ECON7772 - Econometric Methods

Lecture Notes from Arthur Lewbel's lectures

Paul Anthony Sarkis
Boston College

Contents

Chapter 1

Regression, Correlation and Causes

Chapter 2

Finite Sample Properties of Estimators

Throughout this section, we'll define $\mathbf{X}$ and $\mathbf{Z}$ as random vectors of size j and k resp. while a, b will be (following the context) either scalars or vectors and A a matrix. Also, we assume a perfect knowledge of moments of distribution ; this chapter only constitutes a quick review. If you find yourself needing anymore information on these definitions and properties, you should go back to the first semester class notes.

Definition 2.1. *A random vector $\mathbf{X}$ of size j is a vector consisting of j random variables $(X_1, ..., X_j)$. Its expectation, $\mathrm{E}[\mathbf{X}]$ is the vector consisting of the expectations of all its elements, namely $(\mathrm{E}[X_1], ..., \mathrm{E}[X_j])$.*

Definition 2.2. *The variance matrix of a random vector $\mathbf{X}$, denoted $\mathrm{Var}[\mathbf{X}]$ is the $j \times j$ matrix equal to $\mathrm{E}[(\mathbf{X} - \mathrm{E}[\mathbf{X}])(\mathbf{X} - \mathrm{E}[\mathbf{X}])']$*

Definition 2.3. *The covariance $\mathrm{Cov}(\mathbf{X}, \mathbf{Z})$ between two vectors $\mathbf{X}$ and $\mathbf{Z}$ is equal to $E\left[(\mathbf{X} - \mathrm{E}[\mathbf{X}])(\mathbf{Z} - \mathrm{E}[\mathbf{Z}])'\right]$*

Remark 2.1. *The variance of the vector $A\mathbf{X} + b$ is*

$$\mathrm{Var}[A\mathbf{X} + b] = A\,\mathrm{Var}[\mathbf{X}]\,A'$$

The covariance between vectors $A\mathbf{X} + b$ and $C\mathbf{Z} + d$ is

$$\mathrm{Cov}(A\mathbf{X} + b, C\mathbf{Z} + d) = A\,\mathrm{Cov}(\mathbf{X}, \mathbf{Z})\,C'$$

Definition 2.4. *If $\mathbf{X} \sim N(\mu, \Omega)$, we say that that $\mathbf{X}$ follows a joint distribution. This distribution is a multi-variate normal distribution of mean μ and variance Ω.*

Remark 2.2. *The vector $A\mathbf{X} + b$ also follows a joint multi-variate normal distribution, of mean $A\mu + b$ and variance $A\Omega A'$.*
The vector $(\mathbf{X} - \mu)'\Omega^{-1}(\mathbf{X} - \mu)$ follows a chi-squared distribution with j degrees of freedom; we write $(\mathbf{X} - \mu)'\Omega^{-1}(\mathbf{X} - \mu) \sim \chi^2_j$

Differentiation

Differentiating a vector by another is equivalent to taking the gradient of the first vector, wrt the second. * $\frac{\partial a'\mathbf{X}}{\partial a} = \left(\frac{\partial a'\mathbf{X}}{\partial a_1}, ..., \frac{\partial a'\mathbf{X}}{\partial a_j}\right)'$ * $\frac{\partial a'\mathbf{X}a}{\partial a} = (\mathbf{X} + \mathbf{X}')a$, for any square matrix $\mathbf{X}$. * $\frac{\partial \operatorname{E}[g(a\mathbf{X})]}{\partial a} = E\left[\frac{\partial g(a\mathbf{X})}{\partial a}\right]$

Law of iterated expectations

$\operatorname{E}[\mathbf{X}] = E\big[\operatorname{E}[\mathbf{X}|\mathbf{Z}]\big]$

Independence(s) and correlation

Let $\mathbf{X}$ and $\mathbf{Z}$ be any two random vectors with means $\mu_{\mathbf{X}}$ and μ_Z resp.

Def: $\mathbf{X}$ and $\mathbf{Z}$ are said to be independent, $\mathbf{X} \perp \mathbf{Z}$, if nothing can be said about $\mathbf{X}$'s distribution from $\mathbf{Z}$.

Def: $\mathbf{X}$ and $\mathbf{Z}$ are said to be mean-independent if $\operatorname{E}[\mathbf{X}|\mathbf{Z}] = \mu_X$

Def: $\mathbf{X}$ and $\mathbf{Z}$ are said to be uncorrelated, or linearly independent, if $\operatorname{E}[\mathbf{XZ}] = \mu_X\mu_Z \Leftrightarrow cov(\mathbf{X}, \mathbf{Z}) = 0$

These definitions are ranked from the strongest to the weakest.

Proofs

Proposition : Let $Z_1, Z_2, ...$ be a sequence of random variables such that, for all i, $E[Z_i] = \mu$. Let X_n be the sample average over the n first variables. The expectation of the sample average is equal to the expectation of each variable :

$$E[X_n] = E[Z_i] = \mu$$

Proof : $E\big[\frac{\sum_{i=1}^{n} Z_i}{n}\big] = \frac{1}{n}\sum_{i=1}^{n} E[Z_i] = \frac{n\mu}{n} = \mu$

The sample average is said to be an **unbiased estimator** of the random variable Z's expectation.

Definition : Let θ_0 be the true value of a parameter from any distribution. Let $\hat{\theta}$ be an estimator of θ_0. We define the **bias of an estimator** to be the absolute deviation between the true value of the parameter and the expectation of its estimator.

$$\text{Bias}(\hat{\theta}) = |\theta_0 - E[\hat{\theta}]|$$

We say that an estimator is unbiased iff its bias is equal to 0.

Proposition : Proposition : Let $Z_1, Z_2, ...$ be a sequence of **i.i.d. random variables** such that, for all i, $E[Z_i] = \mu$ and $Var[Z_i] = \sigma^2$. Let X_n be the sample average over the n first variables. The variance of the sample average is equal to the variance of each variable, divided by the sample size :

$$Var[X_n] = \frac{Var[Z_i]}{n} = \frac{\sigma^2}{n}$$

Proof : $Var[X_n] = E\big[\big(X_n - E[X_n]\big)^2\big] = E\big[\big(X_n - \mu\big)^2\big] = E\big[\big(\frac{1}{n}\sum_{i=0}^{n} Z_i - \mu\big)^2\big] = \frac{1}{n^2}E\big[\big(\sum_{i=0}^{n} Z_i - \mu\big)\big(\sum_{j=0}^{n} Z_j - \mu\big)\big] = \frac{1}{n^2}E\big[\sum_{i=0}^{n}\sum_{j=0}^{n}(Z_i - \mu)(Z_j - \mu)\big]$

$= \frac{1}{n^2}E\big[\sum_{i=0}^{n}(Z_i - \mu)^2 + \sum_{i=0}^{n}\sum_{j\neq i}^{n}(Z_i - \mu)(Z_j - \mu)\big] = \frac{1}{n^2}\big(\sum_{i=0}^{n} Var[Z_i] + \sum_{i=0}^{n}\sum_{j\neq i}^{n} Cov[Z_i, Z_j]\big) = \frac{Var[Z_i]}{n} = \frac{\sigma^2}{n}$

Proposition : Let $Z_1, Z_2, ...$ be a sequence of **i.i.d. random variables** such that, for all i, $E[Z_i] = \mu$ and $Var[Z_i] = \sigma^2$. Let $\hat{\sigma}_n$ be the sample variance, adjusted for the degrees of freedom, over the n first variables. The sample variance is an unbiased estimator of σ^2.

Proof :

2.7 Efficiency of estimators

Definition : Among a number of estimators of the same class, the estimator having the least variance is called an **efficient estimator**. Let $\hat{\theta}_1$ and $\hat{\theta}_2$ be two

estimators of the same parameter θ, then if $Var[\hat{\theta}_1] < Var[\hat{\theta}_2]$, we say that $\hat{\theta}_1$ is a more efficient estimator than $\hat{\theta}_2$. The estimator with the least possible variance (Cramer-Rao bound at the minimum) is called the **most efficient estimator**.

Chapter 3

Asymptotic Properties of Estimators

Definition 3.1. *Mean squared error, or MSE, is the expectation of the squared deviation between the estimator and the true value of the estimand.*

$$MSE(\hat{\theta}) = E_{\hat{\theta}}\left[(\hat{\theta} - \theta)^2\right]$$

Definition 3.2. *Let $\{X_n\}$ denote a sequence of random variables and c a real number such that $\mathrm{E}[X_i] = \mu_i$ and $Var[X_i] = \sigma_i^2$. We say that X_n converges in mean squared error to c if $\lim_{n\to\infty} X_n = c$ and $\lim_{n\to\infty} \sigma_n^2 = 0$. We write :*

$$X_n \xrightarrow{ms} c$$

Remark 3.1. *Let $\{Z_n\}$ be a sequence of i.i.d. random variables such that, for all i, $\mathrm{E}[Z_i] = \mu$ and $Var[Z_i] = \sigma^2$. Let X_n be the sample average over the n first variables. X_n converges in MSE to the true value of $\mathrm{E}[Z_i] = \mu$.*

Definition 3.3. *Let $\{X_n\}$ denote a sequence of random variables and c a real number. We say that X_n converges in probability to c if, for all ϵ, we have that*

$$\lim_{n\to\infty} Pr\left[|X_n - c| > \epsilon\right] = 0$$

We write $X_n \xrightarrow{p} c$.
Moreover, we say that X_n converges in probability to a random variable X if $(X_n - X) \xrightarrow{p} 0$.

This definition allows us to define a useful characteristic of estimators, namely consistency. An estimator that converges in probability to the true value of its estimand is said to be a consistent estimator.

Remark 3.2. *If a sequence of random variables converges in MSE to a variable c, then it also converges in probability to c. The converse is not true.*

Proof. From Chebychev's inequality, we can write that $\Pr[|X_n - \mu_n| > \epsilon] \leq \frac{\sigma_n^2}{\epsilon}$, for all $\epsilon > 0$. Therefore, we have that

$$0 \leq \lim_{n\to\infty} \Pr[|X_n - \mu_n| > \epsilon] \leq \lim_{n\to\infty} \frac{\sigma_n^2}{\epsilon}$$

and from the assumption of convergence in MSE, we know that $\lim_{n\to\infty} \frac{\sigma_n^2}{\epsilon} = 0$.

We indeed have that $\lim_{n\to\infty} \Pr[|X_n - c| > \epsilon] = 0$ □

Theorem 3.1 (Weak Law of Large Numbers)**.** *Let $\{Z_n\}$ denote a sequence of iid random variables such that $\mathrm{E}[Z_i] = \mu$ and $\mathrm{Var}[Z_i] = \sigma^2$. Let X_n be the sample average of $Z_1, ..., Z_n$, then $X_n \xrightarrow{p} \mu$.*

Proof. We already know that $\mathrm{E}[X_n] = \mu$ and $\mathrm{Var}[X_n] = \frac{\sigma^2}{n}$, therefore $X_n \xrightarrow{\text{ms}} \mu \Rightarrow X_n \xrightarrow{\text{p}} \mu$ □

Theorem 3.2 (Khinchin's WLLN)**.** *Let $\{Z_n\}$ denote a sequence of iid random variables such that $\mathrm{E}[Z_i] = \mu$ and $\mathrm{E}[|Z_i|]$ is finite. Let X_n be the sample average of $Z_1, ..., Z_n$, then $X_n \xrightarrow{p} \mu$.*

These two theorems are pretty useful because they show that for any sequence of iid random variables having a finite variance, the sample average associated will converge in probability to the true mean of the random variables. The next theorem shows what we can say when the sequence is not iid.

Theorem 3.3. *Let $\{Z_n\}$ denote a sequence of random variables such that $\mathrm{E}[Z_i] = \mu_i$ and $\mathrm{Cov}(Z_i, Z_j) = \sigma_{i,j} \forall i \neq j$. Let X_n be the sample average of the first n variables. We denote $\bar{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} \mu_i$ and $\mu_0 = \lim_{n\to\infty} \mu_n$. If μ_0 exists and $\lim_{n\to\infty} \mathrm{Var}\left[\bar{Z}\right] = 0$, then $X_n \xrightarrow{p} \mu_0$.*

Proof. It is trivial to show that $\mathrm{E}[X_n] = \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}[Z_i] = \frac{1}{n}\sum_{i=1}^{n}\mu_i = \mu_n$ and therefore $\lim_{n\to\infty}\mathrm{E}[X_n] = \mu_0$; if μ_0 exists. By assumption, $\lim_{n\to\infty}\mathrm{Var}[X_n] = 0$, therefore, we have that $X_n \xrightarrow{\mathrm{p}} \mu_0$. □

This last theorem relies on two (really) strong assumptions :

1. μ_0 exists : this assumption is true if the sequence of random variables (which are not iid) somehow have convergent means, which is far from guaranteed.
2. $\lim_{n\to\infty}\mathrm{Var}[X_n] = 0$: this assumption relies on the fact that Z_is should tend to be more and more uncorrelated as well as having low variances. This can be shown by the fact that $\mathrm{Var}[X_n] = \frac{1}{n^2}\sum\mathrm{Var}[Z_i] + \frac{1}{n^2}\sum\sum\mathrm{Cov}(Z_i, Z_j)$

Theorem 3.4 (Slutsky's Theorem)**.** *For any continuous function $g(\cdot)$ that does not depend on the sample size n, we have:*

$$\textit{plim}\; g(X_n) = g(\textit{plim}\; X_n)$$

Definition 3.4. *Let $\{X_n\}$ denote a sequence of random variables with cdf $F_n(\cdot)$. We say that X_n converges in distribution to a random variable X if*

$$\lim_{n\to\infty} F_n(x) = F(x)$$

at all continuous points of the cdf $F(x)$. We write $X_n \xrightarrow{d} X$.

Remark 3.3. *If a sequence of random variables converges in probability to a random variable X, then it also converges in distribution to X. The converse is not true.*

Convergence in probability and convergence in distribution can also be used together to give some interesting properties.

Remark 3.4. *Let $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{p} c$ where X is a random variable and c a constant. We have :*

- $X_n Y_n \xrightarrow{d} Xc$
- $X_n + Y_n \xrightarrow{d} X + c$

Theorem 3.5 (Slutsky's Theorem for convergence in distribution)**.** *For any continuous function $g(\cdot)$ that does not depend on the sample size n and can be used to represent a distribution, we have:*

$$X_n \xrightarrow{d} X \Rightarrow g(X_n) \xrightarrow{d} g(X)$$

Now we will turn our heads on one of the most important theorem of statistics (and maybe mathematics) that will be used throughout the course to understand asymptotic behavior of estimators.

Theorem 3.6 (Lindeberg-Lévy Central Limit Theorem)**.** *Let $\{Z_n\}$ denote a sequence of iid random variables such that $\mathrm{E}[Z_i] = \mu$ and $\mathrm{Var}[Z_i] = \sigma^2 < \infty$. As always, consider X_n the sample average of the n first random variables Z_i. As n approaches infinity, the random variable $\sqrt{n}(X_n - \mu)$ converges in distribution to a normal distribution of mean 0 and variance σ^2.*

$$\sqrt{n}(X_n - \mu) \xrightarrow{d} N(0, \sigma^2)$$

We say that X_n asymptotically follows a normal distribution $N(0, \sigma^2)$.

Each random variable defined here could also be a random vector. Then we'd assume $\mathrm{Var}[Z_i] = \Omega$ its variance matrix.

Delta Method

Consider a random variable X_n such that $\sqrt{n}(X_n - a) \xrightarrow{\mathrm{d}} Y$. Moreover, let $g(\cdot)$ be a function such that:

1. $g(\cdot)$ is a continuous function, differentiable in the neigborhood of a.
2. $\frac{\partial g}{\partial a^T} \neq 0$ and is finite.
3. $\frac{\partial g}{\partial n} = 0$.

Then,

$$\sqrt{n}(g(X_n) - g(a)) \xrightarrow{\mathrm{d}} \frac{\partial g}{\partial a^T} Y$$

Ex.1. (Normal and Chi-square distributions). Let Z_n be a sequence of i.i.d. random variables such that $\sqrt{n}(Z_n - \alpha) \xrightarrow{\mathrm{d}} N(0, 1)$. Consider the function $g(x) = x^2$.

- By the Delta method, we have that $\sqrt{n}(Z_n^2 - \alpha^2) \xrightarrow{\text{d}} N(0, \left(\frac{\partial g}{\partial \alpha}\right)^2)$ where $\frac{\partial g}{\partial \alpha} = 2\alpha$

Theorem 3.7 (Lindeberg-Feller Central Limit Theorem). *Let $\{Z_n\}$ denote a sequence of independent (but not necessarily identically distributed) random variables such that $\mathrm{E}[Z_i] = \mu_i$ and $\mathrm{Var}[Z_i] = \sigma_i^2 < \infty$. Consider the sample average X_n and the sample variance $\bar{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^n \sigma_i^2$. Suppose*

$$\lim_{n\to\infty} \max_i \frac{\sigma_i^2}{n\bar{\sigma}_n^2} = 0 \text{ and } \lim_{n\to\infty} \bar{\sigma}_n^2 = \bar{\sigma}^2 < \infty$$

Then, $\sqrt{n}\frac{(X_n - \bar{\mu})}{\bar{\sigma}} \xrightarrow{d} N(0,1)$.

Definition 3.5 (Little-o notation). *Let $\{C_n\}$ be a sequence of constants. We say that:*

- *C_n is $o(1)$ if $\lim_{n\to\infty} C_n = 0$, we write: $C_n = o(1)$.*
- *C_n is $o(n^k)$ if $\frac{C_n}{n^k} = o(1)$, we write: $C_n = o(n^k)$.*

The intuition behind this notation is to convey the meaning that the sequence C_n converges to 0 faster that the function inside the operator (1 or n^k). In words, we could say that C_n is ultimately smaller than 1 or n^k. There exists an equivalent definition for a random variable, related to the convergence in probability.

Definition 3.6 (Convergence in probability). *Let $\{X_n\}$ be a sequence of random variables. $X_n = o_p(1)$ if, for all $\varepsilon > 0$ and $\delta > 0$, there exists an N for which $n > N$ implies $\Pr(|X_n| > \varepsilon) < \delta$. Because it is basically the definition of having a plim of 0, we can say that:*

- *X_n is $o_p(1)$ if $\operatorname{plim}_{n\to\infty} X_n = 0$, we write: $X_n = o_p(1)$.*
- *X_n is $o_p(n^k)$ if $\frac{X_n}{n^k} = o_p(1)$, we write: $X_n = o_p(n^k)$.*

This means that, in the limit, the set of values X_n or $\frac{X_n}{n^k}$ will converge to 0 in probability.

Definition 3.7 (Big-O notation). *Let $\{C_n\}$ be a sequence of constants. We say that:*

- *C_n is $O(1)$ if $|\lim_{n\to\infty} C_n| \leq c$, we write: $C_n = O(1)$.*

- *C_n is $O(n^k)$ if $\frac{C_n}{n^k} = O(1)$, we write: $C_n = O(n^k)$.*

Definition 3.8 (Stochastic boundedness)**.** *Let $\{X_n\}$ be a sequence of random variables. $X_n = O_p(1)$ if for all $\delta > 0$ and associated $K_\delta > 0$, there exists a $\tilde{n}$ such that $n > \tilde{n}$ implies that*

$$\Pr(|X_n| > K_\delta) < \delta$$

$X_n = O_p(n^k)$ if $\frac{X_n}{n^k} = O_p(1)$.

Remark 3.5. *If $X_n = o_p(1)$, then $X_n = O_p(1)$. Trivially, this also means that if $X_n = o_p(n^k)$, then $X_n = O_p(n^k)$.*

This is an equivalent statement that saying that a converging sequence must be bounded.

Proof. □

Definition 3.9 (Influence function)**.** *Let $\hat{\theta}$ be a function of random variables $F(Z_1, ..., Z_n)$. Suppose there exists $R_i = r_i(Z_1, ..., Z_n, \theta_0)$ and $S_n = o_p(1)$ such that*

$$\sqrt{n}(\hat{\theta} - \theta_0) = \sqrt{n}\bar{R} + S_n$$

We can simplify by writing $S_n = o_p(1)$ and then,

$$\hat{\theta} = \theta_0 + \bar{R} + O_p(n^{-\frac{1}{2}})$$

Now suppose that $\sqrt{n}\bar{R} \xrightarrow{d} N(0, \Omega)$, then $\sqrt{n}(\hat{\theta} - \theta_0) = O_p(1)$. We say that $\hat{\theta}$ is a root-n-consistent estimator.

Theorem 3.8. *Consider an extremum estimator $\hat{\theta}$ such that $\hat{\theta} \in \arg\max_\theta Q_n(\theta)$. Define $Q_0(\theta)$ as the limit in probability of $Q_n(\theta)$. Next, we assume:*

- ***A1. Identification:*** *$Q_0(\theta)$ exists and is maximized at the true value of the parameter $\theta = \theta_0$*
- ***A2. Continuity:*** *$Q_n(\theta)$ is differentiable.*
- ***A3. Compactness:*** *The domain of $Q_n(\theta)$ is compact (i.e. there exists θ_L and θ_U such that $\theta_L \leq \theta \leq \theta_U$).*
- ***A4. Stochastic equicontinuity:*** *$|\frac{\partial Q_n(\theta)}{\partial \theta}| = O_p(1)$ where δ and K_δ do not depend on θ.*

If these four axioms are satisfied, then $\hat{\theta}$ is a consistent estimator of θ_0, that is, it converges in probability to the true value θ_0.

Consistency of the OLS estimator

Define a model as

$$Y = b_0 W + e$$

such that $\mathrm{E}\left[Y^2\right]$ and $\mathrm{E}\left[W^2\right]$ are finite and different from 0. Moreover, assume that (Y_i, W_i) are iid and $\mathrm{E}\left[eW\right] = 0$. Finally, we'll assume that while b_0 is unknown, it is smaller in absolute value than a huge number M.

Is $\hat{b}_{OLS}$ a consistent estimator of b_0?

Recall that

$$\hat{b}_{OLS} \in \arg\max_b -\sum_{i=1}^{n}(Y_i - bW)^2$$

We define the sum of squared residuals as our $Q_n(b)$ function.

A1. Does $\operatorname{plim} Q_n$ exist? It might not be clear in the form we just defined since increasing n will make the sum of squares larger and larger. However, we could define Q_n to be the average of the sum of squared residuals. Then, from the law of large numbers, we can be sure that Q_n will converge to its expectation:

$$\lim_{n\to\infty} Q_n(b) = Q_0(b) = \mathrm{E}\left[-(Y - bW)^2\right]$$

Now, is Q_0 maximized at b_0?

By the FOC:

$$\begin{aligned}
\frac{\partial Q_0(b)}{\partial b} = 0 &\Leftrightarrow -2\,\mathrm{E}\left[WY\right] + 2b\,\mathrm{E}\left[W^2\right] = 0 \\
&\Leftrightarrow b = \frac{\mathrm{E}\left[WY\right]}{\mathrm{E}\left[W^2\right]} \\
&\Leftrightarrow b = \frac{\mathrm{E}\left[W(bW + e)\right]}{\mathrm{E}\left[W^2\right]} \\
&\Leftrightarrow b = \frac{b_0\,\mathrm{E}\left[W^2\right]}{\mathrm{E}\left[W^2\right]} + \frac{\mathrm{E}\left[We\right]}{\mathrm{E}\left[W^2\right]} \\
&\Leftrightarrow b = b_0
\end{aligned}$$

A2. Since Q_n is a quadratic function, we know for sure that it is smooth.

A3. By assumption $|b_0| < M$, therefore the domain of Q_n is compact.

A4. Finally,

$$\begin{aligned}|\frac{\partial Q_n(b)}{\partial b}| &= |-\frac{1}{n}\sum_{i=1}^{n} 2(Y_i - bW_i)(-W_i)| \\ &\xrightarrow{\text{p}} |\text{E}\left[2(Y_i - bW_i)(-W_i)\right]| \Rightarrow |\frac{\partial Q_n(b)}{\partial b}| = O_p(1)\end{aligned}$$

We can conclude, by theorem 3.8 that $\hat{b}_{OLS}$ is a consistent estimator of b_0.

Theorem 3.9 (Glevenko-Cantelli Theorem)**.** *Let $\{Z_n\}$ be any sequence of iid random variables with cdf $F(\cdot)$. The observed cumulative distribution*

$$\hat{F}(z) = \frac{1}{n}\sum_{i=1}^{n} \text{I}(Z_i \leq z)$$

is a consistent estimator of the true cdf $F(\cdot)$.

Chapter 4

Classical Regression

4.1 Introduction, model and assumptions

Definition 4.1 (Data sample)**.** *A data sample is a set of n observations, denoted by the subscript i. Its elements are random variables (Y_i, X_i) where X_i can be a vector.*

In econometrics, we commonly view these observations as iid draws from a common distribution called the Data-generating process (or DGP).

Consider the a model of a variable Y explained by k regressors with n observations:

$$Y_i = \beta_1 + \beta_2 X_{2i} + ... + \beta_k X_{ki} + e_i$$

which can be written in its matrix form:

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_{21} & X_{31} & \dots & X_{k1} \\ 1 & X_{22} & X_{32} & \dots & X_{k2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{2n} & X_{3n} & \dots & X_{kn} \end{bmatrix} \cdot \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}$$

Definition 4.2 (OLS estimator)**.** *The OLS estimator $\hat{\beta}$ of the true parameter β is*

the vector that minimizes the sum of squared residuals:

$$\begin{aligned}\hat{\beta} &\in \arg\min_{\beta} e'e \\ &\in \arg\min_{\beta} (Y - X\beta)'(Y - X\beta)\end{aligned}$$

Remark 4.1. *If the matrix given by $X'X$ is invertible, then the value of the OLS estimator $\hat{\beta}$ is:*

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

Proof. In order to solve the model, we can simplify the objective function to $Y'Y - Y'\beta X - \beta' X'Y + \beta' X'X\beta$ where each term is a 1×1 matrix. The FOC gives:

$$\begin{aligned}\frac{\partial}{\partial \beta} = 0 &\Leftrightarrow -2X'Y + 2X'X\hat{\beta} = 0 \\ &\Leftrightarrow X'Y = X'X\hat{\beta} \\ &\Leftrightarrow \hat{\beta} = (X'X)^{-1}X'Y\end{aligned}$$

□

Remark 4.2. *In the particular case of $k = 2$ so that $y = a + bx + e$. We have that:*

$$\begin{aligned}\hat{b} &= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\operatorname{Cov}(x, y)}{\operatorname{Var}[x]} \\ \hat{a} &= \bar{y} - \hat{b}\bar{x}\end{aligned}$$

Proof. □

Definition 4.3 (Fitted values and residuals)**.** *The fitted value $\hat{Y}$ is defined by:*

$$\hat{Y} = X\hat{\beta}$$

This variable is not a predictor of Y since it is a function of the sample only. The residuals are $\hat{e} = Y - \hat{Y} = Y - X\hat{\beta}$. They are different from errors e which are unobservable parameters of the regression.

Remark 4.3. *In a linear regression context, we have:*

$$X'\hat{e} = 0$$

This implies that if X contains a column of ones (i.e. there is a constant in the model), then $\frac{1}{n}\sum_{i=1}^{n}\hat{e}_i = 0$.

Proof. We can easily see that:

$$X'\hat{e} = X'(Y - X\hat{\beta}) = X'Y - X'X\hat{\beta} = X'Y - X'X(X'X)^{-1}X'Y = 0$$
$$\Leftrightarrow \begin{bmatrix} 1 & 1 & \dots & 1 \\ X_{21} & X_{22} & \dots & X_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ X_{k1} & X_{k2} & \dots & X_{kn} \end{bmatrix} \cdot \begin{bmatrix} \hat{e}_1 \\ \hat{e}_2 \\ \vdots \\ \hat{e}_n \end{bmatrix} = 0 \Leftrightarrow \begin{bmatrix} \sum_{i=1}^{n}\hat{e}_i \\ \vdots \end{bmatrix} = 0$$

□

Definition 4.4 (Linear estimator)**.** *An estimator is linear in random variables if it can be written as linear transformation of a random vector. In words, it must be equal to a constant matrix multiplied by a random vector:*

$$\tilde{\beta} = \tilde{C}Y$$

Definition 4.5 (Projection matrix)**.** *The matrix $P = X\tilde{C}$ is a projection matrix. Its properties are that:*

- $PX = X$
- $P = P'$
- $PP = P$
- $\mathrm{tr}(P) = k$

Remark 4.4. *The projection matrix P can be used to recover the fitted values with*

$$PY = \hat{Y}$$

Proof. We have $PY = X(X'X)^{-1}X'Y = X\hat{\beta} = \hat{Y}$ □

Definition 4.6 (Orthogonal projection matrix)**.** *Let $M = I_n - P$, M is called the orthogonal projection matrix since $MX = 0$. Other properties of M include:*

- $MP = 0$
- $\text{tr}(M) = n - k$
- $MY = Y - PY = Y - \hat{Y} = \hat{e}$
- $\hat{e} = MY = M(X\beta + e) = Me$

Definition 4.7 (Estimator of the error variance)**.** *The moment estimator of the variance σ^2 would be:*

$$\tilde{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} e_i^2$$

However, e_i is never observed and cannot be used. Let's substitute for $\hat{e}_i$ after OLS estimation. We get the feasible variance estimator:

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} \hat{e}_i^2$$

Alternatively, we can write $\tilde{\sigma}^2 = n^{-1}e'e$ and $\hat{\sigma}^2 = n^{-1}\hat{e}'\hat{e} = n^{-1}Y'MMY = n^{-1}e'MMe = n^{-1}e'Me$. A nice property of this is that:

$$\begin{aligned}\tilde{\sigma}^2 - \hat{\sigma}^2 = n^{-1}e'e - n^{-1}e'Me &= n^{-1}e'(I_n - M)e \\ &= n^{-1}e'Pe \\ &\geq 0\end{aligned}$$

which means that $\tilde{\sigma}^2 \geq \hat{\sigma}^2$.

Definition 4.8 (R^2 and analysis-of-variance)**.** *We can measure the variance of the model with a variable called R^2. Write*

$$Y = PY + MY = \hat{Y} + \hat{e}$$

It follows that

$$Y'Y = \hat{Y}'\hat{Y} + 2\hat{Y}'\hat{e} + \hat{e}'\hat{e} = \hat{Y}'\hat{Y} + \hat{e}'\hat{e}$$

And hence $Y - \bar{Y} = \hat{Y} - \bar{Y} + \hat{e} \Rightarrow (Y - \bar{Y})'(Y - \bar{Y}) = (\hat{Y} - \bar{Y})'(\hat{Y} - \bar{Y}) + 2(\hat{Y} - \bar{Y})'\hat{e} + \hat{e}'\hat{e}$ which gives

$$\text{Var}\left[Y\right] = \text{Var}\left[\hat{Y}\right] + \text{Var}\left[\hat{e}\right]$$

Finally, we define as R^2 the proportion of variation of Y explained by a variation in $\hat{Y}$:

$$R^2 = \frac{\text{Var}\left[\hat{Y}\right]}{\text{Var}\left[Y\right]} = 1 - \frac{\text{Var}\left[\hat{e}\right]}{\text{Var}\left[Y\right]}$$

We have already seen that, in order to get a solution for our OLS estimator we need the assumption of non-singularity of $X'X$. In the same spirit, we will need other assumptions in order to draw out the properties of $\hat{\beta}$ whether in finite or infinite samples. The assumptions that are going to be described here represent the minimal assumptions that one can make ; we'll see what they imply and how to relax them in the following sections.

Definition 4.9 (Classical assumptions)**.** *The following assumptions on our model are called the classical assumptions:*

A1 ***Linearity and correct specification:*** *the model must be correctly specified as linear in parameters (no β^2). In matrix form, the model must be represented by*

$$Y = X\beta + e$$

A2 ***No randomness in X:*** *the data in X is not random (issued by a random variable). It would be the exact same if we took another sample of the population. Mathematically,*

$$\text{E}\left[e_i^2|X\right] = \text{E}\left[e_i^2\right]$$

A3 ***Non-singularity of $X'X$:*** *for it to be non-singular, it must be that $n > k$ (there are more observations than explanatory variables $\Rightarrow$ non over-idenfication) and $\text{rank}(X) = k$ (no multicollinearity in X).*

A4 ***The errors are spherical:*** *in particular, $\text{E}\left[e_i\right] = \text{E}\left[e\right] = 0$ and $\text{Var}\left[e\right] = \Omega = \sigma^2 I_n$. This property also means that there is no heteroskedasticity nor autocorrelation in the data.*

4.2 Finite sample properties of OLS estimation

Thanks to these four assumptions, we will be able to discuss more in depth the properties of our OLS estimator, first in finite samples.

Remark 4.5. *Under the assumptions A1-A4, the OLS estimator $\hat{\beta}$ is unbiased.*

Proof. We already know that $\hat{\beta} = (X'X)^{-1}X'Y$. Therefore,

$$\begin{aligned}
\mathrm{E}\left[\hat{\beta}\right] = \mathrm{E}\left[(X'X)^{-1}X'Y\right] &= \mathrm{E}\left[(X'X)^{-1}X'(X\beta + e)\right] \\
&= \mathrm{E}\left[(X'X)^{-1}X'X\beta + (X'X)^{-1}X'e\right] \\
&= \mathrm{E}\left[\beta + (X'X)^{-1}X'e\right] \\
&= \beta + (X'X)^{-1}X'\,\mathrm{E}\left[e\right] \\
&= \beta
\end{aligned}$$

□

Remark 4.6. *A linear estimator is unbiased if and only if its associated transformation matrix $\tilde{C}$ is such that*

$$\tilde{C}X = I_k$$

Proof. In the case of linear regression, we'd have

$$\begin{aligned}
\tilde{\beta} = \tilde{C}(X\beta + e) = \tilde{C}X\beta + \tilde{C}e \Rightarrow \mathrm{E}\left[\tilde{\beta}\right] &= \tilde{C}X\beta + \tilde{C}\,\mathrm{E}\left[e\right] \\
&= \tilde{C}X\beta = \beta \text{ if } \tilde{C}X = I_k
\end{aligned}$$

This is verified in the OLS since $\tilde{C}X = (X'X)^{-1}X'X = I_k$ □

Remark 4.7. *The variance of a homoskedastic, non-autocorrelated linear estimator is given by* $\mathrm{Var}\left[\tilde{\beta}\right] = \sigma^2(\tilde{C}\tilde{C}')$. *In the particular case of the OLS estimator, we have* $\mathrm{Var}\left[\hat{\beta}_{OLS}\right] = \sigma^2(X'X)^{-1}$

Proof. The proof is trivial and follows the properties of the variance.

$$\begin{aligned}
\mathrm{Var}\left[\tilde{\beta}\right] = \mathrm{Var}\left[\tilde{C}X\beta + \tilde{C}e\right] = \mathrm{Var}\left[\tilde{C}e\right] &= \tilde{C}\,\mathrm{Var}\left[e\right]\tilde{C}' \\
&= \tilde{C}\Omega\tilde{C}' \\
&= \tilde{C}\sigma^2 I_n\tilde{C}' \\
&= \sigma^2(\tilde{C}\tilde{C}')
\end{aligned}$$

In the case of $\hat{\beta}_{OLS}$, we have $\sigma^2(\tilde{C}\tilde{C}') = \sigma^2((X'X)^{-1}X'((X'X)^{-1}X')') = \sigma^2(X'X)^{-1}$ □

Remark 4.8. *The OLS estimator $\hat{\beta}_{OLS}$ is BLUE: the Best Linear Unbiased Estimator. This property means that, among linear unbiased estimators, the OLS estimator is the most efficient one.*

Proof. □

Practical considerations on the data

Assume the model is

$$y_i = a + bx_i + e_i$$

Then, we can show that

- $\text{Var}\left[\hat{b}\right] =$
- $\text{Var}\left[\hat{a}\right] =$
- $\text{Cov}\left(\hat{b}, \hat{a}\right) =$

Analyzing the data we can find some interesting properties for our model.

For example, if σ^2 is small, all three variances and covariance will be small as well. A lower σ implies a more efficient model.

Now, if n is big, the effect is the same, since all variances will be smaller, our model will be more accurate.

Again, the implications are the same with greater values of $x_i - \bar{x}$.

Finally, we can see that the covariance between the two estimators indicate how their errors are related. If the covariance is high and positive, then a mistake in the estimation of $\hat{b}$ will lead to the same mistakes in $\hat{a}$.

Now we have seen that $\text{Var}\left[\hat{\beta}\right]$ is given by $\sigma^2(X'X)^{-1}$. However, in actual situations, the true value σ is unknown and we would need to estimate it.

Remark 4.9 (Properties of the variance estimator)**.** *Let $\hat{\sigma}^2$ be the sample moment estimator. This estimator is biased and:*

$$\text{E}\left[\hat{\sigma}^2\right] = \sigma^2\left(\frac{n-k}{n}\right)$$

Proof.

$$\mathrm{E}\left[\hat{\sigma}^2\right] = \mathrm{E}\left[n^{-1}e'Me\right] = n^{-1}\,\mathrm{E}\left[\mathrm{tr}(Mee')\right] = n^{-1}\,\mathrm{tr}(M\,\mathrm{E}\left[ee'\right]) = n^{-1}\,\mathrm{tr}(M\Omega)$$
$$= n^{-1}\sigma^2(n-k)$$

□

Definition 4.10 (Adjusted sample variance)**.** *We define s^2 to be the adjusted sample estimator of the variance, in short the adjusted sample variance, such that:*

$$s^2 = \frac{\hat{e}'\hat{e}}{n-k} = \frac{(Y - X\hat{\beta})'(Y - X\hat{\beta})}{n-k}$$

This implies that this time we have: $\mathrm{E}\left[s^2\right] = \sigma^2$. Hence, we can use this estimator to estimate the variance of our OLS estimator $\hat{\beta}$:

$$\widehat{\mathrm{Var}\left[\hat{\beta}\right]} = s^2(X'X)^{-1}$$

Each parameter $\hat{\beta}_k$'s variance would be the (k,k)th element of the matrix.

Again, we find ourselves with more information about $\hat{\beta}$, namely its mean and variance, but not enough information to get the whole distribution of $\hat{\beta}$. We know that $\hat{\beta} = \beta + (X'X)^{-1}X'e$ where the distribution e is the only unknown. We will need a new assumption.

Definition 4.11 (Normality of the error term)**.** *Assuming all classical assumptions hold. We add the assumption (**A5**) that the error term e_i follows a normal distribution of mean 0 and variance $\sigma^2 I_n$.*

Following this assumption, we now know that $Y \sim N(X'\beta, \sigma^2 I_n)$ and $\hat{\beta} \sim N(\beta, \sigma^2(X'X)^{-1})$.

Remark 4.10. *Let V_j denote the (j,j)th element of the matrix $(X'X)^{-1}$. Then, $\hat{\beta}_j \sim N(\beta_j, \sigma^2 V_j)$ where σ^2 can be estimated with s^2.*

Therefore,

$$\frac{\hat{\beta}_j - \beta_j}{\sqrt{\sigma^2 V_j}} \sim N(0,1)$$

while,

$$\frac{\hat{\beta}_j - \beta_j}{\sqrt{s^2 V_j}} \sim t_{n-k}$$

This last fact can be used in interval estimation as it implies that:

$$\Pr(\hat{\beta}_j - t_{\alpha/2}\frac{S}{\sqrt{V_j}}$$

Remark 4.11 (Moments of the residuals)**.** *Let the residuals of the regression be* $\hat{e} = Me$ *as we've seen before. We have that:*

- $\mathrm{E}[\hat{e}] = 0$
- $\mathrm{Var}[\hat{e}] = M\sigma^2$

Proof. We have that: $\mathrm{E}[\hat{e}] = \mathrm{E}[Me] = M\,\mathrm{E}[e] = 0$. And $\mathrm{Var}[\hat{e}] = \mathrm{Var}[Me] = M\,\mathrm{Var}[e]\,M' = M\Omega M = M\sigma^2 I_n M = \sigma^2 MM = \sigma^2$ □

4.3 Asymptotic properties of OLS estimation

Remark 4.12 (Consistency of $\hat{\beta}_{\mathrm{OLS}}$)**.** *Let* $Q_n = \frac{X'X}{n}$*, which is a non-singular, positive definite matrix (from A3). Moreover, let its limit* $Q = \lim_{n\to\infty} Q_n$ *exist.*

Consider that $\mathrm{E}\left[\hat{\beta}\right] = \beta$ *and* $\mathrm{Var}\left[\hat{\beta}\right] = \frac{\sigma^2}{n}Q_n^{-1}$*. Then,*

- $\lim_{n\to\infty} \mathrm{E}\left[\hat{\beta}\right] = \beta$
- $\lim_{n\to\infty} \mathrm{Var}\left[\hat{\beta}\right] = \lim_{n\to\infty} \frac{\sigma^2}{n}Q_n^{-1} = 0$

and therefore, $\hat{\beta} \xrightarrow{ms} \beta$*.*

Remark 4.13 (Root-n-consistency of $\hat{\beta}_{\mathrm{OLS}}$)**.**

Chapter 5

Specification issues

5.1 Non-randomness of X

Starting with the usual model:

$$Y = X\beta + e$$

We assume that:

- (y_i, x_i) are independent but not identically distributed.
- $\mathrm{E}[e_i x_i] = 0$, which, if X contains a constant, implies that $\mathrm{E}[e] = 0$.
- For all $i, j : i \neq j$, $\mathrm{E}[e_i e_j] = 0$ so that off-diagonal elements of Ω are zero.
- $\mathrm{E}[\sigma^2 | x_i] = \sigma^2(x_i)$

The assumption that $\mathrm{E}[e_i | X] = 0$ is not made here, implying that X is now a random variable. The implication of this statement can be visible from the new mean of $\hat{\beta}_{OLS}$:

$$\begin{aligned}
\mathrm{E}\left[\hat{\beta}\right] = \mathrm{E}\left[(X'X)^{-1}X'Y\right] &= \mathrm{E}\left[(X'X)^{-1}X'(X\beta + e)\right] \\
&= \mathrm{E}\left[(X'X)^{-1}X'X\beta + (X'X)^{-1}X'e\right] \\
&= \beta + \mathrm{E}\left[(X'X)^{-1}X'e\right]
\end{aligned}$$

Using our definition of $Q_n = \frac{X'X}{n}$, we can write:

$$\mathrm{E}\left[\hat{\beta}\right] = \beta + \mathrm{E}\left[Q_n^{-1}\frac{X'e}{n}\right]$$

Note that, even if $\mathrm{E}\left[e_i X_i\right] = 0$ we cannot cancel out the expectation term since it might be correlated to Q_n^{-1}.

The same issue arises for $\mathrm{Var}\left[\hat{\beta}\right]$:

$$\begin{aligned}\mathrm{Var}\left[\hat{\beta}\right] = \mathrm{Var}\left[(X'X)^{-1}X'e\right] &= \mathrm{E}\left[(X'X)^{-1}X'ee'X(X'X)^{-1}\right] \\ &= \mathrm{E}\left[Q_n^{-1}\frac{(X'e)(X'e)'}{n^2}Q_n^{-1}\right]\end{aligned}$$

We now want to check if $\hat{\beta}$ is consistent. We have:

$$\mathrm{plim}\,\hat{\beta} = \mathrm{plim}\,\beta + \mathrm{plim}\left[Q_n^{-1}\frac{X'e}{n}\right] = \beta + \mathrm{plim}\,Q_n^{-1} + \mathrm{plim}\,\frac{X'e}{n}$$

If $\mathrm{Var}\left[\frac{X'e}{n}\right] \to 0$, we have that $\mathrm{plim}\,\frac{X'e}{n} = \frac{1}{n}\mathrm{E}\left[X'e\right] = 0$ by assumption 2.

Note that the last part allows us to write:

$$\sqrt{n}(\hat{\beta} - \beta) = Q_n^{-1}\sqrt{n}\frac{X'e}{n} \xrightarrow{\mathrm{d}} N(0, \mathrm{Var}\left[Q_n^{-1}\sqrt{n}\frac{X'e}{n}\right])$$

Since Q_n^{-1} is a constant, the problem reduces to finding $\mathrm{Var}\left[\sqrt{n}\frac{X'e}{n}\right]$:

$$\mathrm{Var}\left[\sqrt{n}\frac{X'e}{n}\right] = \frac{1}{n}\mathrm{E}\left[(X'e)(e'X)\right] =$$

5.2 Non-stationarity of X

5.3 High correlation in the error term

5.4 Collinearity

Definition 5.1 (Strict multicollinearity)**.** *Strict multicollinearity is a consequence of the columns of matrix X being linearly dependent. In particular, there is at least one column (or row) of X which is a linear combination of any other column (row). Algebraically,*

$$\exists \alpha \neq 0 : X\alpha = 0$$

Remark 5.1 (Singularity of strictly multicollinear matrices)**.** *If the matrix X is strictly collinear, then its quadratic form $X'X$ is singular and $\hat{\beta}_{OLS}$ is not defined.*

Definition 5.2 (Near multicollinearity)**.** *A matrix X is said to be near multicollinearity (or simply multicollinear) if the matrix $X'X$ is near singular.*

The issue with near multicollinearity resides in the definition of what is "near" or in other words, what is "collinear enough"? We can work out a few examples to check for this problem.

Multicollinearity in examples

> Let x be the average hourly wage and z the average daily wage. Then, it could be that x and z are strictly multicollinear if everyone in the population worked 8 hours exactly ($z = 8x$). In practice, the number of hours worked per day may vary slightly but the correlation between x and y will be very close to 1, leading to near multicollinearity.
>
> Let h be the number of hours worked in a week and w be the total weekly wage. We have that $w = xh$ so x and w are not strictly multicollinear. However, in logs, $\ln(w) = \ln(xh) = \ln(x) + \ln(h)$ implying that $\ln(w)$ and $\ln(x)$ are strictly multicollinear.

Finally, if we use both x and x^2 in a regression, we increase chances of finding near multicollinearity.

5.5 Coefficient interpretation

5.5.1 Linear vs. log specification

Let us compare two different specifications:

$$Y = a + bX + e \text{ and } \ln(Y) = \alpha + \beta \ln(X) + \varepsilon$$

We know that coefficients should be interpreted as the derivative of the regressed term with respect to the regressor. In this case,

- $b = \frac{dY}{dX}$ is the derivative of Y w.r.t. X.
- $\beta = \frac{d\ln(Y)}{d\ln(X)} = \frac{dY}{dX}\frac{X}{Y}$ is the elasticity of Y w.r.t. X.

However, whether you want to estimate an elasticity or a derivative should not affect what model you should use. One should only care about the true specification of a model, then make the computations necessary to find a certain variable.

5.5.2 Measurement units

Now, consider two models

$$Y = a + bX + e \text{ and } Y = a^* + b^*X^* + e^*$$

where X is measured in thousands of dollars while X^* is directly measured in dollars. We have $X^* = 1000X$. Notice that we can rewrite the second model as:

$$Y = a^* + b^* \cdot (1000X) + e^*$$

Therefore it must be that $a^* = a$, $b = 1000b^*$ and $e = e^*$. This also implies that each t-statistic will be the exact same. Hence, a change of unit in a linear model does not change the fit of the model.

If the change of units happens on a logarithmic model, then the result above is different. In particular,

$$\begin{aligned}\ln(Y) &= \alpha^* + \beta^* \ln(1000X) + e^* \\ &= \underbrace{\alpha^* + \beta^* \ln(1000)}_{\text{new constant}} + \beta^* \ln(X) + e^*\end{aligned}$$

Here, the constant term will change (and its t-statistic too).

5.5.3 Percent change

One should always use a log specification for a percent change variable ($\ln(\frac{X_t}{X_{t-1}})$) instead of computing the actual period percent change ($\frac{X_t - X_{t-1}}{X_{t-1}}$).

5.5.4 Interaction variables

Consider the following model,

$$Y_i = \beta_1 + \beta_2 X_i + \beta_3 Z_i + \beta_4 \underbrace{X_i Z_i}_{\textit{interaction term}} + e_i$$

This specification allows for variables to interact with each other so that $\frac{\partial Y}{\partial X} = \beta_2 + \beta_4 Z$ and $\frac{\partial Y}{\partial Z} = \beta_3 + \beta_4 X$. This means that the effect of X (or Z) on Y also depends on the value that Z (or X) takes. This model is close to the analysis performed in a diff-in-diff model since having this specification almost implies having two models to estimate.

A similar model would be one including a polynomial function of one variable such as,

$$Y_i = \beta_1 + \beta_2 X_i + \beta_3 X_i^2 + \beta_4 X_i^3 + e_i$$

Both these models do not violate any assumptions among the Gauss-Markov assumptions. However, one should consider the fact that interacting variables increase the likelihood of multicollinearity in the variables (since there will be a strong correlation between single and interacted variables).

Predicting sales revenue at CVS

Chapter 6

Maximum Likelihood Estimation

Estimating the probability of a coin flip

Let a coin be flipped a hundred times, with probability p of falling on Heads (H) and $(1-p)$ of falling on Tail (T).

Consider any outcome of this experiment, what can we say about $\hat{p}$?

- If all 100 coins are H? Probably $\hat{p} = 1$.
- If only 99 coins are Heads? Probably $\hat{p} = 0.99$.

But how can we use what we know of the distribution of these outcomes to help us estimate p?

The likelihood of the experiment giving the outcome that 100 H have occurred is p^{100}. What is the value of p that maximizes this probability?
$\Rightarrow \hat{p} = 1$

The likelihood of the experiment giving the outcome that 99 H have occurred is $100p^{99}(1-p)$. What is the value of p that maximizes this probability?

$$\Rightarrow \frac{\partial \mathcal{L}}{\partial p} = 0 \Leftrightarrow 99 \cdot 100 \cdot \hat{p}^{98}(1-\hat{p}) - 100\hat{p}^{99} = 0 \Leftrightarrow 99\hat{p}^{98} = 100\hat{p}^{99}$$
$$\Leftrightarrow \hat{p} = 0.99$$

This method is called Maximum Likelihood Estimation.

6.1 Basic assumptions

We have seen that for a sequence of random variables $Z_1, ..., Z_n$, the joint pdf can be written as $f(Z_1, ..., Z_2|\theta)$ where θ is the vector of parameters that define the joint distribution.

Definition 6.1 (Likelihood function). *Let $\{Z_n\}$ be any sequence of random variables following a joint distribution $f(Z_1, ..., Z_n|\theta)$. The likelihood function is the equivalent of the joint pdf expressed in terms of the parameters θ. We write it as $L(\theta|Z_1, ..., Z_n)$.*

When $Z_1, ..., Z_n$ are iid,

$$L(\theta|Z_1, ..., Z_n) = \prod_{i=1}^{n} f(Z_i|\theta)$$

Definition 6.2 (Maximum Likelihood estimator). *Let $\{Z_n\}$ be any sequence of random variables following a joint distribution $f(Z_1, ..., Z_n|\theta)$. The maximum likelihood estimator of θ is the argument that maximizes the likelihood function $L(\theta|Z_1, ..., Z_n)$.*

$$\hat{\theta}_{ML} = \arg\max_{\theta \in \Theta} L(\theta|Z_1, ..., Z_n)$$

where Θ is the set of all possible values of θ.

Definition 6.3 (Assumptions on MLE). *In order to further analyze the MLE, let's describe a set of additional assumptions:*

- ***A1. Compactness:*** *Let Θ be the set of all possible parameters. We will assume that this set is compact and θ_0, the true value of the parameter lies in this set.*
- ***A2. Name?:*** *For all $\theta \in \Theta$ such that $\theta \neq \theta_0$, we have that,*

$$\mathrm{E}\left[\frac{\partial \ln f(Z_i|\theta)}{\partial \theta}\right] \neq \mathrm{E}\left[\frac{\partial \ln f(Z_i|\theta_0)}{\partial \theta_0}\right]$$

- ***A3. Boundedness:*** *All first-order, second-order and third-order (own and cross) derivatives of $\ln f(Z_i|\theta)$ with respect to θ exist and are bounded.*
- ***A4. Name?:*** *Let Ω_Z be the support of $f(\cdot|\theta)$; either Ω_Z does not depend on θ or $f(Z_i|\theta) = 0$ for all θ on the boundary of Θ.*
- ***A5. Name?:*** *All Z_i are random variables once conditioned on θ.*

6.2 Some properties

Remark 6.1 (Log-likelihood function). *Let $\hat{\theta}_{ML}$ be the MLE for the parameter θ_0 from the distribution $f(Z_1, ..., Z_n|\theta)$. Then, $\hat{\theta}_{ML}$ also solves the logarithm of the likelihood function:*

$$\begin{aligned}\hat{\theta}_{ML} &= \arg\max_{\theta\in\Theta} L(\theta|Z_1, ..., Z_n) \\ &= \arg\max_{\theta\in\Theta} \ln\big(L(\theta|Z_1, ..., Z_n)\big) \\ &= \arg\max_{\theta\in\Theta} \ln\big(\prod_{i=1}^{n} f(Z_i|\theta)\big) \\ &= \arg\max_{\theta\in\Theta} \sum_{i=1}^{n} \ln f(Z_i|\theta)\end{aligned}$$

Definition 6.4 (Score function). *The score function, denoted $s(Z|\theta)$ is defined as the gradient of the log-likelihood function when differentiated wrt θ:*

$$s(Z|\theta) = \frac{\partial \ln f(Z|\theta)}{\partial\theta}$$

Because Z_i are iid, $s_i(Z_i|\theta)$ are also iid.

Remark 6.2 (Maximum of the score function). *Let $f(Z_1, ..., Z_n)$ be the joint pdf of iid random variables $Z_1, ..., Z_n$ such that θ_0 is the true parameter. Then, $\mathrm{E}[s(Z|\theta_0)] = 0$. This fact is very important because, linked to assumption 2 above, it means that the log-likelihood function is maximized at one unique point θ_0.*

Proof. We know that for any θ, $\int_{\Omega_Z} f(Z|\theta)dZ = 1$. By Leibniz rule, we can

differentiate and get:

$$\begin{aligned}
\frac{\partial \int_{\Omega_Z} f(Z|\theta)dZ}{\partial\theta} &= 0 \\
\int_{\Omega_Z} \frac{\partial f(Z|\theta)dZ}{\partial\theta} &= 0 \\
\int_{\Omega_Z} \frac{\partial \ln f(Z|\theta)dZ}{\partial\theta} f(Z|\theta)dZ &= 0 \\
\int_{\Omega_Z} s(Z|\theta) f(Z|\theta)dZ &= 0 \\
\mathrm{E}\left[s(Z|\theta)\right] &= 0
\end{aligned}$$

Hence, in particular for θ_0, $\mathrm{E}\left[s(Z|\theta_0)\right] = 0$. □

Definition 6.5. *The Hessian matrix of the LL function, denoted as $H(Z|\theta)$ is the derivative of the score function or equivalently, the second-order derivative of the LL function.*

$$H(Z|\theta) = \frac{\partial^2 \ln f(Z|\theta)}{\partial\theta\partial\theta'} = \frac{\partial s(Z|\theta)}{\partial\theta'}$$

Remark 6.3 (Variance of the score function)**.** *Let $f(Z_1, ..., Z_n)$ be the joint pdf of iid random variables $Z_1, ..., Z_n$ such that θ_0 is the true parameter. Then,*

$$\mathrm{Var}\left[s(Z|\theta_0)\right] = -\mathrm{E}\left[H(Z|\theta)\right]$$

Proof. □

Definition 6.6 (Information matrix)**.** *The information matrix is the opposite of the Hessian matrix, it can be put in relation to the log-likelihood function of the sequence of rvs as:*

$$-\mathrm{E}\left[\frac{\partial^2 \ln f(Z_1, ..., Z_n|\theta)}{\partial\theta\partial\theta'}\right] = -n\,\mathrm{E}\left[H(Z|\theta)\right] = -nI(Z|\theta)$$

Theorem 6.1 (Consistency of the ML estimator)**.** *Let $\{Z_n\}$ be any sequence of random variables following a joint distribution $f(Z_1, ..., Z_n|\theta)$. Under the assumptions of the MLE, $\hat{\theta}_{ML}$, is a consistent estimator of θ.*

Proof. □

Theorem 6.2 (Asymptotic normality of the ML estimator)**.** *Suppose θ_{ML} is a consistent estimator of a parameter θ, following the previous theorem. Then,*

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, J^{-1}JJ^{-1})$$

where $J = -H$.

Proof. □

6.3 Application of ML estimation to Binary Choice models

6.3.1 Classical BC model

Let Y_i be a binary variable. The data set is (Y_i, X_i) such that Y_i is independent of X_i. We write the true model as:

$$\Pr[Y_i = 1|X] = F(X_i, \beta)$$

From this model, we get:

$$\mathrm{E}[Y_i|X] = \Pr[Y_i = 1|X] \cdot 1 + \Pr[Y_i = 0|X] \cdot 0 = F(X_i, \beta)$$

Assuming Y_i are iid, we can get the likelihood function of the data as:

$$\begin{aligned} L = \Pr[Y_1, ..., Y_n|X, \beta] &= \prod_{i=1}^{n} \Pr[Y_i = 1|X_i, \beta]^{Y_i} \Pr[Y_i = 0|X_i, \beta]^{1-Y_i} \\ &= \prod_{i=1}^{n} F(X_i, \beta)^{Y_i}(1 - F(X_i, \beta))^{1-Y_i} \end{aligned}$$

in log-likelihood form:

$$\ln L = \sum_{i=1}^{n} (Y_i \ln(F(X_i, \beta)) + (1 - Y_i) \ln(1 - F(X_i, \beta)))$$

Its maximum for β is: $s(X_i, \beta) = 0 \Leftrightarrow \frac{\partial \ln f(Y_i|\beta)}{\partial \beta} = 0 \Leftrightarrow \left[\frac{Y_i}{F(X_i,\beta)} - \frac{(1-Y_i)}{1-F(X_i,\beta)}\right] \frac{\partial F(X_i,\beta)}{\partial \beta} = 0$

We can also compute the information matrix

$$
\begin{aligned}
J_0 &= \mathrm{E}\left[s(X|\beta_0)s(X|\beta_0)'\right] \\
&= \mathrm{E}\left[\left[\frac{Y_i}{F(X_i,\beta)} - \frac{(1-Y_i)}{1-F(X_i,\beta)}\right] \frac{\partial F(X_i,\beta)}{\partial \beta} \frac{\partial F(X_i,\beta)}{\partial \beta'} \left[\frac{Y_i}{F(X_i,\beta)} - \frac{(1-Y_i)}{1-F(X_i,\beta)}\right]'\right] \\
&= \mathrm{E}\left[\left[\left(\frac{Y_i}{F(X_i,\beta)}\right)^2 - 2\frac{Y_i}{F(X_i,\beta)}\frac{(1-Y_i)}{1-F(X_i,\beta)} + \left(\frac{(1-Y_i)}{1-F(X_i,\beta)}\right)^2\right] \frac{\partial F(X_i,\beta)}{\partial \beta} \frac{\partial F(X_i,\beta)}{\partial \beta'}\right] \\
&= \mathrm{E}\left[\left[\frac{Y_i}{F(X_i,\beta)^2} + \frac{(1-Y_i)}{1-F(X_i,\beta)^2}\right] \frac{\partial F(X_i,\beta)}{\partial \beta} \frac{\partial F(X_i,\beta)}{\partial \beta'}\right]
\end{aligned}
$$

6.3.2 Threshold model

Chapter 7

Inference and Hypothesis tests

7.1 Recap

In the case of a linear regression model with normal errors $e_i \sim N(0, \sigma^2)$, it is possible to compute the exact distribution of OLS coefficients $\hat{\beta}_{OLS}$ and OLS residuals $\hat{e}_i$.

First, recall that $\hat{\beta} - \beta = (X'X)^{-1}X'e$, a linear projection of the error e. Hence, we can get:

$$\begin{aligned}\hat{\beta} - \beta &\sim (X'X)^{-1}X'N(0, \sigma^2 I_n) \\ &\sim N(0, \sigma^2(X'X)^{-1}X'X(X'X)^{-1}) \\ &\sim N(0, \sigma^2(X'X)^{-1})\end{aligned}$$

Second, with help of $\hat{e} = Me$, we have that

$$\hat{e} \sim N(0, \sigma^2 MM) \sim N(0, \sigma^2 M)$$

The first two points lead to a result on the joint distribution of $\hat{\beta}$ and $\hat{e}$:

$$\begin{bmatrix}\hat{\beta} - \beta \\ \hat{e}\end{bmatrix} = \begin{bmatrix}(X'X)^{-1}X'e \\ Me\end{bmatrix} = \begin{bmatrix}(X'X)^{-1}X' \\ M\end{bmatrix} e$$

Again, this is a linear projection of e, we can guess its mean (0) and covariance matrix. The covariance is 0 since $(X'X)^{-1}X'M = 0$.

Finally, consider the adjusted sample variance estimator s^2. We have that $(n-k)s^2 = \hat{e}'\hat{e} = e'Me$. We can use the spectral decomposition of M, namely $H\Lambda H'$ where $H'H = I_n$ and Λ is a diagonal matrix with the first $n-k$ terms equal to 1, the rest to 0.

Let $u = H'e \sim N(0, I_n\sigma^2)$ and partition it as $u = (u_1, u_2)$. Then,

$$\begin{aligned}(n-k)s^2 = e'Me = e'H\Lambda H'e &= u'\Lambda u \\ &= u_1'u_1 \\ &\sim \sigma^2\chi^2_{n-k}\end{aligned}$$

The main results are:

- $\hat{\beta} \sim N(\beta, \sigma^2(X'X)^{-1}$, and in particular $\hat{\beta}_j \sim N(\beta_j, \sigma^2\left[(X'X)^{-1}\right]_{jj})$.
- $\hat{e} \sim N(0, \sigma^2 M)$
- $\hat{\beta}$ and $\hat{e}$ are independent
- $\frac{(n-k)s^2}{\sigma^2} \sim \chi^2_{n-k}$
- $\hat{\beta}$ and s^2 are independent

7.2 T-statistic

We can use all results of the last section to derive two data statistics.

Definition 7.1 (Standardized statistic)**.** *Define the standardized statistic as:*

$$\frac{\hat{\beta}_j - \beta_j}{\sqrt{\sigma^2\left[(X'X)^{-1}\right]_{jj}}} \sim N(0,1)$$

The issue with this last statistic is that σ^2 is unknown. If we use s^2, the adjusted variance estimator, we can design a more useful statistic (that will be used for hypothesis testing).

Definition 7.2 (T-statistic)**.** *Define the T-statistic as:*

$$\frac{\hat{\beta}_j - \beta_j}{\sqrt{s^2 \left[(X'X)^{-1}\right]_{jj}}} = \frac{\hat{\beta}_j - \beta_j}{s(\hat{\beta}_j)} \sim t_{n-k}$$

where $s(\hat{\beta}_j)$ is the square root of the $j \times j$-th element of the adjusted variance matrix, and t_{n-k} represents the Student's t-distribution of $(n-k)$ degrees of freedom.

Consider a classical linear regression where e is assumed to follow a normal distribution $N(0, \sigma^2)$. Using Student's t-statistic, we can design a test to assess whether the estimated coefficient $\hat{\beta}$ is equal to a specific value β (we are interested in β_0, the true value of the regression).

Proposition 7.1 (Student's t-test)**.** *Define the null hypothesis as $H_0 : \hat{\beta} = \beta$ while the alternative hypothesis will be $H_1 : \hat{\beta} \neq \beta$.*

The statistic used to test H_0 against H_1 is the absolute value of Student's t-statistic:

$$|T| = \left| \frac{\hat{\beta}_j - \beta_j}{s(\hat{\beta}_j)} \right|$$

We reject H_0 if $|T| > c$.

We call c the critical value of the test. We have seen that it is defined as the threshold for the test but its value is in fact determined to control the probability of type-I error. For a given value of c, the probability of type-I is:

$$\begin{aligned}
\Pr\left[\text{Reject } H_0 | H_0 \text{ is true}\right] &= \Pr\left[|T| > c | H_0\right] \\
&= \Pr\left[T > c | H_0\right] + \Pr\left[T < c | H_0\right] \\
&= 1 - t_{n-k}(c) + t_{n-k}(-c) \\
&= 2(1 - t_{n-k}(c))
\end{aligned}$$

We call this probability α, the significance level of the test and hence we choose c such that $t_{n-k}(c) = 1 - \alpha/2$.

7.3 Confidence intervals

We have seen $\hat{\beta}$ as a point estimate for the true parameter β. We could also consider a set of values that have a certain probability of containing the true value β.

Definition 7.3 (Interval estimate)**.** *An interval estimate $\hat{C}$ is a set $\left[\hat{L}, \hat{U}\right]$ which goal is to contain the true value of the parameter β.*

Definition 7.4 (Coverage probability)**.** *The coverage probability is defined as* $\Pr\left[\beta \in \hat{C}\right] = 1 - \alpha$

Proposition 7.2 (Normal regression confidence interval)**.** *Consider the interval based on Student's t-statistic defined as the set of values β such that the t-statistic is smaller than c, the critical value of the associated t-test. Formally,*

$$\hat{C} = \{x : |T(x)| \leq c\} = \left\{x : -c \leq \frac{\hat{\beta} - x}{s(\hat{\beta})} \leq c\right\}$$

7.4 Wald tests

We know that $\hat{\beta}$ is asymptotically normal around β. In particular, if we want to test the null hypothesis $H_0 : A\beta - C = 0$, we can use:

$$A\hat{\beta} - C \overset{a}{\sim} N(0, \Omega)$$

7.4.1 Linear Regression model

This also means that:

$$\left(A\hat{\beta}-C\right)' \text{Var}\left[A\hat{\beta}-C\right]^{-1}\left(A\hat{\beta}-C\right) \sim \chi^2_q$$
$$\left(A\hat{\beta}-C\right)' \left(A\,\text{Var}\left[\hat{\beta}\right]A'\right)^{-1}\left(A\hat{\beta}-C\right) \sim \chi^2_q$$
$$\left(A\hat{\beta}-C\right)' \left(A\sigma^2(X'X)^{-1}A'\right)^{-1}\left(A\hat{\beta}-C\right) \sim \chi^2_q$$
$$\frac{\left(A\hat{\beta}-C\right)' \left(A(X'X)^{-1}A'\right)^{-1}\left(A\hat{\beta}-C\right)}{\sigma^2} \sim \chi^2_q$$

However, σ^2 is unknown so we have to use the adjusted sample variance s^2:

$$\frac{\left[\left(A\hat{\beta}-C\right)' \left(A(X'X)^{-1}A'\right)^{-1}\left(A\hat{\beta}-C\right)\right]/q}{\sigma^2\left[\frac{(n-k)s^2}{\sigma^2}\right]/(n-k)} \sim F_{q,n-k}$$

7.4.2 General Case

7.5 Likelihood Ratio tests

The section on t-tests introduced how to assess the validity of a hypothesis on one single estimated coefficient. However, it may be useful to test the validity of a set of estimated coefficients at once. For that purpose, you might already know the popular F-test which can be derived from the Likelihood Ratio (LR) test discussed in this section.

Consider a partition of the regressor X as $X = (X_1, X_2)$ and in a similar way the partition of $\beta = (\beta_1, \beta_2)$. The new regression model is:

$$Y = X_1\beta_1 + X_2\beta_2 + e$$

Suppose we want to test the significance of the set of parameters β_2, define the null hypothesis as $H_0 : \beta_2 = 0$.

If H_0 is true, then the "restricted" model is $Y = X_1\beta_1 + e$. Under the alternative hypothesis $H_1 : \beta_2 \neq 0$, we keep our "unrestricted" model.

Proposition 7.3 (Likelihood Ratio test)**.** *The statistic used to test the validity of H_0 against H_1 under the LR test is:*

$$LR = -2 \ln \frac{L(\hat{\beta}_1)}{L(\hat{\beta})} \sim \chi^2_q$$

where $L(\cdot)$ is the value of the likelihood function and q is the number of linear restrictions.

7.6 Lagrange Multiplier tests

Chapter 8

Generalized Least-Squares and non-iid errors

8.1 Heteroskedasticity

Heteroskedasticity is the phenomenon when error terms e_i do not have the same variance for all i. Formally, we write $\mathrm{E}\left[e_i e_i'\right] = \sigma^2\Omega$ where Ω is a diagonal matrix different from the identity matrix and by normalization $\mathrm{tr}(\Omega) = n$.

In this particular case, our typical model for $Y = X\beta + e$ does not satisfy all Gauss-Markov assumptions. The next question we should ask ourselves is how does this violation affects our OLS estimates. Note that we still have: $\hat{\beta}_{OLS} = \beta + (X'X)^{-1}X'e$. Hence,

$$\mathrm{E}\left[\hat{\beta}_{OLS}\right] = \beta + (X'X)^{-1}X'\,\mathrm{E}\left[e\right] = \beta$$

Our OLS estimator is still unbiased, is it consistent?

$$\begin{aligned}
\lim_{n\to\infty} \operatorname{Var}\left[\hat{\beta}_{OLS}\right] = \lim_{n\to\infty} \mathrm{E}\left[(\hat{\beta}-\beta)(\hat{\beta}-\beta)'\right] &= \lim_{n\to\infty} \mathrm{E}\left[((X'X)^{-1}X'e)((X'X)^{-1}X'e)'\right] \\
&= \lim_{n\to\infty} \mathrm{E}\left[(X'X)^{-1}X'ee'X(X'X)^{-1}\right] \\
&= \lim_{n\to\infty} (X'X)^{-1}X'\,\mathrm{E}\left[ee'\right]X(X'X)^{-1} \\
&= \lim_{n\to\infty} \frac{1}{n^2}\left(\frac{X'X}{n}\right)^{-1} X'\sigma^2\Omega X\left(\frac{X'X}{n}\right)^{-1} \\
&= \lim_{n\to\infty} \frac{\sigma^2}{n}\left(\frac{X'X}{n}\right)^{-1} \frac{X'\Omega X}{n}\left(\frac{X'X}{n}\right)^{-1} \\
&= \lim_{n\to\infty} \frac{\sigma^2}{n} Q_n^{-1} R_n Q_n^{-1}
\end{aligned}$$

It turns out that the consistency of $\hat{\beta}$ depends heavily on the limiting behavior of the term R_n. Indeed, since Q_n tends to Q_0, a constant, when n grows. We only need that R_n grows at a rate lower than σ^2/n to have a variance that tends to 0 as n tends to infinity.

8.1.1 Generalized Least-Squares

The last result we derived about consistency of the OLS estimator is not satisfying enough, thus we might want to design a better estimator. The intuition needed to build a better one relies on two elements: first, we want an estimator that takes into account the new form of the variance matrix; second, we could transform the variance matrix into an identity matrix and somehow make our OLS estimator work. The Generalized Least-Squares estimator does exactly those two things.

Let P be a matrix such that:

$$\operatorname{Var}\left[Pe\right] = \sigma^2 I_n$$

This implies

$$\mathrm{E}\left[(Pe)(Pe)'\right] = \sigma^2 I_n \Leftrightarrow \mathrm{E}\left[Pee'P'\right] = \sigma^2 I_n \Leftrightarrow \sigma^2 P\Omega P' = \sigma^2 I_n \Leftrightarrow P\Omega P' = I_n$$

This is what we call the spectral decomposition of Ω. Now we have that the term Pe is indeed homoskedastic so we might look for the effects of multiplying our whole model by P.

Let $Y^* = X^*\beta + e^*$ where starred variables are the true variables projected by matrix P (i.e. $Y^* = PY$). This model satisfies all Gauss-Markov assumptions when the true model only violates heteroskedasticity. Therefore, we'll use OLS on the modified model to recover the true effects: $\hat{\beta}_{GLS} = (X^{*'}X^*)^{-1}X^{*'}Y^*$. What can we say about this new estimator in terms of X and Y (we already know that it is consistent in regards to X^* and Y^*)? We know that

$$\hat{\beta} = (X^{*'}X^*)^{-1}X^{*'}Y^* = (X'P'PX)^{-1}X'P'PY = (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}Y$$

Then,

$$\begin{aligned}
\mathrm{E}\left[\hat{\beta}\right] &= \mathrm{E}\left[(X'\Omega^{-1}X)^{-1}X'\Omega^{-1}Y\right] \\
&= (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}\,\mathrm{E}\left[X\beta + e\right] \\
&= (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}X\beta + (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}\,\mathrm{E}\left[e\right] \\
&= \beta
\end{aligned}$$

and:

$$\mathrm{Var}\left[\hat{\beta}\right] = \sigma^2(X^{*'}X^*)^{-1} = \sigma^2(X'P'PX)^{-1} = \sigma^2(X'\Omega^{-1}X)^{-1}$$

And we only need to estimate Ω in order to get our GLS estimate consistent.

Weighted Least Squares

Suppose the true model is

$$y_i = a + bx_i + cz_i + e_i$$

where $\mathrm{Var}\left[e_i\right] = \sigma^2 w_i^2$. In this context we can guess that $\mathrm{Var}\left[\frac{e_i}{w_i}\right] = \sigma^2$ and hence $P_{i\times i} = \frac{1}{w_i}$ (meaning that P is a matrix with diagonal terms equal to $1/w_i$). Then, our new model looks like

$$PY = Pa + PXb + PZc + Pe$$

or in a clearer way:

$$\frac{y_i}{w_i} = \frac{a}{w_i} + b\frac{x_i}{w_i} + c\frac{z_i}{w_i} + \frac{e_i}{w_i}$$

This is called Weighted Least Squares (where the variable w represents the weights put on each variable).

8.1.2 White test

Now that we know what to do in the case of heteroskedasticity, we might want to know how to test if the data is indeed heteroskedastic or not. In order to do this, there are three steps:

1. Regress the original model by OLS and keep the residuals $\hat{e}_i$
2. Regress the OLS residuals on all variables and their possible interactions (again, by OLS):

$$\hat{e}_i = a_0 + a_1 x_i + a_2 z_i + a_3 x_i^2 + a_4 z_i^2 + a_5 x_i z_i$$

3. If we have homoskedasticity, it must be that $\mathrm{E}[eX] = 0$, thus, testing for heteroskedasticity is equivalent to testing whether jointly $a_0 = a_1 = ... = 0$. In order to do that, construct the statistic nR^2 from the previous regression and it should follow a chi-squared distribution of $k + 1 + k!$ degrees of freedom.

$$nR^2 \xrightarrow{\mathrm{d}} \chi^2_{k+1+k!}$$

This procedure is known as the White test for heteroskedasticity.

8.1.3 White standard errors

If we do not have a given specification for heteroskedasticity in our model, we will have to fall back on OLS estimation. This causes issues because, while the OLS estimator is consistent, the variance of $\hat{\beta}$ depends on Ω which is not defined. We'll have to estimate it.

Recall that

$$\mathrm{Var}\left[\hat{\beta}\right] = \frac{\sigma^2}{n}\left(\frac{X'X}{n}\right)^{-1}\frac{X'\Omega X}{n}\left(\frac{X'X}{n}\right)^{-1}$$

which, since Ω is a diagonal matrix, gives

$$\mathrm{Var}\left[\hat{\beta}\right] = \frac{1}{n}\left(\frac{X'X}{n}\right)^{-1}\left[\frac{1}{n}\sum_i x_i x_i' \sigma_i^2\right]\left(\frac{X'X}{n}\right)^{-1}$$

Moreover, we know from the LLN that

$$\frac{1}{n}\sum_i x_i x_i' e_i^2 \to \frac{1}{n}\sum_i x_i x_i' \sigma_i^2$$

Hence we could use the OLS residuals to estimate this and get a consistent estimator for the variance of $\hat{\beta}$, namely:

$$\widehat{\operatorname{Var}\left[\hat{\beta}\right]} = \frac{1}{n}\left(\frac{X'X}{n}\right)^{-1}\left[\frac{1}{n}\sum_i x_i x_i' \hat{e}_i^2\right]\left(\frac{X'X}{n}\right)^{-1}$$

Note that relying on the LLN to get the result implies that while White standard errors give a consistent estimator for large samples, it may still be not consistent for small samples.

8.2 Autocorrelation

Autocorrelation is another type of inconsistency of the error term. This time, instead of variance changing with i, we have that error terms are correlated with each other: $\mathrm{E}\left[e_i e_j'\right] \neq 0$ for $j \neq i$. Because this issue usually arises in temporal contexts, we'll change indexes from i to t and get the following definition of autocorrelation: $\mathrm{E}\left[e_t e_{t-j}\right] \neq 0$ for $j > 0$.

8.2.1 Correlogram

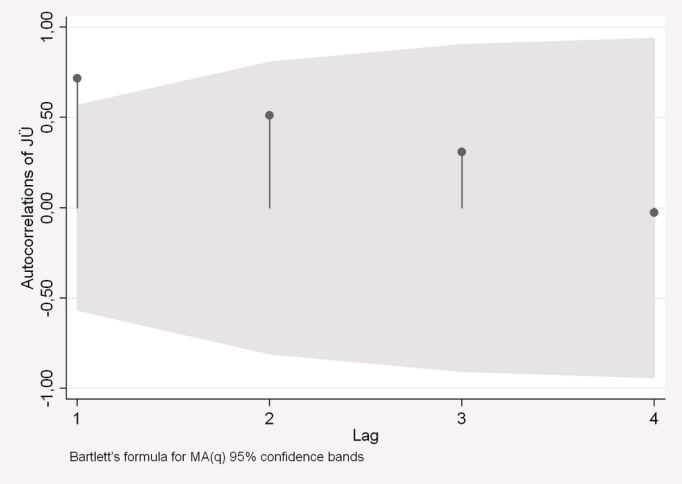
We might be interested first in how this autocorrelation is present in the data. For that purpose we'll use a measure of estimated correlation between two periods t and $t - s$ over the whole sample.

Definition 8.1 (Sample autocorrelation at lag s)**.** *For a given lag s, we define the sample autocorrelation, denoted $\hat{r}_s$ as follows:*

$$\hat{r}_s = \frac{\frac{1}{T-s}\sum_{t=s+1}^{T}\hat{e}_t\hat{e}_{t-1}}{\frac{1}{T}\sum_{t=1}^{T}\hat{e}_t^2}$$

If $\hat{r}_s$ is big in absolute value, then there is autocorrelation. If $\hat{r}_s$ is positive, then the autocorrelation is positive, and vice-versa.

We can represent sample autocorrelation graphically using a correlogram. For each lag, the correlogram will plot the value of the sample correlation in order to compare each one of them. For example, the following graph shows a 4-lag correlogram where sample autocorrelation seems to be decreasing over time:



After analysis of sample autocorrelations, one question remains: how many lags are significant in our data? In other words, for how many j do we have autocorrelation with the current error term? In order to answer that question, we define the Ljung-Box Q-statistic that will be used to test the number of significant lags.

Definition 8.2 (Ljung-Box Q-statistic)**.** *The Ljung-Box Q-statistic is defined as follows:*

$$Q = \sum_{s=1}^{L} \frac{(T+2)(T+s)}{T} \hat{r}_s^2$$

Under the null hypothesis (no autocorrelation in the first L lags), we have $Q \xrightarrow{d} \chi_L^2$. Hence it is possible to reject the null if Q does not follow this distribution. Note that in order to carry the test, you should have decided on a L to test in the first place. This could be done with the correlogram for example.

8.2.2 First-order autocorrelation

In this part of the section on autocorrelation, we'll study the case of a first-order autocorrelation. This model implies that only the first lag ($s = 1$) has positive correlation with the instant error. As such, we model our regression as

$$Y_t = X_t\beta + e_t$$

$$e_t = \rho e_{t-1} + v_t$$

We assume that v_t is a Gauss-Markov type of error term such that $\mathrm{E}[v_t] = 0$, $\mathrm{E}[v_t v_{t-s}] = 0$ for all $s \neq 0$, $\mathrm{E}[v_t^2] = \sigma_v^2$ and hence $\mathrm{E}[vv'] = \sigma^2 I_n$. Moreover we assume that the errors are not explosive, meaning that $|\rho| < 1$. From those assumptions, we can write:

$$\begin{aligned} e_t = \rho e_{t-1} + v_t = \rho(\rho e_{t-2} + v_{t-1}) + v_t &= \rho^2(\rho e_{t-3} + v_{t-2}) + \rho v_{t-1} + v_t \\ &= \dots \\ &= \sum_{s=0}^{\infty} \rho^s v_{t-s} \end{aligned}$$

and therefore,

$$\mathrm{E}[e_t] = \sum_{s=0}^{\infty} \rho^s \,\mathrm{E}[v_{t-s}] = 0$$

$$\begin{aligned} \mathrm{Var}[e_t] = \mathrm{E}\left[\left(\sum_{s=0}^{\infty} \rho^s v_{t-s}\right)^2\right] = \sum_{s=0}^{\infty} \rho^{2s} \,\mathrm{E}\left[v_{t-s}^2\right] &= \sigma_v^2 \sum_{s=0}^{\infty} (\rho^2)^s \\ &= \frac{\sigma_v^2}{1-\rho^2} \end{aligned}$$

The two last equations imply that the error term e_t is in fact homoskedastic. Because of that fact we can rewrite:

$$\mathrm{E}[e_t e_{t-s}] = \mathrm{E}\left[\left(\rho^s e_{t-s} + \sum_{s=0}^{\infty} \rho^s v_{t-s}\right) e_{t-s}\right] = \rho^s \sigma_e^2$$

In matrix form,

$$\mathrm{E}\left[ee'\right] = \sigma_e^2 \begin{bmatrix} 1 & \rho & \dots & \rho^{T-1} \\ \rho & 1 & \dots & \rho^{T-2} \\ \vdots & \vdots & \ddots & \vdots \\ \rho^{T-1} & \rho^{T-2} & \dots & 1 \end{bmatrix}$$

8.2.3 GLS and feasible GLS

As we have seen in the case of heteroskedasticity, knowing the value of $\mathrm{E}\left[ee'\right]$ will help us design a matrix P such that $\mathrm{Var}\left[Pe\right] = \sigma_e^2 I_T$. Here, the coefficient ρ is the key to having a modelthat satisfies GM assumptions.

Suppose $Y_t = a + bX_t + e_t$ is the true model with autocorrelation as presented in the beginning of this section. Then, take the first lag and multiply by ρ: $\rho Y_{t-1} = \rho a + \rho b X_{t-1} + \rho e_{t-1}$. By taking the difference:

$$Y_t - \rho Y_{t-1} = a - \rho a + bX_t - \rho b X_{t-1} + e_t - \rho e_{t-1}$$
$$\Leftrightarrow Y_t^* = a^* + bX_t^* + v_t$$

which satisfies the Gauss-Markov assumptions. The issue here is that in practice, we do not know the value of ρ. Hence we must turn to estimations of this value using a technique called feasible GLS.

The feasible GLS revolves around four steps:

1. Estimate $\hat{e}_t$ by performing OLS on the original model.
2. Estimate $\hat{\rho}$ by doing OLS on the error regression.
3. Estimate $\hat{\beta}$ and $\hat{e}_t$ by GLS.
4. Repeat steps 2 to 4 until the estimated value ρ has converged.

8.2.4 Other lag models

There are other specifications for the error lags. In particular, three types of models are often used:

AR(p) processes

These models function in the same way as the first-lag model described earlier, only this time we allow for $p \geq 1$ lags in the model:

$$e_t = \rho_1 e_{t-1} + ... + \rho_p e_{t-p} + v_t$$

MA(q) processes

Here, the errors are considered as moving averages of iid shocks that occurred in the last q periods.

$$e_t = v_t + \theta_1 v_{t-1} + ... + \theta_q v_{t-1}$$

ARMA(p, q) processes

These processes are combinations of AR(p) and MA(q) processes.

8.2.5 Newey-West standard errors

Newey-West standard errors are the autocorrelation analog of White standard errors in the heteroskedastic case. In that sense, they estimate the term $\frac{X'\Omega X}{T}$.

Again, recall that:

$$\text{Var}\left[\hat{\beta}\right] = \frac{\sigma^2}{T} Q_T^{-1} \frac{X'\Omega X}{T} Q_T^{-1}$$

Since Ω is not a diagonal matrix anymore, we have that

$$\frac{X'\Omega X}{T} = \frac{1}{T} \sum_{t=1}^{T} \sum_{s=1}^{T} \left(\text{Cov}\left(e_t, e_s\right) \cdot \left(x_t x_s' + x_s x_t'\right)\right)$$

$$\widehat{\frac{X'\Omega X}{T}} = \frac{1}{T} \sum_{t=1}^{T} \sum_{s=1}^{T} \left(\hat{e}_t \hat{e}_s \cdot \left(x_t x_s' + x_s x_t'\right)\right)$$

and finally, because after L lags, $e_t e_{t-L} = 0$, we have:

$$\widehat{\frac{X'\Omega X}{T}} = \frac{1}{T}\sum_{t=1}^{T}\sum_{s=T-L+1}^{T}(\hat{e}_t\hat{e}_s \cdot (x_t x_s' + x_s x_t'))$$

Chapter 9

Dynamic models and Time Series models

In this chapter we will cover a number of models and concepts related to estimation of temporal relationships in the data. The reasoning behind this kind of models is that sometimes, variables do not respond only to contemporaneous variables but also to previous realizations of these variables (i.e. their own past realizations or other variables' past realizations).

9.1 Dynamic Regression Models

9.1.1 Lagged effects in a dynamic model

Consider the following model:

$$y_t = a + b_0 x_t + b_1 x_{t-1} + ... + e_t$$

In this model, a one-time change in the variable x will affect the expectation of y in all subsequent periods. This is what we call a lagged effect. We consider two types of lagged effects: those which continue to effect y for an infinite amount of periods but with fading impact are called infinite lag models, those which cease to have an effect after a finite amount of periods are called finite lag models.

In such dynamic models, we measure the effect of a change in x_t by the variation on the equilibrium value of y_t. Assuming that there exists such an equilibrium, we define it as:

$$\bar{y} = a + \sum_{i=0}^{\infty} b_i \bar{x} = a + \bar{x} \sum_{i=0}^{\infty} b_i$$

Here you can clearly see that for this value to exist we need that the sum of b_i be finite.

Definition 9.1 (Short-run effect)**.** *In a dynamic model, we define the short-run effect or impact effect as the current-time coefficient of the model:* b_0.

Definition 9.2 (Cumulated effect)**.** *The cumulated effect of a dynamic model after* T *periods is defined as the sum of the first* T *coefficients of the model:* $\sum_{i=0}^{T} b_i$.

Definition 9.3 (Long-run effect)**.** *Finally, we define the long-run effect or equilibrium effect as the sum of all coefficients of the model:* $\sum_{i=0}^{\infty} b_i$.

Definition 9.4 (Lag weight)**.** *The lag weight* w_i *of a lag coefficient* b_i *is defined as:*

$$w_i = \frac{b_i}{\sum_{j=0}^{\infty} b_j}$$

Hence, we can rewrite our model as:

$$y_i = a + b \sum_{i=0}^{\infty} w_i x_{t-i} + e_t$$

Two other useful statistics of the lag weights are the median lag and the mean lag. They are defined respectively as:

$$t_{1/2} = \inf \left\{ t : \sum_{i=0}^{t} w_i \geq 0.5 \right\} \text{ and } \bar{t} = \sum_{i=0}^{\infty} i w_i$$

$$t_{1/2} = \inf \left\{ t : \frac{\sum_{i=0}^{t} b_i}{\sum_{i=0}^{\infty} b_i} \geq 0.5 \right\} \text{ and } \bar{t} = \frac{\sum_{i=0}^{\infty} i b_i}{\sum_{i=0}^{\infty} b_i}$$

9.1.2 Lag and difference operators

A convenient tool for manipulating lagged variables is the lag operator, denoted L. Placing L before a variable means taking its lag of one period. As an example, $Lx_t = x_{t-1}$. It is useful to define some properties of this operator:

- The lag of a constant is the constant: $La = a$.
- The lag of a lag is the second lag: $L(Lx_t) = L^2 x_t = x_{t-2}$.
- Thus, it works like a power: $L^p x_t = x_{t-p}$, $L^q(L^p x_t) = L^{q+p} x_t = x_{t-p-q}$, $(L^p + L^q)x_t = x_{t-p} + x_{t-q}$. Finally, $L^0 x_t = x_t$.

A related useful operation is the difference operator Δ such that:

$$\Delta x_t = (1 - L)x_t = x_t - x_{t-1}$$

9.2 Simple Distributed Lag Models

9.3 Autoregressive Distributed Lag Models

9.4 Issues with Dynamic Models

Chapter 10

Instrumental Variables, 2SLS, Endogeneity and Simultaneity

10.1 Correlation between errors and regressors

We have discussed many ways that our data could not satisfy Gauss-Markov assumptions for OLS. Now, we'll study the case of $\mathrm{E}[Xe] \neq 0$. How can this be? There are three main reasons why:

1. The specification is different from the true model. For example, if a variable is omitted from the model.
   ex. Let the true model be $y_i = a + bx_i + cz_i + e_i$ but we regress the model without z_i. Then, if $\mathrm{Cov}(x_i, z_i) \neq 0$ putting z_i in the error term will imply that $\mathrm{Cov}(X, e) \neq 0$.

2. The true model suffers from simultaneity of equations. This issue will be discussed later in the course but we'll show a quick example here.
   ex. Let the true model be $y_i = a + bx_i + e_i$ and $x_i = c + dy_i + u_i$. Then, because x_i both determines y_i and is determined by it, we'll have that $\mathrm{E}[Xe] \neq 0$.

3. Finally, if there is measurement error in X this could also lead to a non-null covariance between the errors and the regressors.

ex. Suppose the true model be $Y = \beta X^* + u$. However, suppose that X^* is not observed and instead we only have $X = X^* + v$. Assuming that u and v have nice properties (namely $\mathrm{E}[uX^*] = \mathrm{E}[vX^*] = \mathrm{E}[u] = \mathrm{E}[v] = \mathrm{E}[uv] = 0$), then you could regress $Y = \beta X + e$ and get $e = u - \beta v$. Hence, $\mathrm{E}[Xe] = -\beta \mathrm{E}[v^2] \neq 0$.

In general, suppose the model is $y = a + bx + e$, then $\hat{b} = b + \frac{\widehat{\mathrm{Cov}(x,e)}}{\widehat{\mathrm{Var}[x]}}$. Therefore,

$$\mathrm{E}[Xe] \neq 0 \Rightarrow \lim_{n\to\infty} \widehat{\mathrm{Cov}}(x, e) \neq 0 \Rightarrow \mathrm{plim}\, \hat{b} \neq b$$

10.2 Measurement errors

We have seen that under measurement errors of the form $X = X^* + v$ where X^* is the true value of the variable, $\mathrm{Cov}(X, e) = -\beta \mathrm{E}[v^2]$. Moreover, it is trivial to show that $\mathrm{Var}[X] = \mathrm{Var}[X^*] + \mathrm{Var}[v]$. Hence, we can show that,

$$\mathrm{plim}\, \hat{\beta} = \beta + \frac{-\beta \mathrm{Var}[v]}{\mathrm{Var}[X^*] + \mathrm{Var}[v]}$$

There are two important issues about this result: first, it shows an asymptotic bias of our OLS estimator ; second, the bias is leans strongly towards 0 and is positively correlated with β (the bigger β is, the bigger the bias).

This problem quickly becomes more important as more variables are subject to measurement errors. Indeed, while the direction of the bias is straightforward on the mismeasured variable's coefficient, the effect on other variables can go any direction! Hence, when multiple variables are mismeasured, then it is impossible to identify the direction of the bias for any of the coefficients.

10.3 Instrumental variables

10.3.1 Intuition

Suppose we find a variable Z such that $\text{Cov}(Z, Y) = b\,\text{Cov}(X, Z) + \text{Cov}(Z, e)$. Then, if $\text{Cov}(Z, e) = 0$, we have that:

$$b = \frac{\text{Cov}(Z, Y)}{\text{Cov}(X, Z)} \Rightarrow \hat{b} = \frac{\widehat{\text{Cov}(Z, Y)}}{\widehat{\text{Cov}(X, Z)}}$$

This estimator is called the IV estimator (for Instrumental Variable) while Z is called the instrument. This result shows two important facts:

- OLS estimation is a special of IV estimation when $Z = X$.
- In order to get a consistent $\hat{b}_{IV}$, we need that:
  1. $\text{Cov}(Z, e) = 0$: this requirement is described as the validity (or exogeneity) of the instrument Z.
  2. $\text{Cov}(Z, X) \neq 0$: this requirement is the relevance of the instrument.

  Together, these two requirements mean that a valid instrument has to affect Y only through its effect on X.

10.3.2 Generalization

We can generalize IV estimation in matrix form. Suppose the true model is $Y = X\beta + e$. We need that Z is the exact same dimensions of X (in practice, we do not have to instrument every column of X). Then,

$$Z'Y = Z'X\beta + Z'e \Leftrightarrow (Z'X)^{-1}Z'Y = \beta + (Z'X)^{-1}Z'e$$

We'll define $\hat{\beta}_{IV} = (Z'X)^{-1}Z'Y$.

Hence, we can rewrite the previous equation as:

$$\hat{\beta}_{IV} = \beta + \left(\frac{Z'X}{n}\right)^{-1}\left(\frac{Z'e}{n}\right)$$

This estimator is consistent if $\operatorname{plim} \frac{Z'X}{n}$ is non-singular and $\operatorname{plim} \frac{Z'e}{n} = 0$.

Notice that we have $\sqrt{n}(\hat{\beta} - \beta) = \sqrt{n} \left(\frac{Z'X}{n}\right)^{-1} \left(\frac{Z'e}{n}\right)$. We can therefore try to prove root-n consistency and asymptotic normality ($\sqrt{n}$-CAN). First, using CLT, we'll show that $\sqrt{n} \left(\frac{Z'e}{n}\right) \xrightarrow{\text{d}} N(0, \sigma^2 \operatorname{E}\left[\frac{Z'Z}{n}\right])$:

Then, using the law of large numbers (LLN), we can show that $\frac{Z'X}{n} \xrightarrow{\text{p}} \operatorname{E}\left[\frac{Z'X}{n}\right]$. Hence, by the properties of convergence, we have that:

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{\text{d}} N\left(0, \sigma^2 \underbrace{\operatorname{E}\left[\frac{Z'X}{n}\right]^{-1}}_{\Sigma_{ZX}} \underbrace{\operatorname{E}\left[\frac{Z'Z}{n}\right]}_{\Sigma_{ZZ}} \underbrace{\operatorname{E}\left[\frac{X'Z}{n}\right]^{-1}}_{\Sigma_{XZ}}\right)$$

And hence, $\hat{\beta}_{IV} \xrightarrow{\text{d}} N\left(\beta, \frac{\sigma}{n}\Sigma_{ZX}^{-1}\Sigma_{ZZ}\Sigma_{XZ}^{-1}\right)$.

10.4 Multiple IVs and 2SLS

Now, suppose that our true model is: $Y = a + bX + e$ as before but this time you observe two valid instruments Q and R... From what we know, we could either estimate by IV $\hat{b}_Q$, $\hat{b}_R$ or even any $\hat{b}_{QR}$ which would be a any linear combination of both instruments. Indeed, because $Z = \alpha_0 + \alpha_1 Q + \alpha_2 R$ is also a valid instrument (however not always relevant), we have at our disposition a continuum of valid instruments. The obvious question that we'll answer in this section is how do we choose between all those instruments.

The intuition for how to choose our instrument relies on the probability limit of b when instrumenting with Z. We have seen in the previous section that this value is:

$$\hat{b}_{IV} = b + \frac{\widehat{\operatorname{Cov}(Z, e)}}{\widehat{\operatorname{Cov}(Z, X)}}$$

From this equation we see that we want the covariance of Z and X being the highest possible while maintaining a small covariance with the error term. This boils down to finding Z such that its correlation with X is the highest. Hence we'll use an OLS estimation.

The OLS regression performed here will be of X on Q and R:

$$X = \alpha_0 + \alpha_1 Q + \alpha_2 R + u \Rightarrow Z = \hat{X}$$

Then we use Z as the instrument for an IV regression in the true model. This process is called two-stage least-squares or 2SLS (even though the second stage is not an OLS regression). Then we can rewrite our 2SLS estimator as:

$$\hat{b}_{2SLS} = \frac{\operatorname{Cov}\left(\hat{X}, Y\right)}{\operatorname{Cov}\left(\hat{X}, X\right)} = \frac{\operatorname{Cov}\left(\hat{X}, Y\right)}{\operatorname{Var}\left[\hat{X}\right]}$$

which is seemingly close to the OLS estimator using $\hat{X}$ but it is not.

In matrix form, let our true model be:

$$\underbrace{Y}_{n\times 1} = \underbrace{X}_{n\times k} \cdot \underbrace{\beta}_{k\times 1} + \underbrace{e}_{n\times 1}$$

and let our instruments matrix be Q, a $n \times l$ matrix where $l \geq k$ (i.e. there are more instruments than regressors). Then, the 2SLS process follows the following two steps:

1. We estimate $Z = \hat{X} = Q(Q'Q)^{-1}Q'X$ by OLS.
2. We estimate $\beta_{2SLS} = (Z'X)^{-1}Z'Y$ by IV.

Notice that all issues regarding inference, the values of α_j do not matter because Z is as valid as a single instrument (same inference) and any combination will do the job.

10.5 Testing IVs

The testing of instrumental variables revolves around two main questions:

- Does the model need instruments? We can test this statement by looking at $\mathrm{E}[Xe]$ and verifying how it compares to 0.

- Are the instruments provided valid? This question is equivalent to looking at $\mathrm{E}[Qe] = 0$

In order to perform those tests, you need an over-identified model (more instruments than regressors).

10.5.1 Hausman test

The Hausman test is the name of the procedure done to test if $\mathrm{E}[Xe] = 0$ or not. In order to perform this test, we will assume that regardless of the need for instruments, the instruments are valid (i.e. $\mathrm{E}[Qe] = 0$). Then, by assumption, if the model does not need any instrument, the results of OLS and 2SLS should be the same. In order to compare the two models, we'll separate X in two partitions: the potentially endogenous regressors $\tilde{X}$ and the rest. Then we estimate $\hat{\tilde{X}} = Q(Q'Q)^{-1}Q'\tilde{X}$.

Under the null hypothesis (the model does not need any instruments) the OLS regression on

$$Y = X\beta + \hat{\tilde{X}}\gamma + u$$

should give $\hat{\gamma} = 0$. Notice that $\hat{\tilde{X}}\gamma$ actually represents the error term that would be included in u if there were no instruments.

To test $\hat{\gamma} = 0$ we can use a F-test (or a t-test if γ is unidimensional). However, the test power is very low, hence non-rejection does not mean that the model without instrument is perfect.

10.5.2 Hansen-Sargan test

The Hansen-Sargan test procedure has the goal of determining if $\mathrm{E}[Qe] = 0$. The procedure is divided in three steps:

1. Estimate by 2SLS the residuals $\hat{e} = Y - X\hat{\beta}_{2SLS}$.
2. Regress the estimated residuals on Q the matrix containing the instruments: $\hat{e} = Q\delta + v$.

3. Test the value of δ with the statistic:

$$J = nR^2 \sim \chi^2_{l-k}$$

Notice that the residuals estimated by 2SLS use only k regressors while Q provides l; this is why we need that $l > k$ to test the validity of instruments: k regressors are used in estimating $\hat{e}$, $l - k$ are left to test the validity of our instruments.

10.6 Simultaneity

10.6.1 IV/2SLS

The issue of simultaneity arises when two equations to estimate depend on each other as a system. For example, it could be that $Y = X\beta + e$ and $X = y\gamma + u$ and GM assumptions would be violated because of the non-zero covariance between the error terms and the regressors.

We'll see how to deal with this issue by working on a frequent example in IO: estimating a demand-supply system. Let the supply and demand equations be:

$$S : Q = \alpha_2 P + \varepsilon$$

$$D : Q = \beta_2 P + \beta_3 Y + u$$

These two equations together are called the structural model, they are directly derived from theory and can contain relations with each other. As we've seen, because of simultaneity, this model cannot be estimated by OLS.

We could try and solve for P. From the supply function, we have that $P = \frac{Q-\varepsilon}{\alpha_2}$. Plugging it into the demand function we get $Q = Q\frac{\beta_2}{\alpha_2} - \frac{\beta_2}{\alpha_2}\varepsilon + \beta_3 Y + u$ which gives:

$$\left[1 - \frac{\beta_2}{\alpha_2}\right] Q = \beta_3 Y + u - \frac{\beta_2}{\alpha_2}\varepsilon$$

$$Q = \frac{\beta_3 \alpha_2}{\alpha_2 - \beta_2} Y + \frac{\alpha_2 u - \beta_2 \varepsilon}{\alpha_2 - \beta_2}$$

$$P = \frac{\beta_3}{\alpha_2 - \beta_2} Y + \frac{u - \varepsilon}{\alpha_2 - \beta_2}$$

Notice here that the new system does not rely on any endogenous variable and hence can be estimated by OLS, although the parameters will not be consistent. This new system is called the reduced-form and can serve the purpose of forecasting variables.

Now, going back to our structural model, we have seen that OLS cannot be performed because of the covariance between the regresssor and the error term. Indeed,

$$\operatorname{plim} \hat{\alpha}_2 = \alpha_2 + \frac{\operatorname{Cov}(P, \varepsilon)}{\operatorname{Var}[P]} \neq \alpha_2$$

Hence we need to use an instrumental variable to estimate the supply properly. It turns out that in this setting Y_i makes a perfect instrument because it is related to supply uniquely via it correlation with P_i. This variable is what we call a demand-shifter. Because it shifts demand and demand only, it allows us to identify the slope of the supply curve. Notice that Y_i is a valid instrument because it appears only in the structural equation of the demand function. Consequently, we can guess that we cannot estimate the slope of demand (there is no supply-shifter in the supply equation).

10.6.2 Seemingly unrelated regression

Suppose that we have two different models for n individuals, represented as:

$$Y_{1i} = a + bX_{1i} + u_{1i}$$

$$Y_{2i} = c + dX_{2i} + u_{2i}$$

where the two models both satisfy all Gauss-Markov assumptions. Nevertheless, both error terms are correlated across models for a given individual only, i.e. $\operatorname{Cov}(u_{1i}, u_{2i}) \neq 0$ for all i. Why not use OLS then? Of course, OLS estimation is actually interesting because both models separately respect GM assumptions, thus yielding $\sqrt{n}$-CAN estimators. However, the last point about correlation across

models can help us achieve a more efficient estimator (it is indeed additional information, why not use it?). Consider stacking the two equations as:

$$\begin{bmatrix} Y_{11} \\ \vdots \\ Y_{1n} \\ Y_{21} \\ \vdots \\ Y_{2n} \end{bmatrix} = a \begin{bmatrix} 1 \\ \vdots \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} + c \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ \vdots \\ 1 \end{bmatrix} + b \begin{bmatrix} X_{11} \\ \vdots \\ X_{1n} \\ 0 \\ \vdots \\ 0 \end{bmatrix} + d \cdot \begin{bmatrix} 0 \\ \vdots \\ 0 \\ X_{21} \\ \vdots \\ X_{2n} \end{bmatrix} + \begin{bmatrix} u_{11} \\ \vdots \\ u_{1n} \\ u_{21} \\ \vdots \\ u_{2n} \end{bmatrix}$$

where the variance matrix is:

$$\Omega = \begin{bmatrix} \sigma_1^2 & \dots & 0 & \sigma_{12} & \dots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & \sigma_1^2 & 0 & \dots & \sigma_{12} \\ \sigma_{12} & \dots & 0 & \sigma_2^2 & \dots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & \sigma_{12} & 0 & \dots & \sigma_2^2 \end{bmatrix}$$

We can then design a feasible GLS estimator on that system:

1. Start with regressing both models separately by OLS to get the estimates $\hat{u}_1$ and $\hat{u}_2$. Construct $\hat{\Omega}$ using $\hat{\sigma}_1^2 = \widehat{\text{Var}\,[\hat{u}_1]}$, $\hat{\sigma}_2^2 = \widehat{\text{Var}\,[\hat{u}_2]}$ and $\hat{\sigma}_{12} = \widehat{\text{Cov}\,(\hat{u}_1, \hat{u}_2)}$.
2. Use the matrix $\hat{\Omega}^{-1}$ in the GLS estimator:

$$\hat{\beta}_{GLS} = (X'\hat{\Omega}^{-1}X)^{-1}X'\hat{\Omega}^{-1}Y$$

3-stage least-squares

Now, further suppose that your system of SUR does not satisfy Gauss-Markov assumptions, then you could instrument it to estimate the residuals. This method is called 3SLS, as it requires that you estimate the residuals $\hat{u}_1$ and $\hat{u}_2$ by 2SLS, and then do GLS with the covariance matrix calculated then.

Chapter 11

Non-linear models, GMM and extremum estimators

11.1 Nonlinear Least Squares

11.1.1 Model

Suppose our model is $Y_i = g(X_i, \theta)$ where $g(\cdot)$ is a nonlinear function of parameters θ. With what we know, we could still use a least-squares approach to find the best estimator, that is:

$$\hat{\theta} = \arg\max_{\theta} \sum_{i=0}^{n} \hat{e}_i^2 = \arg\max_{\theta} \sum_{i=0}^{n} [Y_i - g(X_i, \theta)]^2$$

We take the first-order condition:

$$\frac{\partial \mathcal{L}}{\partial \theta} = 0 \Leftrightarrow \sum_{i=0}^{n} \left(2[Y_i - g(X_i, \hat{\theta})] \frac{\partial g(X_i, \hat{\theta})}{\partial \theta} \right) = 0$$

But the issue becomes finding $\hat{\theta}$ such that this condition is satisfied (a harder problem that will be treated in the following section). For the time being, we can ask ourselves what are the properties of this estimator.

By looking at minimizing the average sum of squared residuals instead, we find that the FOC is:

$$\frac{1}{n}\sum_{i=1}^{n}\left([Y_i - g(X_i,\hat{\theta})]\frac{\partial g(X_i,\hat{\theta})}{\partial\theta}\right) = 0$$

Hence, by the law of large numbers,

$$\mathrm{E}\left[(Y - g(X,\hat{\theta}))\frac{\partial g(X,\hat{\theta})}{\partial\theta}\right] = 0$$

By expanding Y and iterated expectations:

$$\mathrm{E}\left[\mathrm{E}\left[(g(X,\theta_0) + e - g(X,\hat{\theta}))\frac{\partial g(X,\hat{\theta})}{\partial\theta}|X\right]\right] = 0$$

which gives, when you notice that $\mathrm{E}[e|X] = 0$:

$$\mathrm{E}\left[(g(X,\theta_0) - g(X,\hat{\theta}))\frac{\partial g(X,\hat{\theta})}{\partial\theta}\right] = 0$$

Therefore, two types of estimators might be unbiased: the obvious $\hat{\theta} = \theta_0$ but also the undesired $\hat{\theta}$ such that $\frac{\partial g(X,\hat{\theta})}{\partial\theta} = 0$. For a perfect identification of the parameters θ we need the assumption that there is a unique value θ_0 for which $\frac{\partial g(X,\theta_0)}{\partial\theta} = 0$ (i.e. a similar assumption to the one we made about extremum estimators).

11.1.2 Estimation

Let's go back to the question of how to estimate $\hat{\theta}$. By using a first-order Taylor expansion of $g(X,\theta_0)$ around $\hat{\theta}$, we have that:

$$g(X,\theta_0) \approx g(X,\hat{\theta}) + \left(\frac{\partial g(X,\hat{\theta})}{\partial\theta}\right)'(\theta_0 - \hat{\theta})$$

implying that we could rewrite the true model as:

$$Y_i \approx g(X_i,\hat{\theta}) + \left(\frac{\partial g(X_i,\hat{\theta})}{\partial\theta}\right)'(\theta_0 - \hat{\theta}) + e_i$$

$$Y_i - g(X_i, \hat{\theta}) + \left(\frac{\partial g(X_i, \hat{\theta})}{\partial \theta} \right)' \hat{\theta} \approx \left(\frac{\partial g(X_i, \hat{\theta})}{\partial \theta} \right)' \theta_0 + e_i$$

This last equation is essentially a linear model now, with θ_0 being the coefficient that could be estimated by simple OLS. However, you do not have the first value of $\hat{\theta}$, hence you cannot do this regression, there are many ways to find suitable values for $\hat{\theta}$, two of them being interesting and useful enough to discuss here: the gradient method and the grid search.

Gradient method

Grid-search method

11.2 Extremum Estimators

11.3 Generalized Method of Moments

11.3.1 Method of moments

Moment equation models are models where the population parameters solve a system of moment equations. This class of models is much broader than the model that we have considered so far, and we will see how OLS, IV or MLE fall into the GMM estimation class.

Let $g_i(\theta)$ be a l-dimensional observed function of a k-dimensional parameter vector θ. A moment equation model is summarized by the moment equations:

$$\mathrm{E}\left[g_i(\theta)\right] = 0$$

which is a $l \times k$ equation system. Hence, all properties of equation systems hold here: $l \geq k$ implies the model is just (or over) identified, leading to potentially unique solutions for θ ; $l < k$ implies underidentification. In this chapter, we will only consider the first case.

Estimation

Define the sample estimator for our moment equation as:

$$\bar{g}_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} g_i(\theta)$$

The method of moments estimator (MME), denoted $\hat{\theta}_{MM}$ for θ is the value for which $\bar{g}_n(\hat{\theta}) = 0$. This last equation system is named the estimating equation system.

Although this estimator seems very simple to use, it is not guaranteed that a (unique) solution can be found analytically, or even numerically. Before going further in the topic of GMM, we will look at its estimators in known cases.

Mean and variance of a population

Here, set $g_i(\mu) = y_i - \mu$, we clearly have that $\mathrm{E}\left[g_i(\mu)\right] = 0$. Hence, the MME is $\hat{\mu}$ for which

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\mu}) = 0 \Leftrightarrow \hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

In the more complex case where we want to estimate the mean and variance of a population, then

$$g_i(\mu, \sigma^2) = \begin{pmatrix} y_i - \mu \\ (y_i - \mu)^2 - \sigma^2 \end{pmatrix}$$

which also gives $\mathrm{E}\left[g_i(\mu, \sigma^2)\right] = 0$. Here the MME would be $(\hat{\mu}, \hat{\sigma}^2)$ such that:

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^{n} g_i(\hat{\mu}, \hat{\sigma}^2) &= 0 \\ \frac{1}{n}\sum_{i=1}^{n} \begin{pmatrix} y_i - \hat{\mu} \\ (y_i - \hat{\mu})^2 - \hat{\sigma}^2 \end{pmatrix} &= 0 \\ \begin{pmatrix} \hat{\mu} \\ \hat{\sigma}^2 \end{pmatrix} &= \begin{pmatrix} \frac{1}{n}\sum_{i=1}^{n} y_i \\ \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\mu})^2 \end{pmatrix} \end{aligned}$$

OLS

In the OLS case, the population distribution of variables follow the simple following model:

$$Y_i = X_i\beta + e_i \text{ where } e_i \sim N(0, \sigma^2) \text{ and } e_i \perp X_i$$

There are k parameters to estimate in this regression (i.e. the k lines in the β vector). Thus, we require at least k moment conditions to identify these parameters. Let's try and find what those moments can be. From our model, a good condition could be that $e_i \perp X_i$, implying the condition $\mathrm{E}\left[X_i e_i\right] = \mathrm{E}\left[X_i(Y_i - X_i\beta)\right] = 0$. Here, $g_i(\beta) = X_i(Y_i - X_i\beta)$. This is a $k \times 1$ vector, meaning that it contains in fact k conditions: the model is exactly identified.

Hence the MME is $\hat{\beta}$ that satisfies:

$$\frac{1}{n}\sum_{i=1}^{n} X_i(Y_i - X_i'\hat{\beta}) = 0$$

$$\frac{1}{n}\sum_{i=1}^{n}[X_iY_i - X_iX_i'\hat{\beta}] = 0$$

$$\frac{1}{n}\sum_{i=1}^{n} X_iY_i = \frac{1}{n}\sum_{i=1}^{n} X_iX_i'\hat{\beta}$$

$$\sum_{i=1}^{n} X_iY_i = \sum_{i=1}^{n} X_iX_i'\hat{\beta}$$

which finally gives: $\hat{\beta} = \left(\sum_{i=1}^{n} X_iX_i'\right)^{-1}\left(\sum_{i=1}^{n} X_iY_i\right)$, the OLS estimator that we know.

IV

In an instrumental variables setting, the moment condition we used in the OLS case is not valid anymore. Indeed, $\mathrm{E}\left[Xe\right] \neq 0$, and it's exactly for that reason that we are using the l instruments Z. Nevertheless, recall that $\mathrm{E}\left[Z_i e_i\right] = 0$ for instruments to be valid, or alternatively $\mathrm{E}\left[Z_i(Y_i - X_i\beta)\right] = 0$. Again, this implies that $g_i(\beta) = Z_i(Y_i - X_i\beta)$, a $l \times 1$ vector, thus a l-moment conditions vector. If

by chance $l = k$, the model is then exactly identified and we get the following method of moments estimator:

$$\hat{\beta}_{IV} = \left(\sum_{i=1}^{n} Z_i X_i' \right)^{-1} \left(\sum_{i=1}^{n} Z_i Y_i \right)$$

11.3.2 Generalized Method of Moments

Using the previous example of instrumental variables regression, it might be the case that the number of instruments provided is actually greater than the number of parameters to estimate. In this case, it means that we are trying to find the value of k parameters that solve $l > k$ equations: the solution might not be unique. Then we need to find another way to estimate the parameters.

Recall that our estimation process relied on finding β such that $\bar{g}_n(X_i, \beta) = 0$. But now, this value of β might not exist. Using the IV case as a motivating example, think of $\bar{g}_n(X_i, \beta)$ as being equal to $\frac{1}{n}(Z'Y - Z'X\beta)$. Further define $\mu = Z'Y$ and $G = Z'X$. Then, $Z'Y - Z'X\beta = \mu - G\beta \equiv \eta$ which we'll call the error term. Because this η represents the difference that we are trying to set the closest possible to 0, we might want to use a well-known technique to minimize distance: the least-squares method. Write our model as $\mu = G\beta + \eta$, a simple estimate would then be:

$$\hat{\beta} = (G'G)^{-1}G'\mu$$

However, we might have that η is not well-behaved enough for a classical least-squares regression. In particular, it could be that $\mathrm{E}[\eta\eta'] \neq \sigma^2 I_n$. In this case, it might be better to use the generalized least-squares estimators. For a given weight matrix W,

$$\begin{aligned}
\hat{\beta} &= (G'WG)^{-1}G'W\mu \\
&= ((Z'X)'W(Z'X))^{-1}(Z'X)'WZ'Y \\
&= (X'ZWZ'X)^{-1}X'ZWZ'Y
\end{aligned}$$

This estimator minimizes the weighted sum-of-squares: $\eta' W \eta$. Now, recall that $\eta = \bar{g}_n(X_i, \beta)$ and hence we can rewrite our estimator $\hat{\beta}$ as:

$$\hat{\beta} \in \arg\min_{\beta} [\bar{g}_n(X_i, \beta)]' W [\bar{g}_n(X_i, \beta)]$$

for a given weight matrix W.

11.3.3 GMM Estimator properties

11.3.4 Efficient GMM

11.3.5 Estimation of the Efficient Weight Matrix

11.3.6 Variance Estimation

Chapter 12

Non-parametric estimators

12.1 Introduction

The goal of this whole chapter is to understand the implications of non and semi parametric methods in typical econometrics models. For the rest of this chapter, we will assume that observations in the data are i.i.d.

First, let's review the differences between what those concepts mean:

- As we have seen, a parametric regression is exactly what we have done since the beginning of the class: you presuppose a model that is fully specified in its parameters. This includes of course the linear model, but also more general distributions of parameters (GMM). In this type of regressions, the parameters have finite dimensions.

- A nonparametric regression would imply a model of infinite dimensional parameters: $Y_i = m(X_i) + e_i$ where $m(\cdot)$ is a function that could basically be anything.
  - ✓ A nonparametric regression does not require a fully specified model for estimation: this can be useful if the particular distribution of a variable is not given (i.e. who says errors are i.i.d. normal)
  - x The extremely high dimensionality of nonparametric models can make them very hard to compute.

- A semiparametric regression is between both, restricting parameters of interest to finite dimensions while allowing other parameters to have infinite dimensions.
  - ✓ A semiparametric regression can overcome the high-dimensionality issue of nonparametric models.
  - ✓ A semiparametric regression only focuses on variables of interest, allowing free movements of other variables.
  - ✓ A semiparametric regression is increasingly popular among econometricians.

Let X be a random variable (a scalar for now), x is a realization of X. As before, X_i and x_i are respectively iid random variables and their realizations. Suppose $X \sim F(X)$ for a given $F(\cdot)$ and each X_i has the distribution F.

Definition 12.1 (Empirical distribution fucntion)**.** *Define $\hat{F}(x)$, the empirical distribution function, evaluated at x as:*

$$\hat{F}(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}[X_i \leq x]$$

where $\mathbb{I}$ is the indicator function, taking the value 1 if the condition inside the bracket is met, 0 else. In words, empirical distribution function is the sample proportion of observations lower than or equal to x.

Graphically, if we plot $\hat{F}(x)$ against x, we can see it representing an step-wise approximation of the true distribution F. Below is an example of this for a random sample of 100 observations drawn from the standard normal distribution.

12.2 Estimation of the EDF

From what the graph in the previous section showed us, it seems natural to consider the EDF as a nonparametric estimator for $F(x)$. What are its properties?

For any real number x,

$$\begin{aligned}
\mathrm{E}\left[\hat{F}(x)\right] &= \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n}\mathbb{I}[X_i \leq x]\right] \\
&= \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}\left[\mathbb{I}[X_i \leq x]\right] \\
&= \mathrm{E}\left[\mathbb{I}[X \leq x]\right] \\
&= \int_{-\infty}^{\infty}\mathbb{I}[X \leq x]f(X)\mathrm{d}X \\
&= \int_{-\infty}^{x} f(X)\mathrm{d}X \\
&= F(x)
\end{aligned}$$

Hence the EDF estimator is unbiased. In the same way, we have:

$$\begin{aligned}
\mathrm{Var}\left[\hat{F}(x)\right] = \mathrm{E}\left[(\hat{F}(x) - F(x))^2\right] &= \mathrm{E}\left[\left(\frac{1}{n}\sum_{i=1}^{n}\mathbb{I}[X_i \leq x] - F(x)\right)^2\right] \\
&= \frac{F(x)(1-F(x))}{n}
\end{aligned}$$

implying that the EDF estimator is also consistent. Finally, since $\hat{F}(x)$ is also an average, we can apply the CLT and show that it is $\sqrt{n}$-consistent and asymptotically normal:

$$\sqrt{n}\left(\hat{F}(x) - F(x)\right) \xrightarrow{\mathrm{d}} N[0, F(x)(1-F(x))]$$

12.3 Estimation of the empirical pdf

Now suppose we want to estimate $f(x)$, the probability density function instead. We have seen the parameterized equivalent of this problem with Maximum likelihood estimators. Without any parameters, we can use an approximation by a histogram. Let h represent half of the width of each bin, the histogram of our data can be represented by:

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{I}[x - h \leq x_i \leq x + h]$$

Note that this is equivalent to the following equation:

$$\begin{aligned}\hat{F}(x+h) - \hat{F}(x-h) &= \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}[X_i \leq x+h] - \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}[X_i \leq x-h] \\ &= \frac{1}{n}\sum_{i=1}^{n} \left(\mathbb{I}[X_i \leq x+h] - \mathbb{I}[X_i \leq x-h]\right) \\ &= \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}[x-h \leq X_i \leq x+h] \\ &= \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\left[\left|\frac{x-X_i}{h}\right| \leq 1\right]\end{aligned}$$

The probability at the exact point x is the value of the proportion when the bin width tends to 0. This gives:

$$\hat{f}(x) = \lim_{h\to 0} \frac{\hat{F}(x+h) - \hat{F}(x-h)}{2h} = \lim_{h\to 0} \frac{1}{h}\frac{1}{n}\sum_{i=1}^{n} \frac{1}{2}\mathbb{I}\left[\left|\frac{x-X_i}{h}\right| \leq 1\right]$$

But since we cannot compute a zero binwidth, we will allow for our estimator to vary depending on h, we get formally that:

$$\hat{f}_h(x) \equiv \frac{1}{nh}\sum_{i=1}^{n} \left(\frac{1}{2}\mathbb{I}\left[\left|\frac{x-X_i}{h}\right| \leq 1\right]\right)$$

The part in the sum can in fact be written as a function of $\frac{x-X_i}{h}$, call it $K(\cdot)$ where

$$K(u) = \frac{\mathbb{I}\left[|u| \leq 1\right]}{2}$$

so that $K(u)$ is in fact the pdf of a uniform distribution on the interval $[-1, 1]$. It turns out that this function $K(\cdot)$ is called a kernel. This particular one has simple properties as it gives the same weight to any observations in the h-neighborhood of x and a zero weight to those outside of the h-neighborhood. There are many other existing kernels that might give more information about a pdf. Let's look at a few of them.

The Gaussian kernel estimator has the following kernel:

$$K(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}}$$

The Epanechikov kernel estimator has the following kernel:

$$K(u) = \frac{3}{4}(1 - u^2) \cdot \mathbb{I}[|u| \leq 1]$$

12.3.1 Properties of the kernel estimator

The kernel estimator is unfortunately biased. In particular, we can show that:

$$\mathrm{E}\left[\hat{f}_h(x)\right] = f(x) + h^2 b(x)$$

Nevertheless, the estimator's variance will tend to 0 for a constant h as:

$$\mathrm{Var}\left[\hat{f}_h(x)\right] \approx \frac{1}{nh} v(x)$$

The two previous properties introduce a typical issue in nonparametric models: the MSE trade off. In fact, we can see that choosing a small h will reduce the bias, but increase the variance. The opposite is true as well, a great h will make the estimator more biased while it will reduce its variance. The MSE equation is given by

$$\mathrm{MSE} \approx h^4 b^2(x) + \frac{1}{nh} v(x)$$

The h that minimizes the MSE satisfies the following FOC:

$$4h^3 b^2(x) - \frac{1}{nh^2} v(x) = 0$$

$$h^5 = \frac{v(x)}{4b^2(x)} \cdot \frac{1}{n}$$

$$h = \left[\frac{v(x)}{4b^2(x)}\right]^{1/5} n^{-1/5}$$

This shows an interesting feature: the bigger the number of observations n, the smaller the best h gets. This means that for huge datasets, we can get away with setting h really small to get unbiased estimates.

12.4 Kernel regressions

12.5 Series regressions

12.6 Semiparametric regressions

Chapter 13

Program Evaluation and Treatment Effects

13.1 Introduction

This chapter focuses on a particular type of situation where we observe an outcome Y that is a consequence of a treatment T. Hence, T causes Y but not the other way around.

We will also restrict our attention to binary treatments meaning that:

- $T_i = 1$ if i is in the treatment group.
- $T_i = 0$ if i is in the control group.

Hence, following the Rubin causal notation, for each individual i there are two outcomes: $Y_i(1)$ if individual i was in the treatment group and $Y_i(0)$ if individual i was in the control group. Clearly, you can see that only one of those two outcomes can be observed! In fact, it is impossible that an individual was in both groups at the same time. We call the unobserved outcome the counterfactual.

The actual outcome can be written as:

$$Y_i = Y_i(0) + [Y_i(1) - Y_i(0)] \cdot T_i$$

where $[Y_i(1) - Y_i(0)]$ is called the treatment effect. Note that for the same reason exposed earlier, this treatment effect cannot be observed or computed exactly. That's why the literature has focused on two related but different problems: what is the average treatment effect (ATE)? And what is the average treatment effect on treated individuals (ATT)? These two questions help us formalize the two concepts:

$$\text{ATE} = \text{E}\left[Y_i(1) - Y_i(0)\right]$$

$$\text{ATT} = \text{E}\left[Y_i(1) - Y_i(0)|T_i = 1\right]$$

13.2 Average Treatment Effect (ATE)

From the data, we can identify two elements: $\text{E}\left[Y_i(1)|T_i = 1\right]$, the average outcome of the treatment group and $\text{E}\left[Y_i(0)|T_i = 0\right]$, the average outcome of the control group. Notice that here

$$\text{E}\left[Y_i(1)|T_i = 1\right] = \text{E}\left[Y|T_i = 1\right]$$

$$\text{E}\left[Y_i(0)|T_i = 0\right] = \text{E}\left[Y|T_i = 0\right]$$

If the treatment is truly random, then we have that:

$$\text{ATE} = \text{E}\left[Y_i(1)|T_i = 1\right] - \text{E}\left[Y_i(0)|T_i = 0\right]$$

It is trivial to understand this relation by intuition, but can be helpful to put it in more formal terms. This result depends on two assumptions grouped as assumptions for mean unconditional unconfoundedness.

Definition 13.1 (Mean unconditional unconfoundedness)**.** *A treatment dataset satisfies MUU if:*

1. $\text{E}\left[Y(1)|T = 1\right] = \text{E}\left[Y(1)|T = 0\right] = \text{E}\left[Y(1)\right]$*: this implies that regardless of the group we consider, the average outcome of receiving the treatment is the same.*

2. $\text{E}\left[Y(0)|T = 1\right] = \text{E}\left[Y(0)|T = 0\right] = \text{E}\left[Y(0)\right]$*: this implies that regardless of the group we consider, the average outcome of not receiving the treatment is the same.*

If these assumptions hold, then the ATE written as $\text{ATE} = \text{E}[Y_i(1)|T_i = 1] - \text{E}[Y_i(0)|T_i = 0]$ is a consistent estimator. But what do those assumptions rely on? How can we interpret them in terms of a real-life situation.

For example, when the treatment T is randomly assigned, meaning that T is independent of the outcomes: $(Y(0), Y(1)) \perp T$, then we have unconditional unboundedness. The difference between this and the assumptions above is that "mean" unconfoundedness implies that on average the treatment is as if it was random. Taking out the "mean" implies that it is actually truly random for every individual. Already, we can see how these assumptions can be described as strong since in reality it is very rare to observe true randomness.

To solve this issue, consider the weaker statement that within groups of individuals sharing a characteristic, the treatment is random. That is, conditional on an observable characteristic X, the treatment is randomly assigned or formally, $(Y(0), Y(1)) \perp T|X$. This property is called confoundedness and allows us to get the analog of the previous results, accounting for X. As such, define

$$\text{CATE}(X) = \text{E}[Y(1) - Y(0)|X]$$

as the conditional average treatment effect. If unconfoundedness is true, then

$$\text{E}[Y(1)|T = 1, X = x] = \text{E}[Y(1)|T = 0, X = x] = \text{E}[Y(1)|X = x]$$

$$\text{E}[Y(0)|T = 1, X = x] = \text{E}[Y(0)|T = 0, X = x] = \text{E}[Y(0)|X = x]$$

meaning that we can use the following estimator:

$$\text{CATE}(X) = \text{E}[Y(1)|T = 1, X = x] - \text{E}[Y(0)|T = 0, X = x]$$

Moreover, if we assume that there is overlap in the population by the characteristic X (i.e. for any $x \in X$, some people have been treated and some not), then we can also write

$$\text{ATE} = \text{E}[\text{CATE(X)}]$$

meaning that with unconfoundedness and overlap, we can estimate the ATE by averaging the CATE over the X.

Now consider the true model $Y = Y(0)+[Y(1)-Y(0)]\cdot T$ and define $a = \text{E}[Y(0)]$ and $b = \text{E}[Y(1) - Y(0)]$. Then, you can rewrite the model as:

$$Y = a + u_a + (b + u_b) \cdot T$$

where u_a, u_b are mean-zero errors. This reduces to

$$Y = a + bT + (u_a + u_b T)$$

If we consider the last term in parentheses as as an error term, can we estimate the model by OLS? To verify that, let's check the Gauss-Markov assumptions.

Define $e = (u_a + u_b T)$, then:

- $\mathrm{E}[e] = \mathrm{E}[u_a + u_b T] = \mathrm{E}[u_a] + \mathrm{E}[u_b T]$
- $\mathrm{E}[eT] = \mathrm{E}[(u_a + u_b T)T] = \mathrm{E}[u_a T] + \mathrm{E}[u_b T]$

If both terms are equal to 0, then the model satisfies GM assumptions and we can perform OLS on it. This implies that $\mathrm{E}[u_a T] = \mathrm{E}[u_b T] = 0$.

If $\mathrm{E}[u_a T] = 0$, then:

$$\begin{aligned}
\mathrm{E}[u_a T] = 0 &\Leftrightarrow \mathrm{E}[(Y(0) - \mathrm{E}[Y(0)]\, T] = 0 \\
&\Leftrightarrow \mathrm{E}[Y(0)T] - \mathrm{E}[\mathrm{E}[Y(0)]\, T] = 0 \\
&\Leftrightarrow \mathrm{E}[Y(0)|T = 1]\, \Pr[T = 1] - \mathrm{E}[Y(0)]\, \mathrm{E}[T] = 0 \\
&\Leftrightarrow \mathrm{E}[Y(0)|T = 1]\, \mathrm{E}[T] - \mathrm{E}[Y(0)]\, \mathrm{E}[T] = 0 \\
&\Leftrightarrow \mathrm{E}[T] \cdot [\mathrm{E}[Y(0)|T = 1] - \mathrm{E}[Y(0)]] = 0 \\
&\Leftrightarrow \underbrace{\mathrm{E}[Y(0)|T = 1] = \mathrm{E}[Y(0)]}_{\text{Mean Unconditional Unconfoundedness}}
\end{aligned}$$

This means that OLS also depends on the MUU property of the data!

However, this last result implied a known specification for the effect of the treatment on outcome. In fact, T interacts with Y in a linear manner, but what if it is not the case? We look at three different ways to estimate the relationship between Y and T.

Nonparametric regression

Suppose the relationship between Y and T is not known, you can run a kernel regression on it as: $Y = \hat{m}(T, X) + \mathrm{error}$. Then, you can evaluate the nonpara-

metric ATE as:

$$\widehat{\text{ATE}} = \frac{1}{n}\sum_{i=1}^{n}[\hat{m}(1, X) - \hat{m}(0, X)]$$

Matching estimators

An alternative to the OLS or the nonparametric regression is to find a matching algorithm such that for each individual i in the treatment group, you look for a or multiple individuals j in the control group such that for a given characteristic X, we have $X_i \approx X_j$. Then the ATE is defined for each individual in the treatment group and the estimator for the ATE is the average of the difference in outcome for all i and j matched. Formally,

$$\text{ATE} = \text{E}\left[Y_i - Y_j\right]$$

$$\widehat{\text{ATE}} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - Y_j)$$

In such a setting, the estimation follows three distinct steps:

1. Draw values of x randomly from the population distribution of X.
2. Randomly draw an individual i in the treatment group such that $x_i = x$ and match him with a randomly drawn individual j in the control group such that $x_j \approx x_i$.
3. Calculate $Y_i - Y_j$.

Then, you repeat this process many times and average the difference in outcomes to get the estimator for the ATE.

Propensity score

A different approach to matching would be to look at the probability of being in the treatment group, conditional on some characteristic X. In this situation, we

define this probability as the propensity score. Formally, we define the propensity score of an individual $P(X)$ such that

$$P(X) = \mathrm{E}\left[T|X\right] = \Pr\left[T = 1|X\right]$$

The intuition behind this value is that if the outcome is independent of the group an individual is in, conditional on the characteristic X, then it must also be independent on the treatment while conditioning on the propensity score. Therefore, you could write the outcome as a nonparametric function of T and $P(X)$ instead of X:

$$Y = \hat{m}(T, P(X)) + \text{error}$$

Then the process of estimating the ATE relies also on three steps:

1. Estimate the function $P(X)$ (can be done with a known model or again by nonparametric regression if the process of determining T is unknown).
2. Estimate Y using the nonparametric form with the estimated P, that is $Y = \hat{r}(T, \hat{P}(X))$.
3. The ATE is given by:

   $$\text{ATE} = \mathrm{E}\left[\hat{r}(1, \hat{P}(X)) - \hat{r}(0, \hat{P}(X))\right]$$

   while the estimator for it is

   $$\widehat{\text{ATE}} = \frac{1}{n}\sum_{i=1}^{n}[\hat{r}(1, \hat{P}(X)) - \hat{r}(0, \hat{P}(X))]$$

The remaining question is why would you use $P(X)$ instead of X directly? Well, that depends on the specification of $P(\cdot)$. If it is known very well (probit or logit model), then you can find easily the value for ATE.

Propensity score matching

Finally, the propensity score matching process is trivially the combination of both previous algorithms in the sense that instead of matching a particular characteristic X, you try and match the treatment group with the control group by their similarity in $P(X)$.

13.3 Local Average Treatment Effect (LATE)

What if the assignment to either the treatment or the control group is not perfectly random. That is, what if T is not independent to Y. Then, all of our previous results do not hold, and we need to put in more work to find a suitable solution. As we did in the OLS case when e and X were correlated, let's try to find a binary instrument Z which is randomly assigned.

We define four main types of individuals:

- The compliers are individuals for which $T = Z$ as observed BUT ALSO if T was different, we would get $T = Z$ anyway.
- The deniers are individuals for which $T \neq Z$ as observed BUT ALSO if T was different, we would get $T \neq Z$ anyway.

Chapter 14

Regression Discontinuity Design

to be continued...