling for more smoothing. If we are interested in overall performance, we could also minimize the mean integrated square error

$$\int_{-\infty}^{\infty} \mathbb{E}\big(\hat{f}(x) - f(x)\big)^2 dx,$$

and the solution will again decrease with n but also with some measure of overall smoothness of f.

The obvious catch is that we don't know f, and therefore the "oracle" bandwidth is not known in practice. This problem gave rise to a considerable literature and we will think about it much more. Also, the optimal choice of h as function of n will imply that bias and variance are of the same order in n. This is simply because we want to minimize the larger of two rates, and the maximum of two functions is typically minimized at a point of equality. So the bandwidth choice that makes for the "best" estimator also makes for an asymptotic distribution that is noncentered, much complicating inference. Therefore, it is common to choose a "too small" bandwidth so that variance dominates bias and the asymptotic distribution is centered. This is called [deliberate] **undersmoothing**.

We will next formalize these intuitions. The most important addition will be that I so far picked a very particular kernel. While bandwidth choice tends to be more consequential than kernel choice, we will formally think about both, and we will certainly generalize beyond uniform kernels (which are not common in practice).

10.1.2 Formal Properties of Kernels

A kernel function $k : \boldsymbol{R} \mapsto \boldsymbol{R}$ must integrate to 1: $\int_{-\infty}^{\infty} k(u)du = 1$. No other property is strictly required, but in practice kernel functions are usually symmetric: $k(u) = k(-u)$.[18] The most popular kernels are furthermore nonnegative, i.e. $k(u) \geq 0$ for all u, but (maybe surprising at first) it can be advisable to use kernels that are not. The **order** of a kernel is the index of its first nonzero moment. Thus, defining $\kappa_j(k) = \int_{-\infty}^{\infty} u^j k(u)du$, the order of k equals $\min\{j > 0 : \kappa_j(k) \neq 0\}$. As we restrict attention to symmetric kernels and all odd moments of symmetric kernels are zero, any kernel that we look at has an even order that is at least 2. A kernel has **higher order** if its order strictly exceeds 2. Note that a nonnegative kernel necessarily has $\kappa_2 > 0$ and so cannot be higher order.

Any nonnegative kernel can be interpreted as a probability density, and so any kernel density estimate using a nonnegative kernel is a weighted average in the everyday sense (i.e., excluding negative

[18] There are exceptions to this if estimation of a density near the boundaries of the random variable's support is the goal. We will ignore that.

weights) of the empirical p.m.f. Two simple kernels are the uniform and Gaussian ones:

$$\begin{aligned} k^{uni}(u) &= \frac{1}{2}\mathbf{1}\{|u| \leq 1\} \\ k^{gau}(u) &= (2\pi)^{-1/2}\exp(-u^2/2). \end{aligned}$$

With these or any other nonnegative kernels, the kernel density estimator can be visually intuited as taking each of the n realizations of x_i and replacing it with $1/n$ times a (uniform, standard normal,...) p.d.f. centered at that realization. This makes it immediately obvious that $\hat{f}(x)$ is itself a proper p.d.f., but also that the r.v. described by $\hat{f}$ is a mean-preserving spread of the r.v. described by the empirical distribution. In contrast, with higher-order kernels, $\hat{f}(x)$ can take negative values and therefore need not be a density; but as we will see, it replicates the second (and possibly higher, depending on order of the kernel) moment of the empirical distribution.

While intuitively obvious for nonnegative kernels, we next show that $\hat{f}$ necessarily integrates to 1. To see this formally, first write

$$1 = \int_{-\infty}^{\infty} k(u)du = \int_{-\infty}^{\infty} \frac{1}{h}k\left(\frac{x_i - x}{h}\right)dx,$$

where the last step uses the change-of-variables $u(x) = \frac{x_i - x}{h}$. Next,

$$\int_{-\infty}^{\infty} \hat{f}(x)dx = \int_{-\infty}^{\infty} \frac{1}{nh}\sum_{i=1}^{n} k\left(\frac{x_i - x}{h}\right)dx = \frac{1}{n}\sum_{i=1}^{n}\int_{-\infty}^{\infty} \frac{1}{h}k\left(\frac{x_i - x}{h}\right)dx = 1.$$

An obvious consequence is that if k is nonnegative, then $\hat{f}(x)$ is a valid density. Because this is a desirable property – you probably don't want to report negative probabilities to clients – with higher order kernels we may force it by reporting

$$\widetilde{f}(x) = \frac{\left|\hat{f}(x)\right|_+}{\int_{-\infty}^{\infty}\left|\hat{f}(x)\right|_+ dx}$$

where $|t|_+ = \max\{t, 0\}$. The effect of this manipulation will vanish asymptotically under any assumptions that justify kernel density estimation to begin with.

The "estimator" of $\mathbb{E}(x_i)$ implied by our density estimate is what you'd expect:

$$\begin{aligned} \int_{-\infty}^{\infty} x\hat{f}(x)dx &= \frac{1}{n}\sum_{i=1}^{n}\int_{-\infty}^{\infty} x\frac{1}{h}k\left(\frac{x_i - x}{h}\right)dx \\ &= \frac{1}{n}\sum_{i=1}^{n}\int_{-\infty}^{\infty}(x_i - uh)k(u)du \\ &= \frac{1}{n}\sum_{i=1}^{n}\int_{-\infty}^{\infty} x_i k(u)du - \frac{1}{n}\sum_{i=1}^{n}\int_{-\infty}^{\infty} uhk(u)du \\ &= \frac{1}{n}\sum_{i=1}^{n} x_i \underbrace{\int_{-\infty}^{\infty} k(u)du}_{=1} - \frac{1}{n}\sum_{i=1}^{n} h\underbrace{\int_{-\infty}^{\infty} uk(u)du}_{=0} = \overline{x}, \end{aligned}$$

where we used $u = \frac{x_i - x}{h} \Leftrightarrow x = x_i - uh$. In words, the expectation induced by the estimated density is just the sample average.

The same is not true for the estimated density's second uncentered moment and, therefore, its variance:

$$\begin{aligned}
\int_{-\infty}^{\infty} x^2 \hat{f}(x)dx &= \frac{1}{n}\sum_{i=1}^{n}\int_{-\infty}^{\infty} x^2 \frac{1}{h}k\left(\frac{x_i - x}{h}\right)dx \\
&= \frac{1}{n}\sum_{i=1}^{n}\int_{-\infty}^{\infty} (x_i - uh)^2 k(u)du \\
&= \frac{1}{n}\sum_{i=1}^{n}\int_{-\infty}^{\infty} x_i^2 k(u)du + \frac{1}{n}\sum_{i=1}^{n}\int_{-\infty}^{\infty} (uh)^2 k(u)du - \frac{2}{n}\sum_{i=1}^{n}\int_{-\infty}^{\infty} x_i huk(u)du \\
&= \frac{1}{n}\sum_{i=1}^{n} x_i^2 \underbrace{\int_{-\infty}^{\infty} k(u)du}_{=1} + \frac{1}{n}\sum_{i=1}^{n} h^2 \underbrace{\int_{-\infty}^{\infty} u^2 k(u)du}_{=\kappa_2} - \frac{2}{n}\sum_{i=1}^{n} x_i h \underbrace{\int_{-\infty}^{\infty} uk(u)du}_{=0} \\
&= \frac{1}{n}\sum_{i=1}^{n} x_i^2 + h^2\kappa_2.
\end{aligned}$$

Thus, with nonnegative kernels, the density estimator implies a variance that exceeds the sample variance by $h^2\kappa_2$ (and the df-adjusted estimator by slightly more). In particular, the variance corresponding to the estimated density coincides with the sample variance for higher-order but not for nonnegative kernels. For nonnegative kernels, this is also clear from the representation mentioned earlier, i.e. $\hat{f}$ describes a mean-preserving spread of the empirical p.m.f.

10.1.3 Asymptotic Bias, Variance, and MSE

We next develop expressions for the asymptotic bias and variance of $\hat{f}$. The goal is to (i) figure out a good choice of k and h and (ii) conduct inference. Getting there requires modest algebra.

We start by writing

$$\begin{aligned}
\mathbb{E}\hat{f}(x) = \mathbb{E}\left(\frac{1}{nh}\sum_{i=1}^{n} k\left(\frac{x_i - x}{h}\right)\right) &= \mathbb{E}\left(\frac{1}{h}k\left(\frac{x_i - x}{h}\right)\right) \\
&= \int_{-\infty}^{\infty} \frac{1}{h}k\left(\frac{t - x}{h}\right) f(t)dt = \int_{-\infty}^{\infty} k(u)\, f(x + hu)du.
\end{aligned}$$

The last expression has a clear intuition: $\hat{f}(x)$ effectively averages the estimated density over a neighborhood of order h of x. Averages being unbiased estimators of the corresponding expectations, $\mathbb{E}\hat{f}(x)$ is the true expectation corresponding to this average. Note that this argument was not yet asymptotic; however, without asymptotic approximations we are typically stuck here.

Next, a Taylor expansion to the order ν of our kernel k yields

$$f(x + hu) = f(x) + f'(x)hu + \frac{1}{2}f''(x)h^2u^2 + \cdots + \frac{1}{\nu!}f^{(\nu)}(x)h^\nu u^\nu + o(h^\nu)$$

assuming the derivatives up to $f^{(\nu+1)}$ exist. This is not "just a regularity condition": Higher-order kernels can only be used if one is willing to assume that the true density is correspondingly smooth, and this assumption gets seriously restrictive rather quickly. Next, we use linearity of integrals to get

$$\begin{aligned}
\mathbb{E}\hat{f}(x) &= \int_{-\infty}^{\infty} f(x)k(u)\,du + \int_{-\infty}^{\infty} k(u)\,f'(x)hu du + \frac{1}{2}\int_{-\infty}^{\infty} k(u)\,f''(x)h^2u^2du \\
&\quad + \cdots + \frac{1}{\nu!}\int_{-\infty}^{\infty} k(u)\,f^{(\nu)}(x)h^\nu u^\nu du + o(h^\nu) \\
&= f(x) + f'(x)h\int_{-\infty}^{\infty} uk(u)\,du + \frac{1}{2}f''(x)h^2\int_{-\infty}^{\infty} u^2k(u)\,du \\
&\quad + \cdots + \frac{1}{\nu!}f^{(\nu)}(x)h^\nu\int_{-\infty}^{\infty} u^\nu k(u)\,du + o(h^\nu) \\
&= f(x) + \frac{1}{2}f''(x)h^2\kappa^2 + \cdots + \frac{1}{\nu!}f^{(\nu)}(x)h^\nu\kappa_\nu + o(h^\nu) \\
&= f(x) + \frac{1}{\nu!}f^{(\nu)}(x)h^\nu\kappa_\nu + o(h^\nu)
\end{aligned}$$

because the kernel is of order ν. This makes it immediately obvious why one might *in theory* want to use higher order kernels. For the most salient case of nonnegative kernels, we have

$$\mathbb{E}\hat{f}(x) = f(x) + \frac{1}{2}f''(x)h^2\kappa_2 + O(h^4)$$

and hence

$$\operatorname{bias}\big(\hat{f}(x)\big) = \frac{1}{2}f''(x)h^2\kappa_2 + O(h^4).$$

We see that for a given (nonnegative) kernel, the bias is of order $O(h^2)$. Furthermore, it increases (in absolute value) with the dispersion of the kernel and also with the curvature of f at x. Its sign depends on whether f is convex or concave at x and is of lower order where the curvature of f vanishes. This is as it should be: We effectively average over a symmetric neighborhood of order h of x. As this neighborhood becomes small, f is well approximated on it as either linear or parabolic. Under the former approximation, the average becomes unbiased at the relevant rate of localization. Else, the bias is positive [negative] if f is locally convex [concave], meaning that it is locally above [below] the tangent.

Next,

$$\begin{aligned}
\operatorname{var}\big(\hat{f}(x)\big) &= \operatorname{var}\left(\frac{1}{nh}\sum_{i=1}^{n} k\left(\frac{x_i - x}{h}\right)\right) \\
&= \frac{1}{nh^2}\operatorname{var}\left(k\left(\frac{x_i - x}{h}\right)\right) \\
&= \frac{1}{nh^2}\left(\mathbb{E}\left(k\left(\frac{x_i - x}{h}\right)^2\right) - \left(\mathbb{E}\left(k\left(\frac{x_i - x}{h}\right)\right)\right)^2\right) \\
&= \frac{1}{nh^2}\mathbb{E}\left(k\left(\frac{x_i - x}{h}\right)^2\right) - \frac{1}{n}\left(\mathbb{E}\left(\frac{1}{h}k\left(\frac{x_i - x}{h}\right)\right)\right)^2 \\
&= \frac{1}{nh^2}\mathbb{E}\left(k\left(\frac{x_i - x}{h}\right)^2\right) - O(1/n),
\end{aligned}$$

using that, from above, $\mathbb{E}\left(\frac{1}{h}k\left(\frac{x_i - x}{h}\right)\right) = \mathbb{E}\hat{f}(x) = f(x) + o(1)$. Plugging in from the following Taylor expansion:

$$\begin{aligned}
\frac{1}{h}\int_{-\infty}^{\infty}\left(k\left(\frac{z-x}{h}\right)\right)^2 f(z)dz &= \int_{-\infty}^{\infty}\left(k\left(u\right)\right)^2 f(x+uh)du \\
&= \int_{-\infty}^{\infty}\left(k\left(u\right)\right)^2 \left(f(x) + O(h)\right)du = f(x)\int_{-\infty}^{\infty} k\left(u\right)^2 du + O(h) = f(x)R(k) + O(h),
\end{aligned}$$

where the last step defines the **roughness** $R(g) \equiv \int_{-\infty}^{\infty} g\left(t\right)^2 dt$ of a function $g : \boldsymbol{R} \to \boldsymbol{R}$, we find

$$\operatorname{var}\big(\hat{f}(x)\big) = \frac{1}{nh}\left(f(x)R(k) + O(h)\right) + O(1/n) = \frac{1}{nh}f(x)R(k) + O(1/n). \tag{17}$$

While we avoid writing h_n to economize on subscripts, we know that we'll send $h \to 0$ as $n \to \infty$, so $O(1/n) = o(1/(nh))$ and the first term dominates. We also find that the variance is of order $O(1/(nh))$ and is also proportional to the density at x.

Next, the quality of our estimator is plausibly assessed by its mean squared error (MSE):

$$\operatorname{MSE}\big(\hat{f}(x)\big) \equiv \big(\operatorname{bias}\big(\hat{f}(x)\big)\big)^2 + \operatorname{var}\big(\hat{f}(x)\big) = \left(\frac{1}{\nu!}f^{(\nu)}(x)h^{\nu}\kappa_{\nu} + o(h^{\nu})\right)^2 + \frac{1}{nh}f(x)R(k) + O(1/n).$$

We will always choose h s.t. in the above, all terms other than the *asymptotic mean square error*

$$\operatorname{AMSE}\big(\hat{f}(x)\big) \equiv \left(\frac{1}{\nu!}f^{(\nu)}(x)h^{\nu}\kappa_{\nu}\right)^2 + \frac{1}{nh}f(x)R(k)$$

are dominated. However, which of these terms dominates depends on our exact choice of h. Note that as long as $h \to 0$ but slow enough that $nh \to \infty$, AMSE vanishes and the estimator is thus (pointwise) consistent. These requirements make sense intuitively: Bandwidth must vanish for bias to vanish, but slowly enough so that in expectation the sample size effectively used at x diverges.

10.1.4 Asymptotically Optimal Bandwidth

We could try to specify k and h to optimize $\text{AMSE}(\hat{f}(x))$. However, the solution would usually depend on x. If interest is in overall performance of the estimator, we usually integrate to get the asymptotic mean integrated square error

$$\text{AMISE}(\hat{f}) = \int_{-\infty}^{\infty} \left[\left(\frac{1}{\nu!} f^{(\nu)}(x) h^\nu \kappa_\nu \right)^2 + \frac{1}{nh} f(x) R(k) \right] dx = \left(\frac{1}{\nu!} h^\nu \kappa_\nu \right)^2 R(f^{(\nu)}) + \frac{R(k)}{nh}.$$

In the salient case of nonnegative kernels, this is

$$\text{AMISE}(\hat{f}) = \frac{1}{4} h^4 \kappa_2^2 R(f'') + \frac{R(k)}{nh}.$$

The obvious next question is optimization of kernel and bandwidth. We will first think of optimization of bandwidth given a specific kernel. Then we can write

$$\frac{d}{dh}\text{AMISE}(\hat{f}) = 2\nu \left(\frac{\kappa_\nu}{\nu!} \right)^2 R(f^{(\nu)}) h^{2\nu-1} - \frac{R(k)}{nh^2} \stackrel{!}{=} 0,$$

which is solved by

$$h^* = \left(\frac{R(k)}{2n\nu \left(\frac{\kappa_\nu}{\nu!} \right)^2 R(f^{(\nu)})} \right)^{1/(2\nu+1)} = \left(\frac{R(k)}{2\nu \left(\frac{\kappa_\nu}{\nu!} \right)^2} \right)^{1/(2\nu+1)} R(f^{(\nu)})^{-1/(2\nu+1)} n^{-1/(2\nu+1)}.$$

Here, the last step just isolates some terms of interest. In particular, the optimal bandwidth is proportional to $n^{-1/(2\nu+1)}$ and therefore vanishes more slowly (in terms of rate and not just constant!) as we move to higher order kernels. This comes from the fact that higher order kernels have lower order bias, so the bias-variance trade-off changes. Of course, this is in turn because higher order kernels impose more smoothness on f – no free lunch here.

Next, we can evaluate $\text{AMISE}(\hat{f})$ at the optimal bandwidth and therefore evaluate the value of our optimization problem. After simplification, this yields

$$\text{AMISE}^* = (1+2\nu) \left(\frac{R(k)^{2\nu} \kappa_\nu^2 R(f^{(\nu)})}{(2\nu)^{2\nu} (\nu!)^2} \right)^{1/(2\nu+1)} n^{-2\nu/(2\nu+1)}.$$

This is not a further approximation: The two components of AMISE are of the same order and the addition of two parts is reflected in the + sign. We already mentioned that intuitively, both terms are of the same order at the optimum, but also that this will lead to complications later.[19]

We see that the optimal bandwidth and rate of convergence depend on the order of the kernel. For nonnegative kernels, we get

$$\begin{aligned} h^*\|_{\nu=2} &= \left(\frac{R(k)}{\kappa_2^2 R(f'')} \right)^{1/5} n^{-1/5} \\ \text{AMISE}^*\|_{\nu=2} &= \frac{5}{4} \left(R(k)^4 \kappa_2^2 R(f'') \right)^{1/5} n^{-4/5} \end{aligned}$$

[19] We note in passing that this rate can be shown to be minimax, so it cannot in general be improved upon by a fundamentally different estimation strategy.

At the other extreme, as $\nu \to \infty$, the optimal bandwidth vanishes ever more slowly and the AMISE approaches order n^{-1}. So if the density is infinitely smooth, we can in principle approximate the parametric rate! Of course, this is a purely theoretical consideration.

10.1.5 Practical Approaches to Bandwidth Choice

The obvious next idea would be to simply calculate the optimal bandwidth. We cannot do that because the answer depends on $R(f^{(\nu)})$, which we don't know. The next obvious idea would be to pre-estimate $R(f^{(\nu)})$ and then choose the bandwidth accordingly. However, and unlike with the vaguely similar step of pre-estimating a variance matrix in two-stage GMM, we encounter a regress here: Nonparametrically estimating the roughness of the 2^{nd} or higher derivative of a density is not possible at fast enough rates.

Three possible ways forward are as follows:

Silverman's Rule of Thumb The most popular approach is to resort to a parametric estimator of $R(f^{(\nu)})$ that forces f to be in a narrow parametric class, ideally one such that the implied estimator of $R(f^{(\nu)})$ is easily computed. One such class is the class of centered normal densities, parameterized only by the variance, which we can estimate by the sample variance (Silverman, 1986). Some algebra reveals that the resultant bandwidth scales with the sample standard deviation $\hat{\sigma}$:

$$h^{sil} = \hat{\sigma} c_k n^{-1/(2\nu+1)},$$

where tabulations of the kernel-dependent constant c_k are widely available. Indeed, canned packages will readily implement this rule for you, often as a default option.

Plug-In Estimation of $R(f'')$ Focusing on the case of nonnegative kernels, the only hurdle in estimating

$$h^* = \left(\frac{R(k)}{\kappa_2^2 R(f'')} \right)^{1/5} n^{-1/5}$$

is estimation of $R(f'')$. This can in principle be done with a plug-in estimator $\hat{R}(f'') = R(\hat{f}'')$. But there is a catch: As we will discuss later, optimal bandwidths depend on the estimand. Indeed, the optimal bandwidth h^* will typically be too small (by order compared to n) to even guarantee consistency of the implied estimator of f''. This gives rise to the following iterative algorithm:[20] (i) Initialize with a bandwidth $\widetilde{h}$, e.g. using Silverman's Rule of Thumb. (ii) Compute $R(\hat{f}'')$ using an inflated bandwidth, e.g. $\overline{h}(\widetilde{h}, n) = \widetilde{h} n^{1/10}$. (iii) Estimate the optimal bandwidth by plugging into the expression for h^*. (iv) If the estimate is close to $\widetilde{h}$, declare convergence. Else, iterate using the most

[20] This follows Gasser, Kneip, and Köhler (1991). See Sheather and Jones (1991) for a related proposal.

recent estimate as $\widetilde{h}$. Similar proposals exist for related settings although not (to my knowledge) for very complicated ones. They perform well for densities with moderate $R(f'')$ but tend to oversmooth with very wiggly densities (Jones, Marron, and Sheather, 1996).

Cross-Validation A slightly different approach is to directly estimate the MISE and then choose the bandwidth to minimize the estimate. Write

$$\mathrm{MISE}(h) = \int \mathbb{E}\big(\hat{f}(x) - f(x)\big)^2 dx = \mathbb{E}\left(\int \hat{f}(x)^2 dx\right) - 2\mathbb{E}\left(\int \hat{f}(x) f(x) dx\right) + \int f(x)^2 dx,$$

where for unity of notation I suppress that $\hat{f}$ depends on h. The third term on the r.h.s. does not depend on h and can be ignored. The other terms can be estimated, leading to the cross-validation criterion

$$\mathrm{CV}(h) = \frac{1}{n}\sum_{i=1}^{n} \int \hat{f}_{-i}(x)^2 dx - \frac{2}{n}\sum_{i=1}^{n} \hat{f}_{-i}(x_i) \approx \int \hat{f}(x)^2 dx - \frac{2}{n}\sum_{i=1}^{n} \hat{f}_{-i}(x_i)$$

where $\hat{f}_{-i}$ is the **leave-one-out** estimator of f that discards observation x_i. Note the following details:

- We seem to be estimating the ISE, i.e. the realization of MISE caused by our data, but it can be shown that $\mathrm{CV}(h)$ is unbiased and consistent for MISE; hence, this procedure is also called unbiased cross-validation.

- In the second sum above, the estimated density at x_i is computed not using x_i itself. What is the idea here? Letting f be the true density of x_i and g be some known function, it would be obvious to estimate $\int g(x) f(x) dx = \mathbb{E}g(x_i)$ through the sample analog $\frac{1}{n}\sum_{i=1}^{n} g(x_i)$. However, in our case we have $g = \hat{f}$ and therefore the nontrivial complication that the function evaluated at x_i depends on x_i beyond taking x_i as an argument. The leave-one-out estimator shuts down that dependency. This is a trick that you will see again and again, including in neighboring fields like Machine Learning.

- Strictly speaking, this logic gives rise to the expression left of $\approx$ above (Bowman, 1984). However, the computationally much simpler r.h. expression (Hall, 1983) is asymptotically equivalent and is now the industry standard; e.g., it is implemented in R option `bw="bw.ucv"` in `density`. Note that the leave-one-out estimator is avoided only for computation of the integral.

The cross-validation estimator $\hat{h}$ of the optimal bandwidth is consistent under fairly general conditions and also in more general settings than considered here. Indeed, cross-validation is by now an extremely widely used tool wherever so-called tuning parameters (here: a bandwidth) have to be chosen. The rate of convergence, however, is painfully slow, with $\frac{\hat{h}-h^*}{h^*} = O_P(n^{-1/10})$ in this example.

10.1.6 Asymptotically Optimal Kernel

The choice of kernel order is principally constrained by assumptions one is willing to make on smoothness. However, and despite results about order of bias above, I do not advise to use the highest order kernel that you can maybe justify. Indeed, your default should be to use nonnegative kernels. We have more to say on the choice of kernels within a given order. In particular, the kernel affects $\text{AMISE}^*(\hat{f})$ through a positive power of $R(k)^\nu \kappa_\nu$. Let's normalize $\kappa_\nu = 1$. This is justified because optimal kernels are "identified" only up to horizontal scale; the bandwidth can be used to implicitly rescale any kernel along the u-axis and thereby freely rescale one nonzero moment. Hence, the asymptotically optimal kernel solves

$$\min R(k) \quad \text{s.t.} \quad \int_{-\infty}^{\infty} k(u)du = 1; \int_{-\infty}^{\infty} u^j k(u)du = 0, j = 1, \ldots, \nu - 1; \int_{-\infty}^{\infty} u^\nu k(u)du = 1,$$

where the choice of 1 in the last constraint makes the problem well-defined but is substantively w.l.o.g. (see homework). Müller (1984) showed that the solution to this problem is the Epanechnikov kernel. For $\nu = 2$, this kernel is[21]

$$k^{epa}(u) = \frac{3}{4}\left|1 - u^2\right|_+.$$

Note this kernel's simple geometry as the positive truncation of a parabola. Higher order kernels can be generated from nonnegative kernels by multiplication with certain polynomials, yielding the higher order analogs

$$\begin{aligned} k_4^{epa}(u) &= \frac{15}{8}\left(1 - \frac{7}{3}u^2\right)k^{epa}(u) \\ k_6^{epa}(u) &= \frac{175}{64}\left(1 - 6u^2 + \frac{33}{5}u^4\right)k^{epa}(u). \end{aligned}$$

The *efficiency* of any other kernel is defined as

$$\text{eff}(k) = \left(\frac{\text{AMISE}^*(k)}{\text{AMISE}^*(k_\nu^{epa})}\right)^{\frac{1+2\nu}{2\nu}},$$

where ν is the order of the kernel under consideration. This number always greater than 1 and is taken to a power that compensates the rate at which AMISE decays with n. This makes for a simple interpretation: The AMISE of the Epanechnikov kernel with n observations equals the AMISE of any other kernel k with $\text{eff}(k) \times n$ observations. The efficiency loss can therefore be interpreted as excess data needed compared to the efficient kernel.

Numerically, efficiency losses of many kernels are tiny. The simplest kernels, i.e. Gaussian and uniform, stand out as inefficient among frequently used kernels and yet their efficiencies are 1.05

[21] This presentation follows my preferred convention to have bounded kernels be supported on $[-1, 1]$. This kernel has $\kappa^2 = 1/5$, and so the version that respects the normalization $\kappa_\nu = 1$ stretches it out horizontally by $\sqrt{5}$ as in the Stata manual.

respectively 1.08. This suggests – and it is also borne out in practice – that no "reasonable" kernel choice should hurt too much and that maybe bandwidth is more important. That said, it is typically easy to just use the Epanechnikov kernel, which is available in any canned kernel density estimation codes (e.g. R `density`, Matlab `fitdist`) and is frequently the default (e.g. Stata `kdensity`).[22]

10.1.7 Asymptotic Distribution and Inference

The asymptotic distribution of the estimator is straightforward to derive but, compared to the parametric case, raises its own novel problems. Recalling we assume i.i.d. data, we can use (**??**) and invoke a Central Limit Theorem to get

$$\sqrt{nh}\big(\hat{f}(x) - \mathbb{E}\hat{f}(x)\big) \xrightarrow{d} N(0, f(x)R(k)).$$

So far, so good, but note that I centered $\hat{f}(x)$ at its own expectation, not at the true value of $f(x)$. Whether this matters depends on the bandwidth. In particular, an optimal bandwidth choice implies that bias and variance are of the same order, and then this difference does matter and leads to a noncentered asymptotic distribution.

To work this out a bit, recall that up to very good approximation, the bias is $\frac{1}{\nu!}f^{(\nu)}(x)h^{\nu}\kappa_{\nu}$. Suppose that we optimally choose $h = Cn^{-1(2\nu+1)}$ for some constant C. Then we can solve for $\sqrt{nh}h^{\nu} = C^{\nu+1/2}$ and write

$$\sqrt{nh}\big(\hat{f}(x) - f(x)\big) \xrightarrow{d} N\left(\frac{\kappa_{\nu}C^{\nu+1/2}}{\nu!}f^{(\nu)}(x), f(x)R(k)\right),$$

in particular for nonnegative kernels

$$\sqrt{nh}\big(\hat{f}(x) - f(x)\big) \xrightarrow{d} N\left(\kappa^2 C^{3/2}f''(x)/2, f(x)R(k)\right).$$

This asymptotic bias term cannot be pre-estimated. We might of course try to estimate $f''(x)$, adjust our estimator by a plug-in estimate of its bias, and hope that we now get a centered limiting distribution. But this would require to estimate $f''(x)$ with a bias that is of lower order than the above. In fact, we will later see that the bias is of the same (for the same bandwidth) or higher (for the bandwidth that is actually optimal for derivative estimation) order, and the MSE of the derivative estimator is always of higher order. So this is not going to happen.

In practice, researchers (including this lecturer) usually resolve this by assuming the bias is zero to the relevant order of approximation, thus they conduct inference as if

$$\sqrt{nh}\big(\hat{f}(x) - f(x)\big) \xrightarrow{d} N(0, f(x)R(k)).$$

[22] These results are specific to density estimation. The Gaussian kernel is a popular default because it is applicable and "good enough" in a wide range of applications.

This is justified by claiming that one's bandwidth is **undersmoothed**, i.e. smaller than the optimal one, which increases variance and decreases bias. Under this assumption, the asymptotic approximation is indeed formally justified. However, this trick has at least two downsides:

- The confidence interval can always be interpreted as set-valued estimator. In contrast to parametric settings, this set estimator now converges at a slower rate than the best available point estimator. So in large samples, this estimator is very large compared to the optimal estimator $\pm$ two standard errors (but of course, the latter is not a valid confidence interval!) and is also not guaranteed to contain the optimal estimator. In practice, if we report this CI, we should also report the corresponding estimator, but then of course we are sacrificing an order of estimation precision in order to be able to do inference.

- While we claim a rate for our bandwidth, in a given application, we of course choose a fixed bandwidth. In principle, we could claim for any such bandwidth that it is embedded in a sequence which converges at our rate of choice. This issue arises elsewhere too but is especially tricky here. There is no clear rule on what we're allowed to do, but for example, using a Rule-of-Thumb bandwidth and then claiming asymptotically centered estimates would be overstepping. More generally, while there is a neat theory for choosing the optimal bandwidth, I am not aware of useful guidance for choosing an optimally undersmoothed bandwidth.

Assume now that we did undersmooth and hence have available the centered normal limiting distribution. Then conducting pointwise inference is straightforward. (Uniform and simultaneous confidence bands are not! We omit them here.) In particular, we could report

$$CI = \left[\hat{f}(x) - \Phi^{-1}(1-\alpha/2) \times \sqrt{\frac{\hat{f}(x)R(k)}{nh}}, \hat{f}(x) + \Phi^{-1}(1-\alpha/2) \times \sqrt{\frac{\hat{f}(x)R(k)}{nh}}\right].$$

If $\hat{f}(x)$ is close to zero, this can become awkward as the left boundary may become negative. This can be circumvented by explicitly inverting a hypothesis test:

$$CI = \left\{f : \sqrt{nh}\frac{\left|\hat{f}(x) - f\right|}{\sqrt{fR(k)}} \leq \Phi^{-1}(1-\alpha/2)\right\}.$$

This is computationally more burdensome as the critical value must be recomputed at each f. It also must be taken with a grain of salt because in expectation, it really matters only if the sampling distribution has considerable skewness, in which case a CLT hasn't "kicked in" yet; this undermines justification of Φ^{-1} for critical values.

Notice finally that these confidence intervals are pointwise in a very strong sense. Compared to confidence bands for f, they not only abstract from the multiple hypothesis testing problem inherent

in the latter, but they also are not uniformly valid over regions of parameter space where $f(x) \to 0$ (because the variance term then vanishes). Thus, a confidence band constructed from them may not even be valid uniformly over all points, not to mention for all points simultaneously.

We next discuss two extensions of univariate density estimation: derivatives of densities and multivariate densities. In both cases, the technical development is similar to the above except for being more tedious. We will therefore be cursory about many details. The message is not that these extensions are unimportant – on the contrary! But I hope that the preceding development gave you a good idea of how results were derived, and actually doing the derivations becomes very tedious.

An important insight will be that, while results superficially resemble the above, optimal bandwidth rates, and therefore also rates of convergence of optimal estimators, depend on the estimand. This phenomenon is typical for nonparametric estimation and inference. In the specific case of higher-dimensional density estimation, it gives rise to the **curse of dimensionality**: Rates of consistency deteriorate as the dimension of parameter space increases.

10.1.8 Multivariate Densities

Consider estimation of the density f of the random vector $\boldsymbol{x}_i \in \boldsymbol{R}^D$. Then a kernel is a function $K : \boldsymbol{R}^D \mapsto \boldsymbol{R}$. Conceptually, the only restriction on a kernel is that $\int K(\boldsymbol{u})d\boldsymbol{u} = 1$, but we will restrict attention to kernels of the product form

$$K(\boldsymbol{u}) = k(u_1)k(u_2)\dots k(u_D),$$

where $\boldsymbol{u} = (u_1, \dots, u_D)$ and where k is symmetric. Thus, the kernel function is a product of D identical, component-wise kernels. If k is nonnegative, this can be thought of as the density corresponding to $(u_1, \dots, u_D)$ being distributed independently with density k. Rather than one bandwidth, we now have a bandwidth matrix $\boldsymbol{H} = \mathrm{diag}(h_1, \dots, h_D)$, where we define $|\boldsymbol{H}| = h_1 \times \dots \times h_D$. The density estimator is

$$\hat{f}(x) = \frac{1}{nh_1 \dots h_D} \sum_{i=1}^{n} \prod_{d=1}^{D} k\left(\frac{x_{d,i} - x_d}{h_d}\right) = \frac{1}{n\,|\boldsymbol{H}|} \sum_{i=1}^{n} K\left(\boldsymbol{H}^{-1}(\boldsymbol{x}_i - \boldsymbol{x})\right).$$

One can then derive

$$\begin{aligned}
\mathrm{bias}\big(\hat{f}(\boldsymbol{x})\big) &= \frac{\kappa_\nu}{\nu!} \sum_{d=1}^{D} \frac{\partial^\nu}{\partial x_d^\nu} f(\boldsymbol{x}) h_d^\nu + o(h_1^\nu + ... + h_D^\nu) \\
\mathrm{var}\big(\hat{f}(\boldsymbol{x})\big) &= \frac{f(\boldsymbol{x})R(k)^D}{n\,|\boldsymbol{H}|} + O(1/n) \\
\mathrm{AMISE}\big(\hat{f}\big) &= \left(\frac{\kappa_\nu}{\nu!}\right)^2 \int \left(\sum_{d=1}^{D} \frac{\partial^\nu}{\partial x_d^\nu} f(\boldsymbol{x}) h_d^\nu\right)^2 d\boldsymbol{x} + \frac{R(k)^D}{n\,|\boldsymbol{H}|}.
\end{aligned}$$

The last expression cannot be minimized in closed form. However, we observe that the univariate kernel k – which is still "identified" only up to scale – only enters through $R(k)$, so the same kernel remains optimal. Also, we can easily ascertain the rate of the optimal bandwidth: On the assumption that all components of the bandwidth are of the same order, we can set $h^{2\nu} = cn^{-1}h^{-D}$ (here, c is some constant) and solve for $h \propto n^{-1/(2\nu+D)}$, implying that AMISE* is of order $O(n^{-2\nu/(2\nu+D)})$, specializing to $O(n^{-4/(4+D)})$ if the kernel is nonnegative. This is the **curse of dimensionality**.

In order to optimize bandwidth in practice, we usually do two things:

- The D componentwise bandwidths are constrained to be of the same order, scaled only to adjust to data variability. Thus, $h_d = h\hat{\sigma}_d$, where h is common across components and $\hat{\sigma}_d$ estimates the standard deviation of $x_{d,i}$. This also means the ratios of component bandwidths are scale invariant.
- We then resort to rules of thumb or to cross-validation to determine h.

The algebra comes out somewhat differently from above, but the basic ideas are the same, and for the sake of brevity we'll leave it at that.

10.1.9 Estimation of Derivatives

Derivatives often have interpretation, e.g. (after rescaling) as elasticities in demand estimation. So there is ample motivation for estimating them. To do that, we return to the scalar case but the estimand now is the r'th derivative $f^{(r)}$ of f.

The obvious estimator is the plug-in estimator

$$\widehat{f^{(r)}} = \hat{f}^{(r)} = \frac{1}{nh^{1+r}} \sum_{i=1}^{n} k^{(r)} \left(\frac{x_i - x}{h} \right).$$

At a minimum, that requires $k^{(r)}$ to exist, constraining our choice of kernel. Furthermore, we will see that optimal bandwidth results are yet again different from the previous chapter's reference results. Let's compute the bias:

$$\begin{aligned}
\mathbb{E}\big(\widehat{f^{(r)}}(x)\big) &= \int_{-\infty}^{\infty} \frac{1}{h^{1+r}} k^{(r)} \left(\frac{t-x}{h} \right) f(t)dt \\
&= \int_{-\infty}^{\infty} \frac{1}{h^{r}} k^{(r-1)} \left(\frac{t-x}{h} \right) f'(t)dt \\
&= \ldots = \int_{-\infty}^{\infty} \frac{1}{h} k \left(\frac{t-x}{h} \right) f^{(r)}(t)dt,
\end{aligned}$$

where we integrated by parts r times. In analogy to previous arguments, we get

$$\int_{-\infty}^{\infty} \frac{1}{h} k \left(\frac{t-x}{h} \right) f^{(r)}(t)dt = \int_{-\infty}^{\infty} k(u) f^{(r)}(x+hu)dt = f^{(r)}(x) + \frac{1}{\nu!} f^{(r+\nu)}(x) h^{\nu} \kappa_{\nu} + o(h^{\nu}),$$

so the leading term of the bias equals $f^{(r+\nu)}(x)h^\nu\kappa_\nu/\nu!$. Note that this argument requires existence of derivatives of f up to $r+\nu+1$, so higher order kernels call for even stronger assumptions here. Next,

$$\begin{aligned}
\operatorname{var}\left(\widehat{f^{(r)}}(x)\right) &= \frac{1}{nh^{2+2r}}\operatorname{var}\left(k^{(r)}\left(\frac{x_i-x}{h}\right)\right) \\
&= \frac{1}{nh^{2+2r}}\left[\mathbb{E}\left(k^{(r)}\left(\frac{x_i-x}{h}\right)\right)^2 - \left(\mathbb{E}\left(k^{(r)}\left(\frac{x_i-x}{h}\right)\right)\right)^2\right] \\
&= \frac{1}{nh^{2+2r}}\left[h\int_{-\infty}^{\infty} k^{(r)}(u)^2 f(x+hu)du - \left(h\int_{-\infty}^{\infty} k^{(r)}(u) f(x+hu)du\right)^2\right] \\
&= \frac{f(x)}{nh^{1+2r}}\int_{-\infty}^{\infty} k^{(r)}(u)^2 du - \frac{1}{nh^{2r}}\left(\int_{-\infty}^{\infty} k^{(r)}(u)\, du\right)^2 \\
&= \frac{f(x)R(k^{(r)})}{nh^{2+2r}} + O\left(\frac{1}{nh^{2r}}\right),
\end{aligned}$$

Next, we can solve to get

$$\begin{aligned}
\operatorname{AMSE}\left(\widehat{f^{(r)}}(x)\right) &= \frac{f^{(r+\nu)}(x)^2 h^{2\nu}\kappa_\nu^2}{(\nu!)^2} + \frac{f(x)R(k^{(r)})}{nh^{1+2r}} \\
\operatorname{AMISE}\left(\widehat{f^{(r)}}(x)\right) &= \frac{R(f^{(r+\nu)})h^{2\nu}\kappa_\nu^2}{(\nu!)^2} + \frac{R(k^{(r)})}{nh^{1+2r}},
\end{aligned}$$

thus the variance (the r.h. term) is much larger than before! This changes the bias-variance trade-off and we'll have to use a larger bandwidth. In particular, we find

$$\begin{aligned}
h_r^* &= \left(\frac{(1+2r)(\nu!)^2 R(k^{(r)})}{2\nu\kappa_\nu^2 R(f^{(r+\nu)})}\right)^{1/(1+2r+2\nu)} n^{-1/(1+2r+2\nu)} \\
\operatorname{AMISE}^* &= (1+2r+2\nu)\left(\frac{\kappa_\nu^2}{(1+2r)(\nu!)^2}\right)^{\frac{1+2r}{1+2r+2\nu}}\left(\frac{R(k^{(r)})}{2\nu}\right)^{\frac{2\nu}{1+2r+2\nu}} n^{\frac{-2\nu}{1+2r+2\nu}},
\end{aligned}$$

which for nonnegative kernels specializes to

$$\begin{aligned}
h_r^* &= \left(\frac{(1+2r)R(k'')}{\kappa_2^2 R(f^{(r+2)})}\right)^{1/(5+2r)} n^{-1/(5+2r)} \\
\operatorname{AMISE}^* &= (5+2r)\left(\frac{\kappa_2^2}{4+8r}\right)^{\frac{1+2r}{5+2r}}\left(\frac{R(k^{(r)})}{4}\right)^{\frac{4}{5+2r}} n^{\frac{-4}{5+2r}}.
\end{aligned}$$

As one would have hoped, the rates recover previous rate results for $r=0$, but they also illustrate considerable penalty for trying to estimate derivatives.

The asymptotically optimal nonnegative kernels now are the biweight kernel

$$k^{bi}(u) = \frac{15}{16}\left(\left|1-u^2\right|_+\right)^2$$

for estimating f' and the triweight kernel

$$k^{tri}(u) = \frac{35}{32}\left(\left|1-u^2\right|_+\right)^3$$

for estimating f''. Both can be generalized to higher orders and continue to be optimal for these estimands. Note that the Gaussian kernel turns out to be quite inefficient for these tasks, and the uniform, Epanechnikov, as well as any other nondifferentiable kernels are inapplicable.

10.2 Nonparametric Mean Regression

We next consider estimation of

$$m(\boldsymbol{x}) \equiv \mathbb{E}(y_i|\boldsymbol{x}_i = \boldsymbol{x}).$$

This is called *nonparametric mean regression*. It generalizes OLS and many other models, as becomes espcially obvious upon equivalently writing

$$y_i = m(\boldsymbol{x}_i) + \varepsilon_i, \quad \mathbb{E}(\varepsilon_i|\boldsymbol{x}_i) = 0.$$

Note that in contrast to the linear model,

$$y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + \varepsilon_i, \quad \mathbb{E}(\varepsilon_i|\boldsymbol{x}_i) = 0,$$

the stipulation $\mathbb{E}(\varepsilon_i|\boldsymbol{x}_i) = 0$ is a normalization, not a substantive assumption. We will also not assume homoskedasticity and generally treat $\sigma^2(\boldsymbol{x}) \equiv \mathbb{E}(\varepsilon_i^2|\boldsymbol{x}_i = \boldsymbol{x})$ as nonconstant. Indeed, the only substantive assumption we make is that data are i.i.d., that $m(\boldsymbol{x})$ is well-defined, and (frequently suppressed) smoothness conditions on f and σ^2. We will explicitly analyze mostly the case of scalar x_i and keep notation consistent by using f for the density of x_i. The extension to multidimensional $\boldsymbol{x}_i$ is theoretically (not computationally) routine.

10.2.1 Nadaraya-Watson Estimator

An intuitively obvious estimator that naturally relates to kernel density estimation is the **local constant** or **Nadaraya-Watson** estimator:

$$\hat{m}(x) = \frac{\sum_{i=1}^n y_i k\left(\frac{x_i - x}{h}\right)}{\sum_{i=1}^n k\left(\frac{x_i - x}{h}\right)}.$$

Very similarly to kernel density estimation, this can be intuited as smoothly generalizing the "moving window average"

$$\hat{m}^{uni}(x) = \frac{\sum_{i=1}^n y_i \mathbf{1}(|x_i - x| \leq h)}{\sum_{i=1}^n \mathbf{1}(|x_i - x| \leq h)}$$

which in turn specializes $\hat{m}$ to the case of uniform kernels. Note also that the estimator can be written as

$$\hat{m}(x) = \frac{\frac{1}{nh}\sum_{i=1}^n y_i k\left(\frac{x_i - x}{h}\right)}{\hat{f}(x)},$$

where it is understood that $\hat{f}$ uses the same kernel and bandwidth. We will restrict attention to nonnegative kernels to avoid dealing with $\hat{f}(x) \leq 0$.

Taking a cue from the last representation of $\hat{m}$ and using $y_i = m(x_i) + \varepsilon_i$, write

$$\begin{aligned}
\frac{\frac{1}{nh}\sum_{i=1}^n y_i k\left(\frac{x_i - x}{h}\right)}{\hat{f}(x)} &= \frac{1}{nh\hat{f}(x)}\sum_{i=1}^n k\left(\frac{x_i - x}{h}\right)(m(x) + m(x_i) - m(x) + \varepsilon_i) \\
&= \frac{1}{nh\hat{f}(x)}\sum_{i=1}^n k\left(\frac{x_i - x}{h}\right)m(x) + \frac{1}{nh\hat{f}(x)}\sum_{i=1}^n k\left(\frac{x_i - x}{h}\right)(m(x_i) - m(x)) \\
&\quad + \frac{1}{nh\hat{f}(x)}\sum_{i=1}^n k\left(\frac{x_i - x}{h}\right)\varepsilon_i \\
&= m(x) + \frac{\hat{m}_1(x)}{\hat{f}(x)} + \frac{\hat{m}_2(x)}{\hat{f}(x)},
\end{aligned}$$

where the last step defines m_1 and m_2 and observes simplification in the first term.

A simple application of the Law of Iterated Expectations yields $\mathbb{E}\hat{m}_2(x) = 0$, hence the bias of the estimator equals $\mathbb{E}(\hat{m}_1(x)/\hat{f}(x))$. In contrast, and less obviously, the contribution of $\hat{m}_2(x)$ to the variance dominates, and so in practice this is the "variance term." Write

$$\begin{aligned}
\text{var}\left(\hat{m}_2(x)\right) &= \frac{1}{nh^2}\text{var}\left(k\left(\frac{x_i - x}{h}\right)\varepsilon_i\right) = \frac{1}{nh^2}\mathbb{E}\left(k\left(\frac{x_i - x}{h}\right)\varepsilon_i\right)^2 \\
&= \frac{1}{nh^2}\mathbb{E}\left(k\left(\frac{x_i - x}{h}\right)^2\sigma^2(x_i)\right) = \frac{1}{nh^2}\int k\left(\frac{t - x}{h}\right)^2\sigma^2(t)f(t)dt \\
&= \frac{1}{nh}\int k(u)^2\sigma^2(x + hu)f(x + hu)du = \frac{1}{nh}\int k(u)^2\sigma^2(x)f(x)du + o(1/nh) \\
&= \frac{R(k)\sigma^2(x)f(x)}{nh} + o(1/nh),
\end{aligned}$$

where the last step used smoothness assumptions on σ^2 and f.

Next,

$$\begin{aligned}
\mathbb{E}\hat{m}_1(x) &= \frac{1}{h}\mathbb{E}\left(k\left(\frac{x_i - x}{h}\right)(m(x_i) - m(x))\right) = \frac{1}{h}\int k\left(\frac{t - x}{h}\right)(m(x_i) - m(x))f(t)dt \\
&= \int k(u)(m(x + hu) - m(x))f(x + hu)du \\
&= \int\left(k(u)\left(hum'(x) + \frac{h^2u^2}{2}m''(x)\right)(f(x) + huf'(x))\right)du + o(h^2) \\
&= \underbrace{\int k(u)udu}_{=0}\cdot hm'(x)f(x) + \int k(u)u^2du \times h^2\left(\frac{1}{2}m''(x)f(x) + m'(x)f'(x)\right) + o(h^2) \\
&= \kappa^2h^2f(x)\cdot\underbrace{\left(\frac{1}{2}m''(x) + m'(x)f'(x)/f(x)\right)}_{\equiv B(x)} + o(h^2).
\end{aligned}$$

(If you replicate this, note that the $o(h^2)$ absorbed a $O(h^3)$ cross-product term of the integral.) By similar algebra, it can be verified that $\text{var}(\hat{m}_1(x)) = o(1/nh)$, so that we can use Chebyshev's inequality

to get

$$\begin{aligned} & \sqrt{nh}\left(\hat{m}_1(x) - h^2\kappa_2 B(x) f(x)\right) \xrightarrow{p} 0 \\ \Leftrightarrow \quad & \sqrt{nh}\big(\hat{m}_1(x)/\hat{f}(x) - h^2\kappa_2 B(x)\big) \xrightarrow{p} 0, \end{aligned}$$

where we used $f(x)/\hat{f}(x) \xrightarrow{p} 1$ (if you feel uneasy about that, see below). In sum, and also invoking a CLT,

$$\sqrt{nh}\left(\hat{m}(x) - m(x) - h^2\kappa^2 B(x)\right) \xrightarrow{d} N\left(0, \frac{R(k)\sigma^2(x)}{f(x)}\right).$$

Next, we compute

$$\text{AMSE}(\hat{m}(x)) = h^4\kappa_2^2 B(x)^2 + \frac{R(k)\sigma^2(x)}{nhf(x)}.$$

It would be intuitive to just take the expectation w.r.t. x_i to get an AMISE, but that integral will not in general exist. The culprit is the $\sigma^2(x)/f(x)$ term. We therefore define a weighted integrated MSE that introduces a weighting function w:

$$\text{WIMSE}(\hat{m}) \equiv h^4\kappa_2^2 \int B(x)^2 w(x)dx + \frac{R(k)}{nh}\int \frac{\sigma^2(x)}{f(x)} w(x)dx.$$

The weighting function must discount low values of f. For example, we could use $w(x) = \mathbf{1}\{f(x) \geq \delta\}$ for some tuning parameter δ. Beyond the fact that $\sigma^2(x)/f(x)$ might not be integrable, this also helps with a leap of faith in the above development: We relied on $f(x)/\hat{f}(x) \xrightarrow{p} 1$, but that is not true uniformly over small f.

We will not crank out optimized losses, but note that, by setting $h^4 = 1/nh$, we can easily solve for $h^* \propto n^{-1/5}$, and plugging in we then find that $\text{WIMSE}^* \propto n^{-4/5}$, hence the optimized rate of convergence is $n^{-2/5}$. These rates should look familiar – they are the same as for density estimation. The generalization to multivariate mean regression has this same feature, including the curse of dimensionality. Furthermore, optimal kernel theory is unchanged: Normalizing $\kappa^2 = 1$, we see that WIMSE depends on our choice of kernel through $R(k)$, hence the optimal kernel is Epanechnikov.

Optimal bandwidth is a much more complex question. While rule-of-thumb bandwidths are frequently used, they were not developed for this application and in particular do not try to pre-estimate B. I rather recommend cross-validation.

The Nadaraya-Watson estimator is an important default, and it actually performs pretty well in some settings, notably in somewhat higher dimension if some covariates do not really matter. However, it has some unappealing features:

- The limit case as $h \to \infty$ is not a linear fit but a horizontal fit corresponding to the constant function $y = \overline{y}$. (Hence, "local constant estimation," turning into "global constant estimation" as $h \to \infty$.)

- The estimator will be badly biased, indeed inconsistent, at the boundaries of the support unless $m(x)$ is constant there. The reason is that at the lower boundary, observed x_i will be overwhelmingly above x, so that we have upward [downward] bias if m is increasing [decreasing], and similarly at the upper boundary.

- The estimator transforms a perfectly linear scatterplot into a nonlinear fit unless the x are perfectly evenly spaced. To see this, imagine a uniform kernel and say that the scatterplot is perfectly linearly increasing, i.e. there exist (α, β) s.t. $y_i = \alpha + \beta x_i$ for all $i = 1, \ldots, n$. Then
$$\hat{m}(x) = \frac{\sum_i (\alpha + \beta x_i) 1\{|x_i - x| \leq h\}}{\sum_i 1\{|x_i - x| \leq h\}} = \alpha + \beta \frac{\sum_i x_i 1\{|x_i - x| \leq h\}}{\sum_i 1\{|x_i - x| \leq h\}}$$
does *not* in general equal $\alpha + \beta x$, nor is it in general linear in x. In fact, it's a step function; that could be fixed by smoothing the kernel, but the nonlinearity cannot.

These considerations led researchers to consider generalizations of the local constant estimator and also completely different approaches. We will discuss the former in some detail and the latter very briefly.

10.2.2 Local Polynomial Regression

The class of local polynomial estimators can be intuitively described as follows: At each x, fit a local (to x) polynomial approximation to $m(x)$. "Local" is operationalized by weighting all data points according to a kernel density centered at x. The intuition is again clearest with a uniform kernel, in which case we simply restrict attention to the data in window $[x-h, x+h]$. Polynomial approximation means that we regress y_i on a constant, x_i, x_i^2, and so on. The estimator $\hat{m}(x)$ equals the fitted value of the regression at x. Operationally, other than weighting the data, we will recenter them at x, i.e. replace each x_i with $(x_i - x)$. The fitted value of the regression at x, and therefore the estimator $\hat{m}(x)$, is then just the estimated constant $\hat{\alpha}$. Conceptually, this procedure is repeated at each x; in practice, we do it on a grid.

Computationally, local polynomial regression can utilize existing routines for regression. The algorithm is as follows:

1. At each x where the regression is to be evaluated, replace each x_i with $\tilde{x}_i = x_i - x$ and assign it weight $w(\tilde{x}_i) = k(\tilde{x}_i/h)$.

2. Run weighted least squares (WLS) of y_i on a constant, $\tilde{x}_i$, $\tilde{x}_i^2$, ... up to desired order of polynomial. Thus, we have the closed-form regression coefficients
$$\begin{aligned} \hat{\boldsymbol{\beta}}_x &= \left(\sum_i k(\tilde{x}_i/h) \boldsymbol{z}_i \boldsymbol{z}_i' \right)^{-1} \sum_i k(\tilde{x}_i/h) \boldsymbol{z}_i y_i, \\ \boldsymbol{z}_i &= (1, \tilde{x}_i, \tilde{x}_i^2, \ldots, \tilde{x}_i^k)'. \end{aligned}$$

3. The estimator $\hat{m}(x)$ is the estimated intercept.

If we locally regress y_i on a constant and $\tilde{x}_i$, we have local linear regression. If we regress y_i on the constant only, we have local constant regression, i.e. we recover the Nadaraya-Watson estimator as a special case.

Similarly to the intuition already given for Nadaraya-Watson, all local polynomial approximations collapse to the corresponding global approximation as $h \to \infty$, e.g. local linear approximation collapses to OLS. While one can imagine going to relatively high polynomials with scalar x_i, the number of coefficients estimated in the local fit otherwise increases rapidly with order of polynomials (because of cross-product terms). Also taking into account that effective sample size can only grow slowly with n in higher dimension, our ability to locally fit higher order polynomials in higher dimensional covariate space is severely limited with the sample sizes that economists typically encounter. In practice, local polynomials of order greater than 1 are almost exclusively used in scalar problems.

10.2.3 Asymptotics for Local Polynomial Regression

The derivations of asymptotic bias, variance, etc. become extremely involved. We will state without proof that for local linear estimation,

$$\sqrt{nh}\left(\hat{m}_{LL}(x) - m(x) - h^2\kappa^2 m''(x)/2\right) \xrightarrow{d} N\left(0, \frac{R(k)\sigma^2(x)}{f(x)}\right).$$

Note that the bias is proportional to $m''(x)$; this is intuitive upon visually comparing local linear and local quadratic approximations.

The result can be directly compared to local constant estimation. Undoing the definition of $B(x)$ above, we get

$$\sqrt{nh}\left(\hat{m}_{LC}(x) - m(x) - h^2\kappa^2\left(\frac{1}{2}m''(x) + m'(x)f'(x)/f(x)\right)\right) \xrightarrow{d} N\left(0, \frac{R(k)\sigma^2(x)}{f(x)}\right).$$

Hence, the asymptotic distributions have the same variance and differ by a translation of $h^2\kappa_2 m'(x)f'(x)/f(x)$. While we cannot in general rank the bias terms, the common sense is that "typically" the simpler bias term that does not involve m' will also be smaller. This intuition guides practice and accounts for the popularity of $\hat{m}_{LL}$, but it is not a theorem. Note that $\hat{m}_{LL}$ tends to smooth the data a bit less, which may translate into higher finite sample variance.

With that all said, note that the order of bias has not changed, and optimal rates are still the same as for kernel density estimation with nonnegative kernels. This may be surprising, and because of higher order polynomials' ability to fit all different kinds of curves, one might expect it to change as we increase the polynomial order, though with obvious downsides in terms of finite sample performance. This is indeed so, and the order of bias goes down as we move on to 2^{nd} order polynomials.

In general, the bias and variance of higher order polynomials iterate the above pattern as follows:

- For local polynomial estimation of odd order k, the bias is $O(h^{k+1})$. The leading term is $m^{(k+1)}(x)$.
- For local polynomial estimation of even order k, the bias is $O(h^{k+2})$. The leading terms are $m^{(k+2)}(x)$ and $m^{(k+1)}(x)f'(x)/f(x)$. The second term is sometimes called "design bias."

Thus, we gain an order of bias only as we increase the order from odd to even (in increments of h^2), though we effectively lose one of the two bias terms as we increase the order from even to odd.

10.2.4 Cross-Validation

The leading way to determine bandwidths for local polynomial regression is cross-validation. The intuition is minimization of a least squares criterion, i.e. to choose h so as to minimize an estimator of $\mathbb{E}\varepsilon_i^2$. We have to be careful about two things. First, it is essential to use the leave-one-out estimator because we otherwise get a degenerate result; do you see why? Second, we disregard behavior of the estimator in regions of very low density of x_i, e.g. the tails of a normal distribution, by introducing a weighting or trimming function as with weighted integrated mean square error. The criterion then is

$$CV_{LS}(h) = \frac{1}{n}\sum_i \hat{\varepsilon}_{-i}^2 w(x_i),$$

where $\hat{\varepsilon}_{-i} = y_i - \hat{m}_{-i}(x_i)$ is the leave-one-out residual and $\hat{m}_{-i}(x_i)$ is the leave-one-out estimator.

This is a good estimator because

$$\mathbb{E}CV_{LS}(h) = \text{WIMSE}(h) + \mathbb{E}\left(\varepsilon_i^2 w(x_i)\right),$$

and the right-hand side can be interpreted as weighted integrated squared prediction error that compounds variability of y_i with parameter estimation error. Furthermore, since $\mathbb{E}\left(\varepsilon_i^2 w(x_i)\right)$ does not depend on h, the bandwidth that minimizes $CV_{LS}(h)$ estimates (by usual m-estimator arguments) the bandwidth that minimizes WIMSE(h).

To see the claim, write

$$\begin{aligned}\mathbb{E}CV_{LS}(h) &= \mathbb{E}\big((\varepsilon_i + m(x_i) - \hat{m}_{-i}(x_i))^2 w(x_i)\big) \\ &= \mathbb{E}\big((m(x_i) - \hat{m}_{-i}(x_i))^2 w(x_i)\big) + \mathbb{E}\left(\varepsilon_i^2 w(x_i)\right) - 2\underbrace{\mathbb{E}\left((m(x_i) - \hat{m}_{-i}(x_i))\,\varepsilon_i w(x_i)\right)}_{=0},\end{aligned}$$

where we used $\mathbb{E}(\varepsilon_i|x_i) = 0$. It then remains to observe that

$$\mathbb{E}\big((m(x_i) - \hat{m}_{-i}(x_i))^2 w(x_i)|x_{-i}\big) = \int_{-\infty}^{\infty} (m(t) - \hat{m}_{-i}(t))^2 w(t) f(t) dt$$

for any leave-one-out sample x_{-i}, hence

$$\mathbb{E}\big((m(x_i) - \hat{m}_{-i}(x_i))^2 w(x_i)\big) = \int_{-\infty}^{\infty} \mathbb{E}\,(m(t) - \hat{m}_{-i}(t))^2 w(t) f(t) dt$$
$$= \int_{-\infty}^{\infty} \mathbb{E}\,(m(t) - \hat{m}(t))^2 w(t) f(t) dt = \text{WIMSE}(h).$$

10.2.5 Semiparametric Mean Regression

Semiparametric methods combine parametric and nonparametric aspects. The aim is to get the best of both worlds: nonparametric flexibility where it matters but also avoiding the curse of dimensionality. Some authors further differentiate between semiparametrics and seminonparametrics depending on whether the quantities of interest or the nuisance parameters are nonparametrically estimated.

Classic examples of semiparametric models are as follows.

Example 10.1 (Partially Linear Model)

$$\mathbb{E}(y_i|\boldsymbol{x}_i, \boldsymbol{z}_i) = \boldsymbol{x}_i'\boldsymbol{\beta} + m(\boldsymbol{z}_i)$$

Example 10.2 (Separable Model)

$$\mathbb{E}(y_i|x_{1,i}, \ldots, x_{K,i}) = \sum_{k=1}^{K} m_k(x_{k,i})$$

Example 10.3 (Linear Index Model)

$$\mathbb{E}(y_i|\boldsymbol{x}_i) = m(\boldsymbol{x}_i'\boldsymbol{\beta})$$

Example 10.4 (Generated Regressors)

$$\mathbb{E}(y_i|\boldsymbol{x}_i) = \boldsymbol{m}(\boldsymbol{x}_i)'\boldsymbol{\beta}$$

The examples differ with respect to what's typically of interest. In the partially linear model, the modal application is interested in m, and the $\mathbf{x}_i$ are controls. They are handled linearly in the hope to get "good enough" avoidance of omitted variable bias without encountering a curse of dimensionality.[23] In contrast, in the last example, it is frequently $\boldsymbol{\beta}$, or at least some components thereof, that is of interest.

In the first three examples, we can get a parametric convergence rate, as well as convergence of estimators of nonparametric objects at the rate corresponding to the argument's dimensionality, under reasonable though not innocuous conditions. We will examplarily discuss this using the partially linear model and generalizing to extremum (or m-) estimation with infinite dimensional nuisance parameters. The last case is less benign.

[23] A notable caveat is the linear model with few parameters of interest but many (relative to sample size) controls. If Machine Learning tools are used to estimate the controls, this becomes akin to a partially linear model with $\boldsymbol{\beta}$ the parameter of interest.

10.2.6 Robinson (1988) on Partially Linear Models

Write the Partially Linear Model as

$$\begin{aligned} y_i &= \boldsymbol{x}_i'\boldsymbol{\beta} + m(\boldsymbol{z}_i) + \epsilon_i \\ \mathbb{E}(\epsilon_i|\boldsymbol{x}_i, \boldsymbol{z}_i) &= 0, \end{aligned}$$

noting that we do allow for heteroskedasticity, i.e. $\sigma^2(\boldsymbol{x}, \boldsymbol{z}) = \mathbb{E}(\epsilon_i^2|\boldsymbol{x}_i, \boldsymbol{z}_i = (\boldsymbol{x}, \boldsymbol{x}))$ is not restricted to be constant. The classic analysis of this model is due to Robinson (*Econometrica*, 1988). Identification is not completely trivial but we will take it for granted.

The basic ideas of estimation is to "concentrate out" m. Write

$$\begin{aligned} m_y(\boldsymbol{z}_i) =: \mathbb{E}(y_i|\boldsymbol{z}_i) &= \mathbb{E}(\boldsymbol{x}_i'\boldsymbol{\beta}|\boldsymbol{z}_i) + \mathbb{E}(m(\boldsymbol{z}_i)|\boldsymbol{z}_i) + \mathbb{E}(\epsilon_i|\boldsymbol{z}_i) \\ &= \mathbb{E}(\boldsymbol{x}_i'\boldsymbol{\beta}|\boldsymbol{z}_i) + m(\boldsymbol{z}_i) \\ &=: \boldsymbol{m}_x(\boldsymbol{z}_i)'\boldsymbol{\beta} + m(\boldsymbol{z}_i) \end{aligned}$$

We therefore have

$$y_i - m_y(\boldsymbol{z}_i) = (\boldsymbol{x}_i - \boldsymbol{m}_x(\boldsymbol{z}_i))'\boldsymbol{\beta} + \epsilon_i, \tag{18}$$

and if we knew $(m_y, \boldsymbol{m}_x)$, we could estimate $\boldsymbol{\beta}$ by OLS regression of $y_i - m_y(\boldsymbol{z}_i)$ on $(\boldsymbol{x}_i - \boldsymbol{m}_x(\boldsymbol{z}_i))$. This estimator is infeasible, but it motivates a two-stage plug-in procedure: First use the nonparametric method of choice to compute estimators $(\hat{m}_y, \hat{\boldsymbol{m}}_x)$, then compute

$$\hat{\boldsymbol{\beta}} = \left(\sum_{i=1}^n (\boldsymbol{x}_i - \hat{\boldsymbol{m}}_x(\boldsymbol{z}_i))(\boldsymbol{x}_i - \hat{\boldsymbol{m}}_x(\boldsymbol{z}_i))'\right)^{-1} \sum_{i=1}^n (\boldsymbol{x}_i - \hat{\boldsymbol{m}}_x(\boldsymbol{z}_i))(y_i - \hat{m}_y(\boldsymbol{z}_i)).$$

An immediate problem is that this estimator need not even be consistent because $(\hat{m}_y, \hat{\boldsymbol{m}}_x)$ need not be consistent uniformly over the support of $\boldsymbol{z}_i$, specifically if the density vanishes at some points. (And in practice, a close-enough-to-vanishing density would lead to bad finite sample behavior.) Robinson therefore suggests to trim the data and only use data points for which $\boldsymbol{z}_i$ takes a value with high enough estimated density. This leads to estimator

$$\begin{aligned} \hat{\boldsymbol{\beta}} = \left(\sum_{i=1}^n (x_i - \hat{\boldsymbol{m}}_x(\boldsymbol{z}_i))(\boldsymbol{x}_i - \hat{\boldsymbol{m}}_x(\boldsymbol{z}_i))' \times \mathbf{1}\{\hat{f}_z(z_i) \geq b_n\}\right)^{-1} \\ \sum_{i=1}^n (x_i - \hat{\boldsymbol{m}}_x(\boldsymbol{z}_i))(y_i - \hat{m}_y(\boldsymbol{z}_i)) \times \mathbf{1}\{\hat{f}_z(z_i) \geq b_n\}, \end{aligned}$$

where b_n is a tuning parameter that is presumed to vanish very slowly. We will follow much of the literature and ignore b_n, though strictly speaking it should show up in rate results below.

Ideally, we want to claim that this estimator is "oracle efficient," that is, it asymptotically behaves like the infeasible one. (Of course, even if that holds, we should not assume that finite sample behavior is equally as good.) Thus, we would ideally get

$$\begin{aligned}\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) &\xrightarrow{d} N(\mathbf{0}, \boldsymbol{G}^{-1}\boldsymbol{S}\boldsymbol{G}^{-1}),\\ \boldsymbol{G} &= \mathbb{E}\left[(\boldsymbol{x}_i - \boldsymbol{m}_x(\boldsymbol{z}_i))(\boldsymbol{x}_i - \boldsymbol{m}_x(\boldsymbol{z}_i))'\right],\\ \boldsymbol{S} &= \mathbb{E}\left[(\boldsymbol{x}_i - \boldsymbol{m}_x(\boldsymbol{z}_i))(\boldsymbol{x}_i - \boldsymbol{m}_x(\boldsymbol{z}_i))'\sigma^2(\boldsymbol{x}_i, \boldsymbol{z}_i)\right].\end{aligned}$$

This is indeed true under conditions. These obviously include what we would need to get OLS to work. The crucial addition is that the MSE of our nonparametric estimator has to vanish at a sufficient rate. Assuming nonnegative kernels, that condition is $\sqrt{n}(h^4 + n^{-1}h^{-d}) \to 0$. Assuming optimal bandwidth choice, this obtains if $\sqrt{n} \times n^{-4/(4+d)} \to 0$, i.e. if $d \leq 3$ because d is an integer. Note that the integer problem forces us to create some "slack" here that can be used to handle b_n. In short, this method has the desired asymptotics if the nonparametric component has up to three dimensions.[24]

We finally recall that in most applications, the quantity of true interest is m, and the covariates that enter parametrically are "controls." The idea is that linearly adjusting for their effect may be good enough. If this is the motivation, there will be a third estimation stage in which we nonparametrically regress on $\mathbf{z}_i$ after concentrating out $\mathbf{x}_i'$, i.e. nonparametric mean regression in the model

$$y_i - \boldsymbol{x}_i'\hat{\boldsymbol{\beta}} =: \tilde{y}_i = m(\boldsymbol{z}_i) + \epsilon_i.$$

Here, we can ignore estimation error in $\boldsymbol{\beta}$ because the estimator converges at a faster than the nonparametric rate and therefore is "superconsistent" relative to the relevant rate of localization. Note that this final regression need not mimic the preliminary estimation of m_y in detail. Indeed, it frequently uses other values of the tuning parameters, notably if inference is desired or if eyeball criteria suggest specific bandwidths.

10.2.7 When can we estimate parametric components at parametric rate?

We next think more generally about when we can estimate parametric components of semiparametric models at a parametric rate. Indeed, in favorable cases, the asymptotics for the parametric components look like standard m-estimator asymptotics. However, "favorable" here refers to conditions that apply to some but not all interesting models and so truly must be checked on a case-by-case basis. The development follows Andrews' (*Econometrica*, 1994) work on "**MIN**imum estimation with **P**reliminary **I**nfinite Dimensional **N**uisance," but similar ideas appeared in numerous places at the time.

[24] In principle, higher order kernels can be used to make these asymptotics "work" for arbitrarily high d. Note the scare quotes.

The idea is to develop a general theory of extremum estimation when the objective function contains an infinite dimensional nuisance parameter that must be pre-estimated. Thus, suppose that

$$\boldsymbol{\theta}_0 = \arg\min Q(\boldsymbol{\theta}, \tau_0)$$

is being estimated by

$$\hat{\boldsymbol{\theta}} = \arg\min Q_n(\boldsymbol{\theta}, \hat{\tau}),$$

where $\hat{\tau}$ is consistent for τ_0. (This is not sufficient but obviously necessary.) The question is if under reasonable conditions, we can ignore the difference between $\hat{\tau}$ and τ_0 and have standard extremum estimator asymptotics apply. Again, the answer will be a qualified "yes," i.e. such conditions are neither innocuous nor absurd and must be verified on a case-by-case basis.

To simplify expressions, we will assume m-estimation in the narrow sense, i.e. that Q_n is a sample average that approximates a corresponding expectation. This simply avoids some remainder terms in the below. It will be convenient to write

$$\frac{\partial Q_n(\boldsymbol{\theta}, \tau)}{\partial \boldsymbol{\theta}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{g}_i(\boldsymbol{\theta}, \tau),$$

i.e. use the shorthand $\boldsymbol{g}_i$ for the derivative of the objective function evaluated on one data point. (This is a r.v. due to a hidden argument $\boldsymbol{w}_i$, the data.) Of course, we assume $\mathbb{E}\boldsymbol{g}_i(\boldsymbol{\theta}_0, \tau_0) = \mathbf{0}$ and also all the conditions needed for m-estimation. Our estimator is then (with probability approaching 1) characterized by

$$\mathbf{0} = \overline{\boldsymbol{g}}_n(\hat{\boldsymbol{\theta}}, \hat{\tau}) := \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{g}_i(\hat{\boldsymbol{\theta}}, \hat{\tau}),$$

and we can replicate the standard algebra for m-estimators:

$$\begin{aligned}
\mathbf{0} &= \sqrt{n}\overline{\boldsymbol{g}}_n(\hat{\boldsymbol{\theta}}, \hat{\tau}) \\
&= \sqrt{n}\overline{\boldsymbol{g}}_n(\boldsymbol{\theta}_0, \hat{\tau}) + \frac{\partial}{\partial \boldsymbol{\theta}'}\overline{\boldsymbol{g}}_n(\boldsymbol{\theta}_0, \hat{\tau})\sqrt{n}\big(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\big) + o_P(1) \\
\Rightarrow \sqrt{n}\big(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\big) &= -\left(\frac{\partial}{\partial \boldsymbol{\theta}'}\overline{\boldsymbol{g}}_n(\boldsymbol{\theta}_0, \hat{\tau})\right)^{-1} \sqrt{n}\overline{\boldsymbol{g}}_n(\boldsymbol{\theta}_0, \hat{\tau}) + o_P(1)
\end{aligned}$$

As a reminder, in an "oracle" situation where τ_0 is known and used, the last line would read

$$\sqrt{n}\big(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\big) = -\left(\frac{\partial}{\partial \boldsymbol{\theta}'}\overline{\boldsymbol{g}}_n(\boldsymbol{\theta}_0, \tau_0)\right)^{-1} \sqrt{n}\overline{\boldsymbol{g}}_n(\boldsymbol{\theta}_0, \tau_0) + o_P(1)$$

and we would recover standard m-estimator asymptotics. Since we need continuous differentiability of $\mathbb{E}\boldsymbol{g}_i$, consistency of $\hat{\tau}$, and existence of the inverse in the expression anyway, swapping in $\hat{\tau}$ for τ_0 in said inverse is not a big deal. However, we next have $\sqrt{n}\overline{\boldsymbol{g}}_n(\boldsymbol{\theta}_0, \hat{\tau})$ where we would like $\sqrt{n}\overline{\boldsymbol{g}}_n(\boldsymbol{\theta}_0, \tau_0)$. This gap is much more difficult to handle and indeed cannot be closed without further assumptions.

The first of these assumptions is a statistical regularity condition called *stochastic equicontinuity* that is typically just assumed. To get an idea, define

$$\boldsymbol{\nu}_n(\tau) = \sqrt{n}\big(\overline{\boldsymbol{g}}_n(\boldsymbol{\theta}_0, \tau) - \mathbb{E}\boldsymbol{g}_i(\boldsymbol{\theta}_0, \tau)\big).$$

This is a stochastic process, i.e. a random variable that takes entire functions as realizations. Note that for any fixed value of τ, $\boldsymbol{\nu}_n$ is (under reasonable conditions) subject to a CLT. Also, it is intuitive that for τ and τ' very close to each other, $\boldsymbol{\nu}_n$ will be similarly distributed and maybe even very highly correlated. These last intuitions are formalized by assuming that

$$\hat{\tau} \xrightarrow{p} \tau_0 \Rightarrow \boldsymbol{\nu}_n(\hat{\tau}) - \boldsymbol{\nu}_n(\tau_0) \xrightarrow{p} \mathbf{0}.$$

Stochastic equicontiuity is actually a more technical condition implying this, but this will do for our ourposes. Verification of this assumption can be very tedious, yet at the same time it is not generally considered especially restrictive and can be verified in many applications. In practice, researchers who are not statisticians will likely cite such a verification or just make the assumption.

Imposing this assumption is very helpful because we assume that $\mathbb{E}\boldsymbol{g}_i(\boldsymbol{\theta}_0, \tau_0) = \mathbf{0}$, so that (again under weak conditions) a CLT will give us

$$\boldsymbol{\nu}_n(\tau_0) = \sqrt{n}\big(\overline{\boldsymbol{g}}_n(\boldsymbol{\theta}_0, \tau_0) - \mathbb{E}\boldsymbol{g}_i(\boldsymbol{\theta}_0, \tau_0)\big) \xrightarrow{d} N\left(0, \mathbb{E}(\boldsymbol{g}_i(\boldsymbol{\theta}_0, \tau_0)\boldsymbol{g}_i(\boldsymbol{\theta}_0, \tau_0)')\right),$$

and this conclusion then extends to $\boldsymbol{\nu}_n(\hat{\tau})$. Plugging back into the m-estimator algebra, we have

$$\begin{aligned}
& \sqrt{n}\big(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\big) \\
= & -\left(\frac{\partial}{\partial \boldsymbol{\theta}'}\overline{\boldsymbol{g}}_n(\boldsymbol{\theta}_0, \hat{\tau})\right)^{-1} \sqrt{n}\big(\overline{\boldsymbol{g}}_n(\boldsymbol{\theta}_0, \hat{\tau}) \underbrace{- \mathbb{E}\boldsymbol{g}_i(\boldsymbol{\theta}_0, \hat{\tau}) + \mathbb{E}\boldsymbol{g}_i(\boldsymbol{\theta}_0, \hat{\tau})}_{\text{add-and-subtract}}\big) + o_P(1) \\
= & -\left(\frac{\partial}{\partial \boldsymbol{\theta}'}\overline{\boldsymbol{g}}_n(\boldsymbol{\theta}_0, \hat{\tau})\right)^{-1} \big(\boldsymbol{\nu}_n(\tau_0) + \sqrt{n}\mathbb{E}\boldsymbol{g}_i(\boldsymbol{\theta}_0, \hat{\tau})\big) + o_P(1),
\end{aligned}$$

and we are good to go... if $\sqrt{n}\mathbb{E}\boldsymbol{g}_i(\boldsymbol{\theta}_0, \hat{\tau}) = o_P(1)$.

This is the second important condition, which is notably overlooked in Li and Racine's textbook.[25] It is not at all innocuous: While $\mathbb{E}\boldsymbol{g}_i(\boldsymbol{\theta}_0, \tau_0) = \mathbf{0}$ will imply $\mathbb{E}\boldsymbol{g}_i(\boldsymbol{\theta}_0, \hat{\tau}) = o_P(1)$ under conditions we need anyway, the additional $\sqrt{n}$-factor makes the condition much more demanding. Indeed, this can be intuited as imposing a local orthogonality between the nonparametric and parametric part of the estimation problem. (If τ were a vector, it would require cross-derivatives at $(\boldsymbol{\theta}_0, \boldsymbol{\tau}_0)$ to be zero.) The condition may or may not hold in a given application and must be explicitly verified.

[25] More precisely, they use a nonstandard definition of $\boldsymbol{\nu}_n$ in which true expectations are not subtracted. This hides the condition in the equicontinuity assumption on $\boldsymbol{\nu}_n$, but means that such equicontinuity is not implied by relevant "off-the-shelf" results. In contrast, Andrews (display 4.4) uses our definition of $\boldsymbol{\nu}_n$ and explicitly imposes the condition discussed here as Assumption N(c); observe simplification as per display 4.9.

Assuming it does hold, and under sufficient "m-estimator" regularity conditions so that we can claim $\left(\frac{\partial}{\partial \boldsymbol{\theta}'}\overline{\boldsymbol{g}}_n(\boldsymbol{\theta}_0,\hat{\tau})\right)^{-1} \xrightarrow{p} \left(\mathbb{E}\frac{\partial}{\partial \boldsymbol{\theta}'}\boldsymbol{g}_i(\boldsymbol{\theta}_0,\tau_0)\right)^{-1}$, we finally conclude that

$$\sqrt{n}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0) = -\left(\mathbb{E}\frac{\partial}{\partial \boldsymbol{\theta}'}\boldsymbol{g}_i(\boldsymbol{\theta}_0,\tau_0)\right)^{-1}\boldsymbol{\nu}_n(\tau_0)+o_P(1).$$

This informally proves:

Theorem (Andrews, 1994)

Under "m-estimator" regularity conditions, stochastic equicontinuity of $\boldsymbol{\nu}_n$, and if also $\sqrt{n}\mathbb{E}\boldsymbol{g}_i(\boldsymbol{\theta}_0,\hat{\tau}) \xrightarrow{p} \mathbf{0}$, then

$$\begin{aligned}
\sqrt{n}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0) &\xrightarrow{d} N(\mathbf{0},\boldsymbol{G}^{-1}\boldsymbol{S}\boldsymbol{G}^{-1}) \\
\boldsymbol{G} &= \mathbb{E}\left(\frac{\partial}{\partial \boldsymbol{\theta}'}\boldsymbol{g}_i(\boldsymbol{\theta}_0,\tau_0)\right) \\
\boldsymbol{S} &= \mathbb{E}(\boldsymbol{g}_i(\boldsymbol{\theta}_0,\tau_0)\boldsymbol{g}_i(\boldsymbol{\theta}_0,\tau_0)').
\end{aligned}$$

10.2.8 Checking the MINPIN Assumptions in Applications

We first verify conditions under which the crucial orthogonality condition holds in partially linear regression. In that case, we have

$$\boldsymbol{g}_i(\boldsymbol{\theta}_0,\hat{\tau}) = \big(\boldsymbol{x}_i-\hat{\tau}_x(\boldsymbol{z}_i)\big)\big(y_i-\hat{\tau}_y(\boldsymbol{z}_i)-(\boldsymbol{x}_i-\hat{\tau}_x(\boldsymbol{z}_i))'\boldsymbol{\theta}_0\big).$$

Again, setting this equal to zero in expectation is an analog of the OLS moment condition $\mathbb{E}(\mathbf{x}_i(y_i-\mathbf{x}_i'\boldsymbol{\beta}))=\mathbf{0}$. Now, recall from (**??**) that the true model can be written as

$$y_i = m_y(\boldsymbol{z}_i)+(\boldsymbol{x}_i-\boldsymbol{m}_x(\boldsymbol{z}_i))'\boldsymbol{\theta}_0+\epsilon_i$$

with $\mathbb{E}(\epsilon_i|\boldsymbol{x}_i,\boldsymbol{z}_i)=0$. Equating τ with $\boldsymbol{m}$ and plugging in, we get

$$\begin{aligned}
\boldsymbol{g}_i(\boldsymbol{\theta}_0,\hat{\tau}) &= \big(\boldsymbol{x}_i-\hat{\boldsymbol{m}}_x(\boldsymbol{z}_i)\big)\big(m_y(\boldsymbol{z}_i)+(\boldsymbol{x}_i-\boldsymbol{m}_x(\boldsymbol{z}_i))'\boldsymbol{\theta}_0+\epsilon_i-\hat{m}_y(\boldsymbol{z}_i)-(\boldsymbol{x}_i-\hat{\boldsymbol{m}}_x(\boldsymbol{z}_i))'\boldsymbol{\theta}_0\big) \\
&= \big(\boldsymbol{x}_i-\hat{\boldsymbol{m}}_x(\boldsymbol{z}_i)\big)\big(m_y(\boldsymbol{z}_i)-\hat{m}_y(\boldsymbol{z}_i)-(\boldsymbol{m}_x(\boldsymbol{z}_i)-\hat{\boldsymbol{m}}_x(\boldsymbol{z}_i))'\boldsymbol{\theta}_0+\epsilon_i\big)
\end{aligned}$$

and therefore

$$\mathbb{E}\boldsymbol{g}_i(\boldsymbol{\theta}_0,\hat{\tau}|\boldsymbol{z}_i) = \big(\boldsymbol{m}_x(\boldsymbol{z}_i)-\hat{\boldsymbol{m}}_x(\boldsymbol{z}_i)\big)\big(m_y(\boldsymbol{z}_i)-\hat{m}_y(\boldsymbol{z}_i)-(\boldsymbol{m}_x(\boldsymbol{z}_i)-\hat{\boldsymbol{m}}_x(\boldsymbol{z}_i))'\boldsymbol{\theta}_0\big).$$

We can now use the Law of Iterated Expectations to write

$$\begin{aligned}
&\sqrt{n}\mathbb{E}\boldsymbol{g}_i(\boldsymbol{\theta}_0,\hat{\tau}) \\
= \;& \sqrt{n}\int(\boldsymbol{m}_x(\boldsymbol{z})-\hat{\boldsymbol{m}}_x(\boldsymbol{z}))\left(m_y(\boldsymbol{z})-\hat{m}_y(\boldsymbol{z})-(\boldsymbol{m}_x(\boldsymbol{z})-\hat{\boldsymbol{m}}_x(\boldsymbol{z}))'\boldsymbol{\theta}_0\right)f_z(\boldsymbol{z})d\boldsymbol{z} \\
\leq \;& \sqrt{n}\left(\sup_{\boldsymbol{z}}\|\boldsymbol{m}_x(\boldsymbol{z})-\hat{\boldsymbol{m}}_x(\boldsymbol{z})\|\times\sup_{\boldsymbol{z}}|m_y(\boldsymbol{z})-\hat{m}_y(\boldsymbol{z})|+\sup_{\boldsymbol{z}}\|\boldsymbol{m}_x(\boldsymbol{z})-\hat{\boldsymbol{m}}_x(\boldsymbol{z})\|^2\|\boldsymbol{\theta}_0\|\right).
\end{aligned}$$

A sufficient and "tight" (we don't generally expect the result otherwise) condition for this is that $\boldsymbol{m}_x(\boldsymbol{z}) - \hat{\boldsymbol{m}}_x(\boldsymbol{z}) = o_P(n^{-1/4})$. We therefore find that Robinson's (1988) condition was tight. (We are loose about uniform convergence here, but the trimming step fixes that.)

In the case of generated regressors, we have $\boldsymbol{g}_i(\boldsymbol{\theta}_0, \tau) = \mathbb{E}\big(\boldsymbol{m}(\boldsymbol{x}_i)(y_i - \boldsymbol{m}(\boldsymbol{x}_i)'\boldsymbol{\theta}_0)\big)$, and we will establish in a homework that the condition fails. Indeed, nonparametrically generated regressors are a difficult topic; see Mammen, Rothe, and Schienle (*Annals of Statistics*, 2012).

10.2.9 Summary

MINPIN theory is a reasonably general framework for thinking about semiparametric models. Again, it applies if the model in question can be thought of as m-estimation, except that the estimator of an infinite dimensional nuisance parameter enters the sample objective function. (Of course, as with partially linear models, the MINPIN estimation problem may be auxiliary to the estimation problem of true economic interest.) The theory analyzes whether the nonparametric pre-estimation step may be (first-order asymptotically) ignored. It has three core ingredients:

- We must have enough regularity so that, if the nuisance parameter τ were known, we would have a conventional m-estimation problem. Necessity of this is clear enough. Loosely speaking, the assumptions are some smoothness and uniform convergence restrictions and a CLT for the score at fixed τ, which is frequently readily available.

- The "error process" ν_n must be stochastically equicontinuous. This has been verified for important cases and is frequently just assumed.

- The "near orthogonality condition," i.e. $\sqrt{n}\mathbb{E}\boldsymbol{g}_i(\boldsymbol{\theta}_0, \hat{\tau}) = o_P(1)$, may or may not hold and must be verified. This is the hard part.

For additional intuition about what's going on here, note that without the orthogonality condition, and if τ were fixed at any particular value, we would still have consistency and even asymptotic normality of our estimator for the implied "pseudotrue" value of $\boldsymbol{\theta}$ which minimizes the population criterion function given that τ. However, if τ is being estimated, that pseudotrue value is itself a moving target. This can be ignored if it converges to our intended estimand at a fast enough rate, but this generally requires the condition (which is therefore tight).

Note also the following: If $\sqrt{n}\mathbb{E}\boldsymbol{g}_i(\boldsymbol{\theta}_0, \hat{\tau}) = O_P(1)$, our algebra suggests that we get parametric rate of convergence though with an erratic limiting distribution. One might feel this is good enough to claim subsequent ignorability of the parametric estimation step. This reasoning is dangerous because we were cavalier about the need for the nonparametric estimators to converge uniformly and also ignored b_n.

The knife-edge manner in which $d = 4$ would imply that the MSE of $\hat{\boldsymbol{m}}$ is of order $O_P(n^{-1/2})$ is therefore spurious. With $d = 3$, we have enough slack in the rate to handle these issues.

11 The Bootstrap

11.1 Introduction

The bootstrap (introduced by the statistician Brad Efron in a seminal paper in 1979) has become an important alternative to asymptotic approximation. We will first try to understand how it works and then discuss its advantages and limitations.[26]

At the most general level, the bootstrap can be understood as extending the analog or "plug-in" principle from estimation to inference. To understand what I mean by plug-in principle, consider any quantity of interest that can be defined as $\theta = g(F)$, where g is a known function and F is the true distribution of the data. Obvious examples are the mean $\mu = \mathbb{E}(x_i)$ of a r.v. or the linear projection $\boldsymbol{\beta} = \left(\mathbb{E}(\boldsymbol{x}_i\boldsymbol{x}_i')\right)^{-1}\mathbb{E}\boldsymbol{x}_i y_i$. Imagine one has available an estimator $\hat{F}$ of F. An obvious such estimator, but not the only conceivable one, is the empirical distribution F_n, i.e. the discrete distribution whose probability mass function coincides with sample frequencies. Then it would be natural to estimate θ by $g(\hat{F})$. If $\hat{F}$ is consistent for F in a sufficiently strong sense and g is continuous, the estimator will be consistent. This idea can be used to motivate plug-in estimators of μ and $\boldsymbol{\beta}$; in both examples just given, the resulting estimators should look familiar.

Now, in principle g could also be the standard deviation or c.d.f. of a given test statistic at a specific sample size n. Then a plug-in estimator of this c.d.f. could be used to estimate a standard error or critical value. That is the basic idea of bootstrap inference. I'll now explain it with some more notation. We will continue to assume that data are i.i.d., but the bootstrap has been extended beyond that setting.

We will focus on two estimands. One is the (scaled) standard deviation of an estimator. Indeed, recall from the previous section that standard errors in quantile regression are usually bootstrapped for practical reasons. The other one is a general distribution

$$J_n(t, F) = \Pr(T_n \leq t|F)$$

of a sample statistic (in practice a test statistic, hence the notation)

$$T_n(\boldsymbol{w}_1, \ldots, \boldsymbol{w}_n, F).$$

Here, F is the population distribution of observables. Note that both T_n and J_n are indexed by sample size n.

[26] This section owes much to Bruce Hansen's lecture notes, to chapter 1 of Politis, Romano, and Wolf's "Subsampling" textbook, and to Joel Horowitz' *Handbook of Econometrics* chapter. Note that this section is about bootstrap-based inference, not about other uses of bootstrap techniques. We will only cursorily mention bootstrap bias correction and not at all bootstrap averaging of estimators ("bagging"), which is an important part of some Machine Learning estimators but is not motivated for estimators discussed in this lecture.

The idea is to estimate J_n by a plug-in estimator using $\hat{F}$:

$$\hat{J}_n(t) = J_n(t, \hat{F}).$$

It is actually conventional to write J_n^* rather than $\hat{J}_n$, and we will follow this convention henceforth. Also, the classic nonparametric bootstrap estimates F by the empirical distribution F_n, and we will initially stick with that, but it is not essential. Somewhat more formally,

$$\begin{aligned} J_n^*(t) &= J_n(t, F_n), \\ F_n(\boldsymbol{w}) &= \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{\boldsymbol{w}_i \leq \boldsymbol{w}\}. \end{aligned}$$

By a modest generalization of the famous Glivenko-Cantelli Theorem, $F_n(\boldsymbol{w}) - F(\boldsymbol{w}) \xrightarrow{a.s.} 0$ uniformly over $\boldsymbol{w} \in \boldsymbol{R}^K$. The obvious hope is that this makes J_n^* a good estimator of J_n. We will later investigate conditions under which this is true.

This looks easy enough, but beware that some bootstrap concepts can initially be confusing. The bootstrap distribution of a sample statistic is itself a sample statistic and therefore nonrandom given the data. However, being a distribution, it characterizes a random variable. This random variable will be important and is the bootstrap analog of the sample statistic in question. We will generally denote bootstrap analogs by asterisks. For example, the bootstrap analog of our observable variable $\boldsymbol{w}_i$ is the r.v. $\boldsymbol{w}_i^*$ distributed uniformly on $\{\boldsymbol{w}_1, \ldots, \boldsymbol{w}_n\}$. Note that $\mathbb{E}\boldsymbol{w}_i^* = \bar{\boldsymbol{w}}$ and $\operatorname{var} \boldsymbol{w}_i^* = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{w}_i - \bar{\boldsymbol{w}})(\boldsymbol{w}_i - \bar{\boldsymbol{w}})'$; that is, the sample moments of $\boldsymbol{w}_i$ are the population moments of $\boldsymbol{w}_i^*$. The bootstrap analog of $\bar{\boldsymbol{w}}$ is the r.v. $\bar{\boldsymbol{w}}^* \equiv \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{w}_i^*$ whose distribution is implied; similarly, any test statistic T_n will have a bootstrap analog T_n^*. However, the randomness in these bootstrap quantities is *not* driven by the randomness of the original data. You may find it conceptually easiest to just think of $J_n(\cdot)$ (and implied critical values etc.) as estimand and bootstrap random variables as a by-product. However, we will see that implementation of the bootstrap will typically involve simulation of bootstrap random variables.

11.2 Implementing the (simple, nonparametric) Bootstrap

Our definition precisely characterizes J_n^*, and so in principle it can be computed explicitly. In particular, imagine an $(n^n \times n)$ array $(\boldsymbol{w}_i^b)_{i=1,\ldots,n}^{b=1,\ldots,n^n}$ whose rows correspond to all possible (and equally likely) bootstrap samples $(\boldsymbol{w}_1^*, \ldots, \boldsymbol{w}_n^*)$, or in other words to all possible sequences of n draws from the uniform distribution on $\{\boldsymbol{w}_1, \ldots, \boldsymbol{w}_n\}$. Then for a statistic $T_n(\boldsymbol{w}_1, \ldots, \boldsymbol{w}_n, F)$ we have

$$J_n^*(t) = \Pr(T_n^* \leq t) = \frac{1}{n^n}\sum_{b=1}^{n^n} \mathbf{1}\{T_n(\boldsymbol{w}_1^b, \ldots, \boldsymbol{w}_n^b, F_n) \leq t\}.$$

Evaluating this would lead to a complete, or idealized, bootstrap. Asymptotic theory of the bootstrap is strictly speaking asymptotic theory of this complete bootstrap; that is, we take $J_n(t, F_n)$ to be a known function of F_n and then show that as $n \to \infty$, $J_n(\cdot, F_n)$ and $J_n(\cdot, F)$ become similar in some stochastic sense.

But in practice, $J_n(\cdot, F_n)$ can typically not be computed because n^n is too large.[27] Other than in toy examples and examples that allow closed-form analysis, the actual estimator is, therefore, a simulation based approximation of $J_n^*(t)$. In particular, one would randomly select $B \ll n^n$ rows from the above array and then compute the above expression only for this smaller array.

For the classic nonparametric bootstrap that we currently discuss, this *Monte Carlo bootstrap* is easy to implement, and thinking through the implementation may aid the intuition. For a size B Monte Carlo bootstrap, simulate B bootstrap samples by drawing $B \times n$ data points from the original, empirical sample *with* replacement. (Why would it be pointless do do this without replacement?) You may figuratively think of putting the n observed data points into an urn and drawing from that urn $B \times n$ times with replacement. Next, compute the bootstrap test statistic T_n^b for each of the B simulated samples. Your simulated distribution of T_n^*, which in turn is your estimated distribution of T_n, is discrete and uniform over the B (not necessarily distinct) support points you just computed, so that $\Pr(T_n \leq t)$ is estimated by $\frac{1}{B}\sum_{b=1}^{B} \mathbf{1}\{T_n^b \leq t\}$ and other features of the distribution of T_n, e.g. its standard deviation, are similarly estimated by bootstrap analogs.

Almost all bootstraps reported in practice are Monte Carlo bootstraps. This is usually suppressed in notation, and we will ignore it henceforth. For the remainder of this subsection only, let $J_n^{MC}(\cdot)$ explicitly refer to a simulation estimate of $J_n^*(\cdot)$. Then we note the following:

- We emphasized that $J_n^*(\cdot)$ is a sample statistic, i.e. it is nonrandom given the data. In contrast, $J_n^{MC}(\cdot)$ is random even given F_n because of the random selection of bootstrap samples. This is usually ignored because in principle, simulation error can be made arbitrarily small by increasing B.

- An obvious question is what B is big enough in practice. You may get away with B on the order of a few hundred in exploratory analysis, if computation is truly costly, or if you are after bootstrap standard errors (which tend to converge much faster than quantiles as B grows). If you bootstrap a quantile, e.g. to get a critical value, and computation is not a big concern, a bootstrap size of $B = 10000$ (or maybe $B = 9999$ because the bootstrap quantile then coincides with a support point of the bootstrap distribution) would be reasonably expected.[28]

[27] While in most applications, many different data vectors, e.g. permutations of the same vector, will give the same T_n, this effect is not nearly large enough to resolve the problem.

[28] For a rough intuition about simulation error, say you want to estimate the 95th quantile of T_n. Letting the true quantile be $c_{95\%}$, a ballpark estimate informed by properties of the binomial distribution is $\sqrt{B}\,(F_n^*(c_{95\%}) - 95\%) \approx$

- For internal and exploratory analyses, it can be necessary to set B low and useful to make sure that random number seeds change. This will automatically alert you to too low B as you continue exploring. For replicable final versions of your analyses, choose B as large as is computationally reasonable and set specific seeds.

- Make sure you don't get confused by the multiple layers of randomness involved in the bootstrap. To reiterate: (i) The data $\boldsymbol{w}_i$, and all sample statistics, are random in just the way they always were. (ii) The bootstrap analog $\boldsymbol{w}_i^*$, and all test statistics based on it, are fictitious random variables whose distributions, e.g. $J_n^*(\cdot)$, are sample statistics, i.e. nonrandom given the data. (But the distributions *are* themselves random from a pre-sample point of view.) (iii) In practice, inference is based on simulated bootstrap objects like $J_n^{MC}(\cdot)$ that *are* random given the data, but we ignore this layer by presuming that B is large. Asymptotic results presented below are strictly speaking about the idealized bootstrap.

Here are some examples.

Example 11.1 *You observed three realizations* $(0, 1, 2)$ *of a random variable* x_i*, thus the sample average is* 1*. Then* F_n *is the uniform distribution on* $\{0, 1, 2\}$*. This distribution characterizes the bootstrap analog* x_i^**, which has expectation* $\mathbb{E}(x^*) = \bar{x} = 1$*. The complete bootstrap distribution of a size 3 resample* (x_1^*, x_2^*, x_3^*) *is uniform on* $\{0, 1, 2\}^3$*, i.e. it has 27 equally sized mass points. We can derive the bootstrap distribution of the sample average* $\bar{x}^*$*: Evaluating all 27 possible bootstrap samples, it turns out that this average is supported on* $(0, 1/3, 2/3, 1, 4/3, 5/3, 2)$ *with probability masses* $(1/27, 3/27, 6/27, 7/27, 6/27, 3/27, 1/27)$*.*

Example 11.2 *Same as above, but you observed six realizations* $(0, 1, 2, 3, 4, 5)$ *of* x_i*. Now* F_n *is the uniform distribution on* $\{0, 1, 2, 3, 4, 5\}$*. The bootstrap distribution of* $(x_1^*, x_2^*, x_3^*, x_4^*, x_5^*, x_6^*)$ *has* $6^6 = 46656$ *mass points. An idealized bootstrap cannot be done manually, but implementing it on the computer would not be a problem.*

Example 11.3 *Same as above, but you have a realistic sample size. Now you have to do a Monte Carlo bootstrap.*

Example 11.4 ***Estimating a Proportion.***

We explain this example in some detail as we will return to it. The r.v. x_i *is distributed Bernoulli with parameter* $\pi = \Pr(x_i = 1) \in (0, 1)$*. You observe a size* n *i.i.d. sample. The obvious estimator of* π *is* $\hat{\pi} = \bar{x}$*. Its exact sampling distribution is characterized by* $n\bar{x}$ *being Binomial with parameters* (n, π)*. An empirical sample consists of* n *data points, of which* $n\bar{x}$ *are successes and* $n(1 - \bar{x})$ *are*

$N(0, .05)$, i.e. the standard deviation of the simulated probability is about $1/\sqrt{20B}$.

failures. The empirical distribution F_n therefore equals $F_n(x) = 1 - \bar{x} + \bar{x}\mathbf{1}\{x = 1\}$, or in other words, the bootstrap variable x_i^* is distributed Bernoulli with parameter $\hat{\pi}$. A bootstrap sample $(x_1^*, \ldots, x_n^*)$ is i.i.d. from that distribution. Define the bootstrap estimator $\hat{\pi}^* = \bar{x}^* = \frac{1}{n}\sum_{i=1}^n x_i^*$, then $n\bar{x}^*$ is distributed Binomial with parameters $(n, \hat{\pi})$. In this particular case, the exact bootstrap distribution can be computed up to sample sizes where the error in Normal approximation is negligible. Also, in this case it is intuitively clear that, due to consistency of $\hat{\pi}$, the procedure "works," i.e. the bootstrap c.d.f. becomes similar to the true sampling c.d.f.

11.3 Basic Concepts in Bootstrap Inference

We next go over the most straightforward, but also widely used, examples of bootstrap based inference. Throughout this discussion, we'll assume that $\theta_0^* = \hat{\theta}$.

Standard Errors As a warm-up, let's think through computation of bootstrap standard errors. Consider an estimator that is unbiased, at least to first order of approximation. (For any other estimator, forming confidence intervals by adding and subtracting standard errors is not motivated.) In particular, we have

$$SE^* = \left(\mathbb{E}(\hat{\theta}^* - \hat{\theta})^2\right)^{1/2},$$

where the expectation is with respect to the (bootstrap) distribution of $\hat{\theta}^*$; recall that $\hat{\theta}$ is not random in the bootstrap population. Again, this can be approximated to arbitrary degree of precision by choosing a high B in

$$SE^{MC} = \left(\frac{1}{B}\sum_{b=1}^{B}(\hat{\theta}^b - \hat{\theta})^2\right)^{1/2},$$

where $\hat{\theta}^b$ is the b'th bootstrap realization of $\hat{\theta}^*$. We will ignore the difference between complete and Monte Carlo bootstrap henceforth. To repeat, the hard theoretical question is not whether SE^{MC} approximates SE^*, but whether SE^* estimates SE.

Many estimators are demonstrably $\sqrt{n}$-consistent and asymptotically normal but with hard-to-estimate asymptotic variances. In such cases, bootstrap standard errors can be extremely convenient. A notable example is the sample median.

Our main focus will be on bootstrap quantile confidence intervals. There are several ways to construct these.

Percentile Interval The bootstrap equal-tailed percentile confidence interval is

$$CI_\alpha^{perc-2} = \left[\hat{\theta} - q_n^*(1-\alpha/2), \hat{\theta} - q_n^*(\alpha/2)\right],$$

where q^* denotes bootstrap analog (hence the asterisk) quantiles of $(\hat{\theta} - \theta_0)$, i.e. $q^*(\alpha) = \inf\{t : \Pr(\hat{\theta}^* - \hat{\theta} \leq t) \geq \alpha\}$. To motivate it, let's first write out a so-called "oracle" confidence interval. This is the interval that we'd like to use if all population quantities were known. It should by construction attain a size of exactly 95%. (It is, of course, a pure thought device because if we actually knew all population quantities, we'd not need a confidence interval to begin with.) Letting $T_n = \hat{\theta} - \theta_0$ with exact quantile function q_n, this confidence interval would be

$$CI^{oracle} = \left[\hat{\theta} - q_n(1-\alpha/2), \hat{\theta} - q_n(\alpha/2)\right]$$

because

$$\begin{aligned}
\Pr(\theta_0 \in CI^{oracle}) &= \Pr\left(\hat{\theta} - q_n(1-\alpha/2) \leq \theta_0 \leq \hat{\theta} - q_n(\alpha/2)\right) \\
&= \Pr\left(\hat{\theta} - q_n(\alpha/2) \leq \hat{\theta} - \theta_0 \leq q_n(1-\alpha/2)\right) \\
&= 1 - \alpha,
\end{aligned}$$

where the last step assumes for simplicity that $\hat{\theta} - \theta_0$ is continuously distributed. The bootstrap percentile interval is a plug-in estimator of this object, i.e. q_n is replaced with its bootstrap analog q_n^*. The interval is exceedingly easy to compute:

Algorithm 11.1 *1. Generate B bootstrap realizations $\hat{\theta}^1, ..., \widehat{\theta}^B$. Let the vector $(\hat{\theta}^{[1]}, ..., \hat{\theta}^{[B]})$ collect them in increasing order.*

2. Up to integer constraints, the bootstrap percentile $(1-\alpha)$ confidence interval is

$$\left[2\hat{\theta} - \hat{\theta}^{[(1-\alpha/2)B]}, 2\hat{\theta} - \hat{\theta}^{[\alpha B/2]}\right].$$

Knowing α, you can of course choose B to avoid integer constraints. Another popular trick is to use $B = 9999$ or similar, so the relevant quantiles of the bootstrap distribution are unique.

The construction contrasts with the historically oldest bootstrap percentile interval, which is

$$CI_\alpha^{perc-1} = \left[\hat{\theta} + q_n^*(\alpha/2), \hat{\theta} + q_n^*(1-\alpha/2)\right].$$

This confidence interval is even (very marginally) easier to compute: In the preceding algorithm's notation, it is just $[\hat{\theta}^{[\alpha B/2]}, \hat{\theta}^{[(1-\alpha/2)B]}]$, i.e. we draw B bootstrap estimates, sort them, and report the numbers indexed $\alpha B/2$ and $B - \alpha B/2$. However, it is not well motivated. Being the bootstrap analog of

$$\left[\hat{\theta} + q_n(\alpha/2), \hat{\theta} + q_n(1-\alpha/2)\right],$$

at first glance there is no reason to expect that it will cover correctly. The two percentile intervals (asymptotically) coincide if the distribution of T_n is (asymptotically) symmetric around 0, so that $q_n(\alpha/2) \approx q_n(1-\alpha/2)$. This is why CI_α^{perc-1} does the job in many practical applications. However, to the extent that T_n has skewness, CI_α^{perc-1} is asymmetric in exactly the wrong way, and undercoverage caused by skewness is expected to be exaggerated.

Percentile-t Intervals These intervals are based on the idea that most confidence regions can be thought of as lower contour sets of sample test statistics. In other words, think about the problem of testing $H_0 : \theta = \theta^*$ and define your confidence region as nonrejection region of the test.

For simplicity, consider first the one-sided interval, thus we test $H_0 : \theta \leq \theta^*$ vs. $H_1 : \theta > \theta^*$ using test statistic $T_n = (\hat{\theta} - \theta^*)/SE(\hat{\theta})$. (We could also bootstrap the distribution of a non-studentized test statistic, but we will later encounter good reasons not to do so.) We want to reject the null if $T_n > c_{1-\alpha}$, where $c_{1-\alpha}$ is the $(1-\alpha)$-quantile of T_n, i.e. $c_{1-\alpha} = \inf\{c : \Pr(T_n \leq c) \geq 1-\alpha\}$, under the null. Exact computation of $c_{1-\alpha}$ is typically elusive in practice, but we can again think of $c_{1-\alpha}$ as an "oracle" quantity that we estimate by its bootstrap analog. Thus, define $c^*_{1-\alpha} = \inf\{c : \Pr(T^*_n \leq c) = 1-\alpha\}$, where $T^*_n = (\hat{\theta}^* - \hat{\theta})/SE^*(\hat{\theta})$. This gives rise to the one-sided percentile-t interval. The two-sided $(1-\alpha)$-interval is the intersection of the one-sided $(1-\alpha/2)$ intervals from above and below. It is *equal-tailed* because, up to approximation error, the same noncoverage probability (of $\alpha/2$) is incurred at either end of the interval. The interval is *not* in general symmetric around $\hat{\theta}$. The interval is also difficult to generalize to parameters that are not scalars.

The latter problem is avoided by the *symmetric* percentile-t interval. In the scalar case, this interval is based on $|T_n|$, thus it cannot be used for one-sided testing. On the other hand, it naturally generalizes to higher-dimensional confidence regions, e.g. to the equivalent of confidence ellipsoids in $\mathbb{R}^2$, by thinking in terms of a quadratic form that generalizes T_n^2. The idea is to test $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}^*$ vs. $H_1 : \boldsymbol{\theta} \neq \boldsymbol{\theta}^*$ and to (ideally) reject H_0 if $T_n > c_{1-\alpha}$, where $T_n = (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)'\hat{\boldsymbol{V}}^{-1}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)$ and $c_{1-\alpha} = \inf\{c : \Pr(T_n \leq c) = 1-\alpha\}$ under the null. Again, $c_{1-\alpha}$ is replaced with its bootstrap analog $c^*_{1-\alpha}$. In sum, we reject if $T_n > c^*_{1-\alpha}$, where

$$c^*_{1-\alpha} = \inf\{c : \Pr(T^*_n \leq c) = 1-\alpha\}, \qquad T^*_n = (\hat{\boldsymbol{\theta}}^* - \hat{\boldsymbol{\theta}})'(\hat{\boldsymbol{V}}^*)^{-1}(\hat{\boldsymbol{\theta}}^* - \hat{\boldsymbol{\theta}}).$$

Note in particular that T^*_n evaluates distance from the estimated value $\hat{\boldsymbol{\theta}}$ and not from the hypothesized value $\boldsymbol{\theta}^*$. This is because the former, and not the latter, is the true parameter value in the bootstrap population.

11.4 Summary

The above examples illustrate a good method for constructing a bootstrap test or confidence interval: First, imagine that you know the population distribution of observables and therefore the exact distribution, at sample size n, of any estimators and test statistics. Imagine further that you were charged with constructing a confidence region, i.e. a data-dependent set that covers $\boldsymbol{\theta}_0$ with probability $1-\alpha$. With the full "oracle" knowledge that you imagine to have, you should typically be able to construct an exact confidence region, i.e. a set whose coverage probability is exactly $1-\alpha$. Next, replace all unknown population quantities in the confidence region with bootstrap analogs, and similarly for tests.

More generally, your bootstrap constructions should aim to mimic as closely as possible the true sampling process. If your data are a clustered sample (as is true for many surveys), your bootstrap procedure should resample clusters. If you have time series dependence, the bootstrap should mimic that. (In practice, this is done by the "block bootstrap" which resamples entire pieces of your time series at a time.) If you test nulls, remember that you want to simulate the behavior of your test statistic on the null, so you better ensure the null is true in the bootstrap population. (With very simple nulls, this is automatic by using $\hat{\boldsymbol{\theta}}$ for the hypothesized value. In more complicated cases, this may require to manipulate the empirical distribution so that it fulfills the null.) And so on and so forth, though details are far beyond the scope of this lecture.

Is the Bootstrap Only Used for Inference? No. Under certain conditions, the bootstrap distribution of an estimator estimates its sampling distribution so precisely that bias in the estimator can be diagnosed. In principle, this can be used to debias the estimator. Also, in some contexts, notably Machine Learning, it may be beneficial to replace estimators with their own bootstrap averages; this is called bootstrap aggregating or *bagging*. However, bootstrap inference as explained above is the by far most frequent use of the bootstrap in econometrics and also has by far the easiest theoretical justification, and we therefore focus on it.

12 Asymptotic Theory for the Bootstrap

Other than getting the implementation wrong – frequently by confusing population quantities and bootstrap analogs –, frequent mistakes of practitioners are twofold:

- To think the bootstrap always works, even where relatively simple (of the sort we did in this lecture) asymptotic approximation theory fails (and is not merely intractable),
- To not appreciate the cases in which the bootstrap **outperforms** asymptotic approximation.

Accordingly, two rules of thumb to keep in mind are the following:

- If you know that asymptotic approximation would fail, notably due to discontinuity of limit distributions as functions of parameters, you should assume that the bootstrap also fails. For a simple intuition, go back to the very beginning of the previous chapter. Would you expect plug-in estimation to work if g is not continuous?
- If limit distributions do not depend on underlying parameters – as is the case with most test statistics considered so far in this lecture! – the bootstrap may be better than asymptotic approximation.

To understand these claims, we will first go through (validity of) bootstrap inference in Example (iv) in detail, including proofs; then develop some rough intuitions; and then (without proof) look at some theorems and so-called asymptotic expansions.

12.1 Consistency of Bootstrap Inference in a Simple Example

We first reconsider Example **??**, i.e. estimation of a binomial proportion. In this example, the step-by-step algorithm for computing a (nonparametric, nonstudentized, percentile) MC bootstrap confidence interval is as follows.

Algorithm 12.1 ***Simple Bootstrap Inference for a Proportion***

1. *Generate B bootstrap resamples $(x_1^b, \dots, x_n^b), b = 1, \dots, B$, by resampling the empirical data with replacement.*
2. *Estimate $J_n(t) = \Pr(\bar{x}_n - \pi_0 \leq t)$ by $J_n^*(t) = F_n(\bar{x}_n^b - \bar{x}_n \leq t) = \frac{1}{B}\sum_{b=1}^{B} \mathbf{1}\{\bar{x}_n^b - \bar{x}_n \leq t\}$. Here, $\bar{x}_n^b = \frac{1}{n}\sum_{i=1}^{n} x_i^b$. Note that we really need only $q_n^*(\alpha/2)$ and $q_n^*(1-\alpha/2)$, where q_n^* is the quantile function corresponding to J_n^*. In practice, they can be estimated by ordering the bootstrap realizations of $(\bar{x}_n^b - \bar{x}_n)$ and reading out the $aB/2$ and $(1-\alpha/2)B$ position (up to integer issues).*

3. The bootstrap CI is $[\bar{x}_n - q_n^*(1-\alpha/2), \bar{x}_n - q_n^*(\alpha/2)]$.

I describe an MC bootstrap here because that is what we will have to do in realistic applications. In the specific example, as discussed earlier, a complete bootstrap may actually be feasible.

Also, in writing down the algorithm, I followed a typical bootstrap convention and avoided scaling by $\sqrt{n}$, which does not matter for implementation. To show that this bootstrap construction works, however, we want to look at the scaled test statistic $T_n = \sqrt{n}(\bar{x} - \pi_0)$ with distribution $J_n(t) = \Pr(\sqrt{n}(\bar{x} - \pi_0) \leq t)$. If $\tilde{q}_n$ is the quantile function corresponding to $J_n(t)$ and $\tilde{q}_n^*$ its bootstrap analog, then $q_n^* = \tilde{q}_n^*/\sqrt{n}$, so the algorithm could have equivalently be described in terms of estimating J_n by J_n^*.

In the example, that the bootstrap r.v. $\sqrt{n}(\bar{x}_n^b - \bar{x}_n)$ is likely to be distributed similarly to $\sqrt{n}(\bar{x}_n - \pi_0)$ is intuitively clear and could also be verified with ad hoc arguments using that the binomial distribution is very well understood. We will now formalize "likely to be distributed similarly" and then prove that is more instructive and generalizable than the ad hoc argument.

Theorem 12.2 *The above estimator of* J_n *is consistent:*

$$\sup_{t \in \boldsymbol{R}} |J_n^*(t) - J_n(t)| \stackrel{a.s.}{\to} 0 \quad as \quad n \to \infty.$$

Proof. We prove this by invoking the Berry-Esseen Theorem: For an i.i.d. r.v. y_i, if $\mathbb{E}(y_i, y_i^2, |y_i|^3) = (0, \sigma^2, \gamma)$ and G_n is the c.d.f. of $\sqrt{n}\bar{y}_n/\sigma$, then

$$\sup_{r \in \boldsymbol{R}} |G_n(r) - \Phi(r)| \leq \frac{C\gamma}{\sigma^3\sqrt{n}}$$

for a universal constant C. The important thing is that this *uniformly* bounds the rate at which a Central Limit Theorem "kicks in" depending only on the values of a few moments (these values have to be finite).

Next, write

$$J_n(t) = \Pr(\sqrt{n}(\bar{x}_n - \pi_0) \leq t) = \Pr\left(\sqrt{\tfrac{n}{\pi_0(1-\pi_0)}}(\bar{x}_n - \pi_0) \leq t/\sqrt{\pi_0(1-\pi_0)}\right)$$

Define $y_i = x_i - \pi_0$, then $\mathbb{E}(y_i, y_i^2, |y|_i^3) = (0, \pi_0(1-\pi_0), \pi_0(1-\pi_0)^3 + (1-\pi_0)\pi_0^3)$. The r.h. probability in the above display therefore is the c.d.f. of $\sqrt{n}\bar{y}_n/\sigma$ evaluated at $r = t/\sqrt{\pi_0(1-\pi_0)}$. Noting that $\pi_0(1-\pi_0)^3 + (1-\pi_0)\pi_0^3 < 1/2$, we can invoke Berry-Esseen to write

$$\sup_{t \in \boldsymbol{R}} \left| J_n(t) - \Phi\left(t/\sqrt{\pi_0(1-\pi_0)}\right)\right| \leq \frac{C}{(\pi_0(1-\pi_0))^{3/2}\sqrt{n}}$$

and similarly that

$$\sup_{t \in \boldsymbol{R}} \left| J_n^*(t) - \Phi\left(t/\sqrt{\bar{x}_n(1-\bar{x}_n)}\right)\right| \leq \frac{C}{(\bar{x}_n(1-\bar{x}_n))^{3/2}\sqrt{n}}.$$

We wrap up by writing

$$\begin{aligned}
& \sup_{t\in\boldsymbol{R}} |J_n(t) - J_n^*(t)| \\
= & \sup_{t\in\boldsymbol{R}} \left| J_n(t) - \Phi\left(\frac{t}{\sqrt{\pi_0(1-\pi_0)}}\right) + \Phi\left(\frac{t}{\sqrt{\pi_0(1-\pi_0)}}\right) - \Phi\left(\frac{t}{\sqrt{\bar{x}_n(1-\bar{x}_n)}}\right) + \Phi\left(\frac{t}{\sqrt{\bar{x}_n(1-\bar{x}_n)}}\right) - J_n^*(t) \right| \\
\leq & \sup_{t\in\boldsymbol{R}} \left| J_n(t) - \Phi\left(\frac{t}{\sqrt{\pi_0(1-\pi_0)}}\right) \right| + \sup_{t\in\boldsymbol{R}} \left| \Phi\left(\frac{t}{\sqrt{\pi_0(1-\pi_0)}}\right) - \Phi\left(\frac{t}{\sqrt{\bar{x}_n(1-\bar{x}_n)}}\right) \right| + \sup_{t\in\boldsymbol{R}} \left| \Phi\left(\frac{t}{\sqrt{\bar{x}_n(1-\bar{x}_n)}}\right) - J_n^*(t) \right|.
\end{aligned}$$

Now, first think of $\bar{x}_n$ as a nonstochastic sequence converging to π_0, then the first and last supremum in the last line above vanish by the Berry-Esseen theorem; the middle one vanishes by elementary arguments. Of course, $\bar{x}_n$ is in fact random, but because $\bar{x}_n \overset{a.s.}{\to} \pi_0$, the argument applies with probability 1. ■

Corollary 12.3 *The confidence intervals defined earlier achieve asymptotic size control:*

$$\Pr\big(\pi_0 \in [\bar{x}_n - q_n^*(1-\alpha/2), \bar{x}_n - q_n^*(\alpha/2)]\big) \to 1-\alpha$$
$$\Pr\big(\pi_0 \in [\bar{x}_n + q_n^*(\alpha/2), \bar{x}_n + q_n^*(1-\alpha/2)]\big) \to 1-\alpha.$$

The proof effectively went through an asymptotic approximation: We showed that the (nonstochastic) sequence of distributions of $\sqrt{n}(\bar{x}_n - \pi_0)$ and the (stochastic) sequence of distributions of $\sqrt{n}(\bar{x}_n^* - \bar{x}_n)$ converge to the same limit. This foreshadows a typical feature of bootstrap consistency proofs: Despite the fact that we estimate J_n by J_n^* and therefore do not seem to "send n to ∞," proofs of the method usually do just that. This is also a hint that, if CLT-based approximation fundamentally fails, bootstrap inference might be affected too.

As a final note, this was a relatively involved proof for an intuitively straightforward application of the bootstrap. While the proof is a bit more general than the example (it actually applies to means and sample averages of well-behaved r.v.'s), this is a first pointer that proofs of results regarding the bootstrap quickly escalate in complexity.

12.2 General Results

We next consider general conditions under which the bootstrap works. Consider the following table.

	actual sample size	**asymptotic limit**
bootstrap	$J_n(t, \hat{F})$	$J_\infty(t, \hat{F})$
population	$J_n(t, F)$	$J_\infty(t, F)$

Here, $J_\infty(t, F) = \lim_{n\to\infty} J_n(t, F)$. The true sampling distribution of interest is on the bottom left of this table. Its bootstrap estimator is on the top left. Asymptotic approximations roughly correspond to the top right entry: We usually show that we know the limiting distribution of an object up to

some parameters (e.g., an asymptotic variance) which we can estimate. This is like the top right entry, though $\hat{F}$ is not necessarily the nonparametric estimator.

Our hope is that $J_n(t, \hat{F}) - J_n(t, F) = o_P(1)$. However, whenever this can be shown, the typical chain of proof is $J_n(t, \hat{F}) - J_\infty(t, \hat{F}) = o_P(1)$, $J_\infty(t, \hat{F}) - J_\infty(t, F) = o_P(1)$, $J_\infty(t, F) - J_n(t, F) = o_P(1)$. Therefore, despite the fact that we seem to estimate a finite sample distribution by a finite sample distribution, *the bootstrap is an asymptotic approximation.* That said, there is an intuition that $J_n(t, \hat{F})$ might pick up additional features of the sampling distribution, e.g. skewness, even if the limiting distribution is normal. Could this imply that the bootstrap approximation is better than the asymptotic one? The answer is 'yes' in especially well-behaved cases, namely if the top right and bottom right entries of the table are in fact the same; else, this additional approximation step invalidates the idea.

To formalize this, it will be important to distinguish the following three cases.

1. A statistic T_n is an *asymptotic pivot*, meaning that $J_\infty(t, F)$ is constant in F. Classic examples are all statistics that you can show to be asymptotically distributed as $N(0, 1)$.[29]
2. $J_\infty(t, F)$ is continuous in F.
3. $J_\infty(t, F)$ is not continuous in F.

The take-home message is that the bootstrap improves on asymptotic approximation in case 1, is consistent in 2, and fails (at least without further modification) in 3. Not coincidentally, asymptotic approximation may also fail in 3.

12.2.1 Consistency

For linear functionals of F_0, a classic result due to Mammen (1992) specifies *necessary and sufficient* conditions for the simple nonparametric bootstrap (and also close variations of it, though we do not discuss that here) to consistently estimate the sampling distribution of a statistic. In a nutshell, these state that such statistics can be boostrapped if and only if they are subject to a Central Limit Theorem. Thus, the simple nonparametric bootstrap does *not* have a fundamental advantage of general validity over CLT-based asymptotic approximation. Indeed, this powerful theorem predicts some failures of the bootstrap that we will discuss later.

[29] To motivate the term asymptotic pivot, note the following definition: A statistic T_n is a *pivot* if $J_n(t, F)$ is constant in F. Of course, this is possible only with restrictions on F. Classic examples are the t- and F-test for OLS under an assumption of normal errors and the Kolmogorov-Smirnov statistic for continuous F.

Theorem 12.4 ***Necessary and Sufficient Conditions for Bootstrap Consistency***

Assume i.i.d. sampling. Fix sequences of functions $\{g_n\}$ *and numbers* $\{t_n, \sigma_n\}$. *Write* $\bar{g}_n = \frac{1}{n}\sum_i g_n(\boldsymbol{w}_i)$ *and* $T_n = (\bar{g}_n - t_n)/\sigma_n$ *with bootstrap analogs* $\bar{g}_n^* = \frac{1}{n}\sum_i g_n(\boldsymbol{w}_i^*)$ *and* $T_n^* = (\bar{g}_n^* - \bar{g}_n)/\sigma_n$. *Then*

$$\lim_{n\to\infty} \Pr(\sup_t |J_n(t, F_n) - J_\infty(t, F_0)| > \varepsilon) = 0$$

if, and only if, $T_n \xrightarrow{d} N(0,1)$.

Example 12.1 ***Sample Mean***

Let the r.v. x_i *have finite expectation* μ *and variance* σ^2. *Let* $T_n = (\bar{x}_n - \mu)/\sigma$ *(i.e.,* (g_n, t_n, σ_n) *are constant sequences) and* $T_n^* = (\bar{x}_n^* - \bar{x}_n)/\sigma$. *Then Theorem* **??** *applies and implies consistency of the bootstrap.*

Note that this example is about the non-studentized sample mean. We standardized T_n and T_n^* only so that Theorem **??** applies as written; the conclusion extends to the same quantities multiplied by σ. The result is also true if the test statistic gets studentized, i.e. divided by a standard error.

Example 12.2 ***Sample Median***

Let x_i *be have density* $f(\cdot)$ *that is strictly positive at the median* $m = \text{med}(x_i)$. *The plug-in estimator of the* m *is* $\hat{m} = q_n(1/2)$, *or more intuitively, the* $(n/2+1)$*-largest sample realization if* n *is even and the* $(n+1)/2$*-largest realization otherwise. We showed in Section* **??** *that* $\sqrt{n}(\hat{m}-m) \xrightarrow{d} N\big((0, 1/(4f(m)^2)\big)$.

Explicit estimation of this asymptotic variance would require estimation of a density. There are tools for that, but it is not recommended here. Instead, we could: (i) use simple nonparametric bootstrap confidence intervals, (ii) use bootstrap standard errors. Note that (i) is not justified by Theorem **??**, illustrating that that result's strength owes to a very structured setting. However, this bootstrap has been formally justified as long as f is continuous around m and $f(m) > 0$ (Bickel and Freedman 1981). The more common confidence interval is actually (ii), which usually computes rapidly, though its validity technically needs the additional condition that $\mathbb{E}|x_i|^\alpha < \infty$ for some $\alpha > 0$ (Ghosh, Parr, Singh, and Babu 1984).

12.2.2 Limitations of the Bootstrap

We will now consider some limitations of the bootstrap, specifically by examining a few salient examples where it fails. The examples are related to important sources of bootstrap failure: discontinuity of J_∞, nonstandard convergence rate of an estimator, and nonnormal limiting distributions.

One important class of examples are parameter-on-the-boundary problems. Recall that standard asymptotic approximations do not apply in such cases because they assume interiority of $\boldsymbol{\theta}_0$, or more

generally smoothness of the estimation problem. Furthermore, in the specific example below, it is easy to see that normal approximation fails: In the critical case of $\mu = 0$, the test statistic's limiting distribution has a mass point, and convergence to normal is not uniform as $\mu \to 0$. So it would be great if the bootstrap worked because this would make it more general than asymptotic approximation; but for that same reason there are also red flags.

Example 12.3 ***Sample Mean on the Boundary (Andrews 2000)***

Revisit the example of a sample mean from i.i.d. data. For simplicity, let $F_0 = N(\mu_0, 1)$. *Suppose the researcher knows that* $\mu_0 \geq 0$. *The obvious (and, for the record, ML) estimator is* $\hat{\mu}_n = \max\{\bar{x}_n, 0\}$ *with bootstrap analog* $\hat{\mu}_n^* = \max\{\bar{x}_n^*, 0\}$.

For any fixed $\mu > 0$, this problem becomes similar to the simple sample mean as $n \to \infty$ and Theorem **??** will apply. However, say $\mu_0 = 0$, thus $\sqrt{n}(\hat{\mu} - \mu_0)$ is exactly distributed as $\max\{z_i, 0\}$, where $z_i \sim N(0,1)$. The bootstrap will be inconsistent in this case. Andrews shows this through the following algebra: Fix any $c > 0$, then

$$
\begin{aligned}
& \Pr\left(\sqrt{n}\left(\hat{\mu}^* - \hat{\mu}\right) \leq t | \sqrt{n}\bar{x}_n \geq c\right) \\
= \; & \Pr\left(\sqrt{n}\left(\max\{\bar{x}_n^*, 0\} - \max\{\bar{x}_n, 0\}\right) \leq t | \sqrt{n}\bar{x}_n \geq c\right) \\
= \; & \Pr\left(\sqrt{n}\left(\max\{\bar{x}_n^* - \bar{x}_n, -\bar{x}_n\}\right) \leq t | \sqrt{n}\bar{x}_n \geq c\right) \qquad \text{using } \bar{x}_n \geq 0 \Rightarrow \max\{\bar{x}_n, 0\} = \bar{x}_n \\
= \; & \Pr\left(\max\{\sqrt{n}\left(\bar{x}_n^* - \bar{x}_n\right), -\sqrt{n}\bar{x}_n\} \leq t | \sqrt{n}\bar{x}_n \geq c\right) \\
\geq \; & \Pr\left(\max\{\sqrt{n}\left(\bar{x}_n^* - \bar{x}_n\right), -c\} \leq t | \sqrt{n}\bar{x}_n \geq c\right) \qquad \text{using } -\sqrt{n}\bar{x}_n \leq -c \\
\to \; & \Pr\left(\max\{z_i, -c\} \leq t | \sqrt{n}\bar{x}_n \geq c\right) \qquad \text{using } \sqrt{n}(\bar{x}_n^* - \bar{x}_n) \xrightarrow{d} N(0,1) \text{ by Berry-Esseen as above} \\
\geq \; & \Pr\left(\max\{z_i, 0\} \leq t | \sqrt{n}\bar{x}_n \geq c\right).
\end{aligned}
$$

The last inequality is strict if $-c < t < 0$, and for any given c, $\Pr(\sqrt{n}\bar{x}_n \geq c) \to \Phi(-c)$ because we assumed $\mu_0 = 0$. Therefore, the distribution of $\sqrt{n}(\hat{\mu} - \mu_0)$ is not correctly estimated, and a percentile t-interval will be invalid.

Alternatively, let's think through what the bootstrap distribution will be. By Berry-Esseen as shown above, the (bootstrap) distribution of $\sqrt{n}(\bar{x}_n^* - \bar{x}_n)$ does consistently estimate the one of $\sqrt{n}(\bar{x}_n - \mathbb{E}x_i)$ (i.e., $N(0,1)$). However,

$$
\begin{aligned}
\sqrt{n}(\hat{\mu}^* - \hat{\mu}) \quad &= \quad \sqrt{n}(\max\{\bar{x}_n^*, 0\} - \max\{\bar{x}_n, 0\}) \\
&\underset{\text{if } \bar{x}_n \geq 0}{=} \quad \sqrt{n}\max\{\bar{x}_n^* - \bar{x}_n, -\bar{x}_n\} \\
&= \quad \max\{z_i, -\sqrt{n}\bar{x}_n\} + o_P(1)
\end{aligned}
$$

and

$$\begin{aligned} \dots \underset{\text{if } \bar{x}_n \leq 0}{=} & \sqrt{n}\max\{\bar{x}_n^*, 0\} \\ = & \sqrt{n}\max\{\bar{x}_n^* - \bar{x}_n + \bar{x}_n, 0\} \\ = & \max\{z_i + \sqrt{n}\bar{x}_n, 0\} + o_P(1). \end{aligned}$$

Either approximation differs from the correct limit distribution $\max\{z_i, 0\}$ through a term involving $\sqrt{n}\bar{x}_n$, a quantity that does not vanish and (ex ante) is itself distributed standard normal. Thus the bootstrap distribution does not converge to any nonstochastic limit.

In the specific example, the problem is easy to spot. The deeper point, however, is that the bootstrap fails if the true parameter value is on the boundary of parameter space, and this, as well as ad hoc fixes, will not always be so obvious as in this example.

For example, a generalization of this problem occurs if the population distribution occurs at a nondifferentiability of the population objective problem's value function. This could be literally a boundary point of parameter space but also a "switching point" of KKT conditions, i.e. a corner of a feasible set. In short, we should bootstrap m-estimators only if we know that the population problem has a smooth maximum. Of course, in any such setting we could in principle (though maybe not in practice) use asymptotic approximation as well.

The example also relates to our previous insights because $J_\infty(\cdot)$ is discontinuous in F. To see this, note that $\sqrt{n}(\hat{\mu} - \mu_0)$ converges to z_i for any $\mu_0 > 0$ but to $\max\{z_i, 0\}$ if $\mu_0 = 0$. In contrast, $J_n(\cdot)$ is continuous in F. These claims are consistent because convergence of J_n to J_∞ is not uniform over F. Thus, for any n, there are true parameter values – notably at and near 0 – for which J_n does not look similar to J_∞. As a result, $J_n(t, F_n) \to J_\infty(t, F_0)$, though true for constant sequences $\{F_n\}$ (with limit $F_0 = F_n$), need not hold over nonconstant sequences $\{F_n\}$. Theorem **??** does not apply because $\hat{\mu}$ cannot be expressed as sample average, but it is suggestive of inconsistency at $\mu = 0$ because T_n does not converge to $N(0, 1)$.

Example 12.4 ***Infinite Variance (Athreya 1987)***

Consider again the same example of a sample mean from i.i.d. data. The assumption that x_i has a finite variance is essential for the result. Indeed, if the variance of x_i is infinite, $J_n^*(t)$ converges to a nondegenerate random variable, hence not to $J_\infty(t)$.

The proof of this example is the centerpiece of an *Annals of Statistics* paper, so we'll take it on faith. (But note that the subsequent result by Mammen strongly suggests it.)

Example 12.5 ***Binomial Proportion near Zero***

Return to Example **??** *but say the true parameter value drifts toward zero: $\pi_n = \gamma/n$ for some $\gamma \in (0, 1]$. Then it is easily verified that $n\bar{x}_n$ converges in distribution to a Poisson distribution with*

parameter γ. Hence, $n(\hat{\pi}_n - \pi_n)$ converges to the same distribution translated by $(-\gamma)$ and therefore centered at 0. The distribution of $n\hat{\pi}_n^*$ is binomial with random parameters $(n, n\hat{\pi}_n)$. Recalling that $n\hat{\pi}_n$ remains stochastically bounded as $n \to \infty$, the distribution of $n\hat{\pi}_n^*$ for large n is well approximated by a Poisson distribution with random parameter $n\hat{\pi}_n$. As $n\hat{\pi}_n$ itself is well approximated by a Poisson distribution with parameter γ, the bootstrap distribution does not converge to any limit and in particular not to the estimand.

This example is important because estimation of small probabilities is a relevant problem. For example, rare events are important in economics, finance, and medical and biostatistics, and propensity scores near 0 or 1 are a frequent issue in estimation of treatment effects.

We note that, from arguments made in the example, $T_n = n(\hat{\pi}_n - \pi_n)/\sqrt{\gamma}$ converges to a r.v. with unit variance, but this r.v. is not normal. Because Theorem **??** is otherwise applicable, it establishes the bootstrap failure that we also directly verified. Indeed, that the bootstrap distribution fails to converge anywhere would have been predictable from a close look at Mammen's proof.[30]

In this example, we can furthermore verify that percentile confidence intervals will typically undercover. The empirical distribution is degenerate at 0 with probability $(1 - \gamma/n)^n \to e^{-\gamma}$. Whenever that happens, the bootstrap percentile confidence interval is the degenerate interval $[0, 0]$. Hence, for any nominal size, the true size of this confidence interval converges to at most $1 - e^{-\gamma}$. For example, $1 - e^{-1} \approx .63$, so if $\gamma = 1$, the interval undercovers compared to any conventional level.

Example 12.6 ***Absolute Value of Mean.***

Let x_i be i.i.d. with expected value $\mu_0 = \mathbb{E}x_i$ and variance σ^2. We are interested in $\theta_0 = |\mu_0|$, which we estimate by $\hat{\theta} = |\bar{x}_n|$. The question is if bootstrap inference works; in particular, if we can use $J_n^*(t) = \Pr(\sqrt{n}\hat{\theta}^* - \hat{\theta} \leq t)$ as estimator of $J_n(t) = \Pr(\sqrt{n}\hat{\theta} - \theta_0 \leq t)$.

We first note the true limiting distribution of $\sqrt{n}(\hat{\theta} - \theta_0)$: it is $N(0, \sigma^2)$ if $\theta_0 \neq 0$ but it is the distribution of $|z_i|$, where $z_i \sim N(0, \sigma^2)$, if $\theta_0 = 0$. This discontinuity, and also the nonnormality of the limiting distribution at θ_0, are clear signs that something might go wrong at that parameter value (which is furthermore on the boundary of the relevant parameter space). We also note and use freely that by the Berry-Esseen theorem, $\sqrt{n}(\hat{\mu}^* - \hat{\mu}) \xrightarrow{d} N(0, \sigma^2)$.

The bootstrap distribution $J_n^*(\cdot)$ is the distribution of

$$\sqrt{n}(\hat{\theta}^* - \hat{\theta}) = \sqrt{n}(\max\{\hat{\mu}^*, -\hat{\mu}^*\} - \max\{\hat{\mu}, -\hat{\mu}\}).$$

[30] An important step in that proof can be roughly verbalized as follows: If the bootstrap converges anywhere, then no single observation is influential in the limit, and then a CLT applies.

Suppose first $\hat{\mu} > 0$. Then we can write

$$\begin{aligned}
& \sqrt{n}(\max\{\hat{\mu}^*, -\hat{\mu}^*\} - \max\{\hat{\mu}, -\hat{\mu}\}) \\
= \quad & \sqrt{n}(\max\{\hat{\mu}^*, -\hat{\mu}^*\} - \hat{\mu}) \\
= \quad & \sqrt{n}(\max\{\hat{\mu}^* - \hat{\mu}, -\hat{\mu}^* - \hat{\mu}\}) \\
= \quad & \sqrt{n}(\max\{\hat{\mu}^* - \hat{\mu}, -2\hat{\mu} - (\hat{\mu}^* - \hat{\mu})\}) \\
= \quad & \max\{\sqrt{n}(\hat{\mu}^* - \hat{\mu}), -2\sqrt{n}\hat{\mu} - \sqrt{n}(\hat{\mu}^* - \hat{\mu}))\} \\
\xrightarrow{d} \quad & \max\{z_i, -2\sqrt{n}\hat{\mu} - z_i\} = \max\{z_i, -2\sqrt{n}\hat{\theta} - z_i\},
\end{aligned}$$

For fixed positive value of μ, this case is the relevant one with probability approaching 1 and furthermore the r.h. argument of the max will diverge to $-\infty$, so we are left (with high probability) with z_i as desired. Similarly, for $\hat{\mu} < 0$, we get

$$\begin{aligned}
& \sqrt{n}(\max\{\hat{\mu}^*, -\hat{\mu}^*\} - \max\{\hat{\mu}, -\hat{\mu}\}) \\
= \quad & \sqrt{n}(\max\{\hat{\mu}^*, -\hat{\mu}^*\} + \hat{\mu}) \\
\xrightarrow{d} \quad & \max\{-z_i, 2\sqrt{n}\hat{\mu} + z_i\} = \max\{z_i, -2\sqrt{n}\hat{\theta} - z_i\},
\end{aligned}$$

where we used that in this case, $\hat{\theta} = -\hat{\mu}$ and that the distribution of z_i is symmetric, so it does not change the expression to replace z_i with $-z_i$.

We now see that if $\mu_0 \neq 0$, we are asymptotically in exactly one of the two cases and the limiting distribution is just the one of z_i. However, if $\mu = 0$, then $\sqrt{n}\hat{\theta}$ converges not to any number but is itself asymptotically distributed as $|z_i|$. The max operators in the above expression then remain relevant, and the bootstrap distribution does not converge anywhere; also, its support generally includes some negative numbers, which is obviously wrong as $\hat{\theta} - \theta \geq 0$ in this case.

Example 12.7 ***Order Statistic.***

Let us reconsider the maximum of a uniform distribution example from Section **??**. *Thus, x_i is distributed $U[0, \alpha_0]$, and we want to estimate α_0. Recall that the obvious estimator is the sample maximum, but also that the usual $\sqrt{n}$-asymptotics for m-estimation do not apply due to discontinuity of the objective function. Indeed, the true rate of convergence is n, that is, the estimator is superconsistent, and in particular we have*

$$\begin{aligned}
\Pr(n(\hat{\alpha} - \alpha_0) \leq t) &= \Pr\big(\max_{i=1,\ldots,n}\{x_i\} \leq \alpha_0 + t/n\big) = \Pr\big(x_1 \leq \alpha_0 + t/n, \ldots, x_n \leq \alpha_0 + t/n\big) \\
&= \Pr\big(x_1/\alpha_0 \leq 1 + t/(n\alpha_0), \ldots, x_n/\alpha_0 \leq 1 + t/(n\alpha_0)\big) = (1 + t/(n\alpha_0))^n \to e^{t/\alpha_0}.
\end{aligned}$$

Could be bootstrap this? Informally, both superconsistency and failure of asymptotic normality

are red flags. To see that the bootstrap fails, write

$$\begin{aligned}\Pr(n(\hat{\alpha}^* - \hat{\alpha}) = 0) = \Pr\left(\hat{\alpha}^* = \hat{\alpha}\right) = \Pr\Big(\max_{i=1,\ldots,n}\{x_i^*\} = \max_{i=1,\ldots,n}\{x_i\}\Big) \\ = 1 - \Pr\Big(\max_{i=1,\ldots,n}\{x_i^*\} < \max_{i=1,\ldots,n}\{x_i\}\Big) = 1 - \Pr\Big(x_i^* < \max_{i=1,\ldots,n}\{x_i\}, \ldots, x_n^* < \max_{i=1,\ldots,n}\{x_i\}\Big) \\ = 1 - (1 - 1/n)^n \to 1 - e^{-1} \approx 0.63.\end{aligned}$$

So this distribution has a (rather large) mass point at 0 including in the limit, whereas the true asymptotic distribution is continuous. Indeed, it is a "tedious but trivial" exercise to characterize the entire random c.d.f. of $n(\hat{\alpha}^* - \hat{\alpha})$. This distribution is supported on $n(x_1 - \max_{i=1,\ldots,n}\{x_i\}, \ldots, x_n - \max_{i=1,\ldots,n}\{x_i\})$. The highest of these mass points is at 0 and has mass $1 - (1 - 1/n)^n$ as per our algebra, the next one has mass $(1 - 1/n)^n \times (1 - 1/(n-1))^n$, and the lowest one has mass n^{-n}. While these numbers are not random, the location of the mass points is. Once again, the bootstrap distribution therefore does not converge to any limit.

This example is more removed from the above theorem because that result's assumptions obviously do not apply. It is, however, practically important and also an example of normal approximation failure, so that it would be very nice for the simple bootstrap to work. Unfortunately, we again get bootstrap nonconvergence, illustrating that higher generality compared to normal approximation is *not* an advantage of the bootstrap.

Example 12.8 ***(Smoothed) Maximum Score***

Let's reconsider the Maximum Score estimator from Section **??**. *Analysis of the bootstrap for this example is very involved. It turns out that the simple nonparametric bootstrap is inconsistent here (Abrevaya and Huang, 2005) and fails to converge (contrary to a claim in the paper just cited). Valid inference for this estimator using modified bootstrap techniques is the subject of current research.*

The Smoothed Maximum Score estimator is much better behaved. Not only can it be bootstrapped; Horowitz (2002) shows that for test statistics based on a studentized estimator, the bootstrap achieves asymptotic refinement.