ECON7772 - Econometric Methods

Lecture Notes from Arthur Lewbel's lectures

Paul Anthony Sarkis
Boston College

Contents

1 Properties of Estimators **5**

1.1 Review 5

1.2 Finite Sample Properties of Estimators 9

1.3 Asymptotic Properties of Estimators 13

2 Classical Regression **24**

2.1 Introducing the OLS Estimator 24

2.2 Gauss-Markov Theorem 27

2.3 Finite sample properties of the OLS estimator 31

2.4 Asymptotic properties of the OLS estimator 34

3 Specification issues **36**

3.1 Non-randomness of X 36

3.2 Non-stationarity of X 38

3.3 High correlation in the error term 38

3.4 Collinearity 38

3.5 Coefficient interpretation 39

4 Maximum Likelihood Estimation 42

4.1 Basic assumptions 43

4.2 Properties of the ML estimator 44

4.3 Application of MLE to Binary Choice models 49

5 Inference and Hypothesis Testing 51

5.1 Review 51

5.2 Univariate tests 53

5.3 Multivariate tests 55

5.4 Likelihood Ratio tests 57

5.5 Lagrange Multiplier tests 57

6 Generalized Least-Squares and non-iid errors 58

6.1 Heteroskedasticity 58

6.2 Autocorrelation 63

7 Dynamic models and Time Series models 69

7.1 Dynamic Regression Models 69

7.2 Simple Distributed Lag Models 71

7.3 Autoregressive Distributed Lag Models 71

7.4 Issues with Dynamic Models 71

8 Instrumental Variables, 2SLS, Endogeneity and Simultaneity 72

8.1 Correlation between errors and regressors 72

8.2 Measurement errors 73

8.3 Instrumental variables 74

8.4 Multiple IVs and 2SLS 75

8.5 Testing IVs 76

8.6 Simultaneity 78

9 Non-linear models, GMM and extremum estimators 81

9.1 Nonlinear Least Squares 81

9.2 Extremum Estimators 84

9.3 Generalized Method of Moments 85

10 Non-parametric estimators 90

10.1 Introduction 90

10.2 Estimation of the EDF 91

10.3 Kernel Density Estimation 92

10.4 Kernel Regression Estimation 99

11 Program Evaluation and Treatment Effects 107

11.1 Intuition 107

11.2 Identification 109

12 Regression Discontinuity Design **119**

Chapter 1

Properties of Estimators

1.1 Review

1.1.1 Definitions

Throughout this section, we'll define X and 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 1.1. *A random vector X of size j is a vector consisting of j random variables $(X_1, ..., X_j)$. Its expectation, $\mathrm{E}[X]$ is the vector consisting of the expectations of all its elements, namely $(\mathrm{E}[X_1], ..., \mathrm{E}[X_j])$.*

Definition 1.2. *The variance matrix of a random vector X, denoted $\mathrm{Var}[X]$ is the $j \times j$ matrix equal to $\mathrm{E}[(X - \mathrm{E}[X])(X - \mathrm{E}[X])']$*

Definition 1.3. *The covariance $\mathrm{Cov}(X, Z)$ between two vectors X and Z is equal to $E\left[(X - \mathrm{E}[X])(Z - \mathrm{E}[Z])'\right]$*

Proposition 1.1. *The expectation of the vector $AX + b$ is*

$$\mathrm{E}[AX + b] = A\,\mathrm{E}[X] + b$$

The variance of the vector $AX + b$ is

$$\operatorname{Var}[AX + b] = A \operatorname{Var}[X] A'$$

The covariance between vectors $AX + b$ and $CZ + d$ is

$$\operatorname{Cov}(AX + b, CZ + d) = A \operatorname{Cov}(X, Z) C'$$

Definition 1.4. *The vector X follows a joint distribution denoted $F(\cdot)$. If $F(\cdot) = N(\mu, \Omega)$, we say that that X follows a multi-variate normal distribution of mean μ and variance matrix Ω.*

The mean μ is the vector $\mathrm{E}[X]$ while the variance matrix Ω is the matrix containing variances and covariances of all elements of X, such that:

$$\Omega = \begin{bmatrix} \operatorname{Var}[X_1] & \operatorname{Cov}(X_1, X_2) & \dots & \operatorname{Cov}(X_1, X_j) \\ \operatorname{Cov}(X_2, X_1) & \operatorname{Var}[X_2] & \dots & \operatorname{Cov}(X_2, X_j) \\ \vdots & \vdots & \ddots & \vdots \\ \operatorname{Cov}(X_j, X_1) & \dots & \dots & \operatorname{Var}[X_j] \end{bmatrix}$$

Proposition 1.2. *If X follows a multi-variate normal distribution of mean μ and variance matrix Ω, then the vector $AX + b$ also follows a joint multi-variate normal distribution, of mean $A\mu + b$ and variance $A\Omega A'$.*

Proposition 1.3. *If X follows a multi-variate normal distribution of mean μ and variance matrix Ω, then the "standardized" vector $(X - \mu)'\Omega^{-1}(X - \mu)$ follows a chi-squared distribution with j degrees of freedom; we write*

$$(X - \mu)'\Omega^{-1}(X - \mu) \sim \chi^2_j$$

1.1.2 Differentiation

Definition 1.5. *The derivative of a vector X by a scalar a, denoted $\frac{\partial X}{\partial a}$, is the vector consisting of element-wise derivatives with respect to a:*

$$\frac{\partial X}{\partial a} = \begin{bmatrix} \frac{\partial X_1}{\partial a} & \frac{\partial X_2}{\partial a} & \dots & \frac{\partial X_j}{\partial a} \end{bmatrix}'$$

Definition 1.6. *The derivative of a vector X by a vector $\mathbf{Y}$, denoted $\frac{\partial X}{\partial \mathbf{Y}}$, is a matrix consisting of element-wise derivatives such that:*

$$\frac{\partial X}{\partial \mathbf{Y}} = \begin{bmatrix} \frac{\partial X_1}{\partial Y_1} & \frac{\partial X_1}{\partial Y_2} & \cdots & \frac{\partial X_1}{\partial Y_k} \\ \frac{\partial X_2}{\partial Y_1} & \frac{\partial X_2}{\partial Y_2} & \cdots & \frac{\partial X_2}{\partial Y_k} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial X_j}{\partial Y_1} & \frac{\partial X_j}{\partial Y_2} & \cdots & \frac{\partial X_j}{\partial Y_k} \end{bmatrix}$$

Some properties of derivatives

If A is a matrix and not a function of the vector X, then:

$$\frac{\partial AX}{\partial X} = A$$

$$\frac{\partial X'AX}{\partial X} = 2AX$$

If a is a vector and not a function of the vector X, then:

$$\frac{\partial a'X}{\partial X} = a$$

$$\frac{\partial X'aX}{\partial X} = (X + X')a$$

$$\frac{\partial \operatorname{E}[g(a'X)]}{\partial X} = \operatorname{E}\left[\frac{\partial g(a'X)}{\partial X}\right]$$

1.1.3 Law of iterated expectations

For any two random variables X and Z, we have that:

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

1.1.4 Independence(s) and correlation

Let X and Z be any two random vectors with means μ_X and μ_Z resp.

Definition 1.7 (Independence)**.** *X and Z are said to be independent, denoted $X \perp Z$, if nothing can be said about X's distribution from Z. This also implies that:*

$$f(x, z) = f_X(x) \cdot f_Z(z)$$

Definition 1.8 (Mean-independence)**.** *X and Z are said to be mean-independent if*

$$\mathrm{E}[X|Z] = \mathrm{E}[X] = \mu_X$$

Proposition 1.4 (Independence to mean-independence)**.** *If X and Z are independent, then they are mean-independent. The converse is not true.*

Proof.

$$\begin{aligned}\mathrm{E}[X|Z] = \int x f_{X|Z}(x)\mathrm{d}x = \int x \frac{f_{XZ}(x,z)}{f_Z(z)}\mathrm{d}x &= \int x \frac{f_X(x) f_Z(z)}{f_Z(z)}\mathrm{d}x \\ &= \int x f_X \mathrm{d}x \\ &= \mu_X\end{aligned}$$

□

Definition 1.9 (Linear independence)**.** *X and Z are said to be uncorrelated, or linearly independent, if* $\mathrm{E}[XZ] = \mu_X \mu_Z \Leftrightarrow cov(X, Z) = 0$

Proposition 1.5 (Mean-independence to linear independence)**.** *If X and Z are mean-independent, then they are linearly independent. The converse is not true.*

Proof.

$$\begin{aligned}\mathrm{E}[XZ] = \mathrm{E}[\mathrm{E}[XZ|Z]] = \mathrm{E}[Z\,\mathrm{E}[X|Z]] &= \mathrm{E}[Z\,\mathrm{E}[X]] \\ &= \mathrm{E}[Z]\,\mathrm{E}[X] = \mu_X \mu_Z\end{aligned}$$

□

1.2 Finite Sample Properties of Estimators

An estimator is a rule used for calculating an estimate of a given moment of a population (say the mean, the effect of a variable on another, etc) using only observed data. A good estimator is one that is "close" to the real moment underlying the data. What we mean by "close" is not a set-in-stone definition, as we will see later.

1.2.1 Bias

One straightforward "closeness" relationship is bias. The definition of bias relies on the distance between the expected value of an estimator and the true value of the parameter.

Definition 1.10 (Bias of an estimator)**.** *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.*

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

We say that an estimator is unbiased if and only if its bias is equal to 0.

For example, if one wanted to estimate the expected value of a sequence of random variables, one would look at the average realization of these variables. But is this a good estimator in terms of bias, as it turns out, it is.

Proposition 1.6 (Sample average as an unbiased estimator for the unconditional mean)**.** *Let $Z_1, Z_2, ...$ be a sequence of n i.i.d. random variables such that, for all i, $\text{E}[Z_i] = \mu$. Consider X_n, the sample average of all n Z_i variables, or formally,*

$$X_n = \frac{1}{n}\sum_{i=1}^{n} Z_i$$

The sample average is an unbiased estimator of the mean of Z_i.

Proof. $E\frac{\sum_{i=1}^{n} Z_i}{n} = \frac{1}{n}\sum_{i=1}^{n}\text{E}[Z_i] = \frac{n\mu}{n} = \mu$ □

1.2.2 Variance

We might also be interest in estimating the variance of the sequence of variables. If one takes the sample variance, could that also be an unbiased estimator? In this case, it is not. However, by considering the "adjusted" sample variance, then we get an unbiased estimator.

Proposition 1.7 (Sample variance as a biased estimator for the variance)**.** *Consider the previously defined sequence of i.i.d. random variables $\{Z_i\}$ such that, for all i, $E[Z_i] = \mu$ and $Var[Z_i] = \sigma^2$. Let $\hat{\sigma}_n$ be the sample variance and $\hat{s}_n$ be the "adjusted" sample variance. Formally,*

$$\hat{\sigma}_n = \frac{1}{n}\sum_{i=1}^{n}(Z_i - X_n)^2$$

$$\textit{and } \hat{s}_n = \frac{1}{n-1}\sum_{i=1}^{n}(Z_i - X_n)^2$$

The regular sample variance is a biased estimate for the population variance σ^2. In contrast, the "adjusted" sample variance is an unbiased estimator of σ^2.

Proof. □

Bias might not be a complete description of the performance of an estimator. In fact, while in expectation an unbiased estimator is a good estimate, the actual realizations of the estimator might not be close enough. In order to measure how far, on average, the collection of estimates are from their expected value, we define the variance of an estimator.

Definition 1.11 (Variance of an estimator)**.** *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 variance of an estimator to be the expected value of the squared sampling deviations.*

$$\operatorname{Var}\left[\hat{\theta}\right] = \mathrm{E}\left[\left(\hat{\theta} - \mathrm{E}\left[\hat{\theta}\right]\right)^2\right]$$

Proposition 1.8. *Let $Z_1, Z_2, \ldots$ 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 Z, divided by the sample size :

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

Proof.

$$\begin{aligned}
\operatorname{Var}[X_n] &= \mathrm{E}\left[(X_n - E[X_n])^2\right] = \mathrm{E}\left[(X_n - \mu)^2\right] = \mathrm{E}\left[\left(\frac{1}{n}\sum_{i=0}^{n} Z_i - \mu\right)^2\right] \\
&= \frac{1}{n^2}\mathrm{E}\left[\left(\sum_{i=0}^{n} Z_i - \mu\right)\left(\sum_{j=0}^{n} Z_j - \mu\right)\right] \\
&= \frac{1}{n^2}\mathrm{E}\left[\sum_{i=0}^{n}\sum_{j=0}^{n}(Z_i - \mu)(Z_j - \mu)\right] \\
&= \frac{1}{n^2}\mathrm{E}\left[\sum_{i=0}^{n}(Z_i - \mu)^2 + \sum_{i=0}^{n}\sum_{j\neq i}^{n}(Z_i - \mu)(Z_j - \mu)\right] \\
&= \frac{1}{n^2}\left(\sum_{i=0}^{n}\operatorname{Var}[Z_i] + \sum_{i=0}^{n}\sum_{j\neq i}^{n}\operatorname{Cov}(Z_i, Z_j)\right) \\
&= \frac{n\operatorname{Var}[Z]}{n^2} = \frac{\sigma^2}{n}
\end{aligned}$$

□

1.2.3 Efficiency

Using the variance of estimators, we can compare different unbiased estimators based on how far we can expect them to be from their expected value.

Definition 1.12 (Efficiency of estimators)**.** *Among a number of estimators of the same class, the estimator having the least variance is called an efficient estimator. The lower bound of the variance of an estimator is called the Cramer-Rao bound.*

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 $\hat{\theta}_2$.

As an alternative to simple variance, one can use the mean squared error as a measure of efficiency.

Definition 1.13 (Mean Squared Error)**.** *Let θ_0 be the true value of a parameter and $\hat{\theta}$ be an estimator of this value. We define the mean squared error, or MSE, as the expectation of the squared deviation between the estimator and the true value of the estimand. Formally,*

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

Among estimators of the same class, an estimator with low MSE is more efficient than an estimator with high MSE.

Proposition 1.9 (MSE as a trade-off between bias and variance)**.** *For any estimator $\hat{\theta}$, we have that:*

$$\mathrm{MSE}(\hat{\theta}) = \mathrm{Var}\left[\hat{\theta}\right] + [\mathrm{Bias}(\hat{\theta})]^2$$

Proof.

$$\begin{aligned}
\mathrm{MSE}(\hat{\theta}) &= \mathrm{E}\left[(\hat{\theta} - \theta)^2\right] \\
&= \mathrm{E}\left[\left(\hat{\theta} - \mathrm{E}\left[\hat{\theta}\right] + \mathrm{E}\left[\hat{\theta}\right] - \theta\right)^2\right] \\
&= \mathrm{E}\left[\left(\hat{\theta} - \mathrm{E}\left[\hat{\theta}\right]\right)^2 + \left(\mathrm{E}\left[\hat{\theta}\right] - \theta\right)^2 + 2\left(\hat{\theta} - \mathrm{E}\left[\hat{\theta}\right]\right)\left(\mathrm{E}\left[\hat{\theta}\right] - \theta\right)\right] \\
&= \mathrm{Var}\left[\hat{\theta}\right] + \mathrm{E}\left[[\mathrm{Bias}(\hat{\theta})]^2\right] + 2 \cdot \mathrm{E}\left[\left(\hat{\theta} - \mathrm{E}\left[\hat{\theta}\right]\right)\left(\mathrm{E}\left[\hat{\theta}\right] - \theta\right)\right] \\
&= \mathrm{Var}\left[\hat{\theta}\right] + [\mathrm{Bias}(\hat{\theta})]^2 + 2 \cdot \mathrm{E}\left[\hat{\theta}\,\mathrm{E}\left[\hat{\theta}\right] - \hat{\theta}\theta - \mathrm{E}\left[\hat{\theta}\right]^2 + \mathrm{E}\left[\hat{\theta}\right]\theta\right] \\
&= \mathrm{Var}\left[\hat{\theta}\right] + [\mathrm{Bias}(\hat{\theta})]^2
\end{aligned}$$

□

1.3 Asymptotic Properties of Estimators

1.3.1 Convergence

Convergence in Mean Square

Definition 1.14 (Convergence in MSE). *Let $\{X_n\}$ denote a sequence of random variables such that $\mathrm{E}[X_i] = \mu_i$ and $Var[X_i] = \sigma_i^2$, and c a real number. If*

$$\lim_{n\to\infty} \mathrm{MSE}(X_n) = \lim_{n\to\infty} \mathrm{E}\left[(X_n - c)^2\right] = 0,$$

we say that the sequence $\{X_n\}$ converges in mean squared error to c and we write

$$X_n \xrightarrow{ms} c$$

Proposition 1.10 (Sample average convergence in MSE). *Let $\{Z_n\}$ be a sequence of i.i.d. random variables with mean $\mathrm{E}[Z_i] = \mu$ and variance $\mathrm{Var}[Z_i] = \sigma^2$, for all i. Consider X_n, the sample average as defined in the previous sections. We have that*

$$X_n \xrightarrow{ms} \mathrm{E}[Z_i] = \mu$$

In words, the sample average converges in mean squared error to its expected value.

Proof. Recall from the previous sections that $\mathrm{E}[X_n] = \mu$ since it is an unbiased estimator, and that $\mathrm{Var}[X_n] = \frac{\sigma^2}{n}$.

$$\begin{aligned}
\lim_{n\to\infty} \mathrm{E}\left[(X_n - \mu)^2\right] &= \lim_{n\to\infty} \mathrm{E}\left[(X_n - \mathrm{E}[X_n])^2\right] \\
&= \lim_{n\to\infty} \mathrm{Var}[X_n] \\
&= \lim_{n\to\infty} \frac{\sigma^2}{n} = 0
\end{aligned}$$

□

Convergence in Probability

Definition 1.15 (Convergence in probability). *Let $\{X_n\}$ denote a sequence of random variables and c a real number. If for all $\epsilon > 0$,*

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

we say that X_n converges in probability to c and we write

$$X_n \xrightarrow{p} c$$

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

Proposition 1.11. *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 $\lim_{n\to\infty} \Pr[|X_n - c| > \epsilon] = 0$ □

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.

Convergence in Distribution

Definition 1.16 (Convergence in distribution)**.** *Let $\{X_n\}$ denote a sequence of random variables following distribution F_{X_n} and X a random variable with distribution F_X. If,*

$$\lim_{n\to\infty} F_{X_n}(x) = F_X(x) \text{ for all } x,$$

we say that X_n converges in distribution to X and we write

$$X_n \xrightarrow{d} X$$

Proposition 1.12. *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.*

1.3.2 Consistency

As we have seen in the previous section, convergence can be used to show how close to a parameter a sequence can get. This type of measurement can be interesting to compare estimators and their estimand.

Definition 1.17 (Consistent estimator)**.** *Let $\hat{\theta}$ be an estimator of a parameter θ, we say that $\hat{\theta}$ is a consistent estimator if $\hat{\theta} \xrightarrow{p} \theta$.*

Note that this definition of consistency has no relationship whatsoever to bias. In fact, it is possible to find unbiased and consistent estimator, as it is possible to find biased and consistent estimators or unbiased and inconsistent estimators. Thus, while consistency might be an interesting property, it need to be treated independently of bias.

1.3.3 Law of Large Numbers

Theorem 1.1 (Weak Law of Large Numbers)**.** *Let $\{Z_n\}$ denote a sequence of i.i.d. 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 proved that $X_n \xrightarrow{\text{ms}} \mu$. Moreover, we showed that m.s. convergence implied convergence in probability, thus we also have that $X_n \xrightarrow{\text{p}} \mu$. □

Theorem 1.2 (Khinchin's WLLN)**.** *Let $\{Z_n\}$ denote a sequence of i.i.d. 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 powerful in the sense that they show that for any sequence of i.i.d. random variables having a finite variance or finite expected absolute value, the sample average associated will converge in probability to the true mean of the random variables. Nonetheless, these theorems need that the sequence of $\{Z_n\}$ is i.i.d..

Theorem 1.3. *Let $\{Z_n\}$ denote a sequence of random variables such that:*

- $\mathrm{E}[Z_i] = \mu_i$,
- $\mathrm{Var}[Z_i] = \sigma_i^2$ *and*
- $\mathrm{Cov}(Z_i, Z_j) = \sigma_{i,j}$ *for all* $i \neq j$

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 $\lim_{n\to\infty}\mathrm{Var}[X_n] = 0$, 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 i.i.d.) 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 $\mathrm{Var}[X_n] = \frac{1}{n^2}\sum\mathrm{Var}[Z_i] + \frac{1}{n^2}\sum\sum\mathrm{Cov}(Z_i, Z_j)$

1.3.4 Slutsky's Theorems

Theorem 1.4 (Slutsky's Theorem for convergence in probability)**.** *For any continuous function $g(\cdot)$ that does not depend on the sample size n, we have:*

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

Theorem 1.5 (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:*

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

Proposition 1.13 (Properties of convergence). *Let $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{p} c$ where X is a random variable and c a constant. We have :*

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

Using Slutsky's theorem is quite useful to prove consistency of estimators.

Proposition 1.14 (Consistency of the OLS estimator). *The OLS estimate, as defined by:*

$$\hat{b}_{OLS} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}$$

is a consistent estimate of b in the model $y_i = b \cdot x_i + e_i$.

Proof.

$$\begin{aligned}
\operatorname{plim} \hat{b}_{OLS} = \operatorname{plim} \frac{\sum_{i=1}^{n} x_i (b \cdot x_i + e_i)}{\sum_{i=1}^{n} x_i^2} &= \operatorname{plim} \frac{\sum_{i=1}^{n} b \cdot x_i^2}{\sum_{i=1}^{n} x_i^2} + \operatorname{plim} \frac{\sum_{i=1}^{n} x_i e_i}{\sum_{i=1}^{n} x_i^2} \\
&= b + \frac{\operatorname{plim} 1/n \cdot \sum_{i=1}^{n} x_i e_i}{\operatorname{plim} 1/n \cdot \sum_{i=1}^{n} x_i^2} \\
&= b
\end{aligned}$$

□

1.3.5 Central Limit Theorems

Theorem 1.6 (Lindeberg-Lévy Central Limit Theorem). *Suppose $\{Z_n\}$ is a sequence of i.i.d. random variables with $\mathrm{E}[Z_i] = \mu$ and $\mathrm{Var}[Z_i] = \sigma^2 < \infty$ and X_n is the sample average of the first n elements of the sequence. Then, as n approaches infinity, the random variable $\sqrt{n}(X_n - \mu)$ converges in distribution to a normal distribution $N(0, \sigma^2)$. We write*

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

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

This theorem also holds if $\{Z_n\}$ is a sequence of random vectors of size k, then we'd have that

$$\sqrt{n}(X_n - \mu) \xrightarrow{d} N_k(0, \Omega)$$

where N_k denotes the multivariate normal distribution of size k and Ω is the variance matrix of any vector Z_i.

Theorem 1.7 (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 of Z_i as X_n, and the sample average of the variances σ_i^2 as $\bar{\sigma}_n^2$.*

If

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

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

1.3.6 Delta Method

The delta method is a result concerning the asymptotic distribution of a function of an asymptotically normal estimator. In other words, it is used to recover the asymptotic distribution of a function of an estimator, provided that we know the asymptotic distribution of this estimator.

Proposition 1.15 (Univariate Delta Method)**.** *Consider a sequence of random variables $\{X_n\}$ such that:*

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

then, for any function $g(x)$ such that $g(x)$ is not a function of sample size n, its derivative $g'(x)$ exists and is non-zero valued, we have that:

$$\sqrt{n}(g(X_n) - g(\mu)) \xrightarrow{d} N(0, \sigma^2 \cdot [g'(\mu)]^2)$$

Proposition 1.16 (Multivariate Delta Method)**.** *Consider a sequence of random vectors $\{X_n\}$ of size k such that:*

$$\sqrt{n}(X_n - \mu) \xrightarrow{d} N_k(0, \Omega),$$

then, for any scalar-valued function $h(x)$ such that $h(x)$ is not a function of sample size n, its derivative $h'(x)$ exists and is non-zero valued, we have that:

$$\sqrt{n}(g(X_n) - g(\mu)) \xrightarrow{d} N_k\left(0, \frac{\partial h}{\partial \mu'}\Omega\frac{\partial h'}{\partial \mu}\right)$$

Next, we cover an example on how to use this method.

Consider the estimator X_n such that $\sqrt{n}(X_n - a) \xrightarrow{\text{d}} N(0, 1)$ and the function $g(x) = x^2$. First, by Slutsky's theorem, we can write that:

$$\sqrt{n}(X_n - a) \cdot \sqrt{n}(X_n - a) \xrightarrow{\text{d}} X^2$$

$$\text{or equivalently } n(X_n - a)^2 \xrightarrow{\text{d}} \chi_1^2$$

where $X \sim N(0, 1)$.

By the delta method, we have that:

$$\sqrt{n}(X_n^2 - a^2) \xrightarrow{\text{d}} N(0, 1 \cdot (2a)^2) = N(0, 4a^2)$$

1.3.7 Asymptotic Notation

In order to go further in our discussion of convergence and other asymptotic properties, we need to define another type of notation called asymptotic notation. In particular, we will extend existing notation for asymptotic convergence and boundedness to allow for stochastic processes.

Definition 1.18 (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 a sequence C_n converges to 0 at a rate equivalent to the function inside the operator (1 or n^k).

Definition 1.19 (Little-o_p notation)**.** *Let $\{X_n\}$ be a sequence of random variables.*

We say that $X_n = o_p(1)$ if, for all $\varepsilon > 0$ and $\delta > 0$, there exists an N for which $n > N$ implies:

$$\Pr\left[|X_n| > \varepsilon\right] < \delta$$

One could be tempted to draw the parallel with the property of convergence in probability since, by taking a δ arbitrarily close to 0, we can definitely say that:

$$\lim_{n\to\infty} \Pr\left[|X_n| > \varepsilon\right] < \delta$$

Thus, if $X_n \xrightarrow{p} 0$, we can always say that $X_n = o_p(1)$.

We can also extend the result to higher orders of o_p convergence:

- *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: $X_n = o_p(n^k)$.*

In this case the parallel with convergence in probability shows the extension of little-o convergence clearly. In fact, o_p notation defines the convergence at a rate equivalent to the function inside the operator, in probability only (not surely this time). In other words, it means that as n increases, the probability that X_n does not converge to 0 is getting lower and lower.

Definition 1.20 (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: $C_n = O(n^k)$.*

The intuition behind this notation is not anymore about convergence but more about boundedness. Big-O notation defines a sort of asymptotic boundedness, meaning that the sequence will be bounded after some point.

Definition 1.21 (Stochastic boundedness)**.** *Let $\{X_n\}$ be a sequence of random variables.*

We say that $X_n = O_p(1)$ if for all $\delta > 0$ and associated $K_\delta > 0$, there exists a N such that $n > N$ implies that:

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

We can also extend the result to higher orders of O_p convergence: X_n is $O_p(n^k)$ if $\frac{X_n}{n^k} = O_p(1)$, and we write: $X_n = O_p(n^k)$.

In the same way that o_p extended o notation, O_p is the stochastic extension of O notation. It means that, as n increases, the probability that X_n is not asymptotically bounded goes to 0.

Proposition 1.17 (Relation between o_p and O_p convergence)**.** *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 $X_n = O_p(n^k)$.*

Proof. This comes directly from the fact that a convergent sequence has to be bounded. □

1.3.8 Extremum Estimators

Definition 1.22 (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 1.8. *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 i.i.d. 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,

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

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

Theorem 1.9 (Glevenko-Cantelli Theorem)**.** *Let $\{Z_n\}$ be any sequence of i.i.d. random variables with cdf $F(\cdot)$. The observed cumulative distribution*

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

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

Chapter 2

Classical Regression

2.1 Introducing the OLS Estimator

2.1.1 Linear Model

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

$$Y = \beta_1 + \beta_2 X_2 + ... + \beta_k X_k + e$$

which can be written in its matrix form as:

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

$$\Leftrightarrow Y = X\beta + e$$

Definition 2.1 (OLS estimator). *The OLS estimator $\hat{\beta}$ of the true parameter β is*

the vector that minimizes the sum of squared residuals:

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

The FOC of this optimization problem gives:

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

The SOC is given by $2X'X$.

Thus, if the matrix given by $X'X$ *is invertible, then the value of the OLS estimator* $\hat{\beta}$ *is:*

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

2.1.2 Univariate OLS

Proposition 2.1 (Univariate Linear Regression)**.** *In the particular case of* $k = 2$ *so that* $y = a + bx + e$*. We have that:*

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

Consider the univariate model described just above 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 $\hat{a}$.

2.1.3 Fit of the model

Definition 2.2 (Fitted values and residuals)**.** *The fitted values of the model, denoted $\hat{Y}$ are defined by:*

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

These are not predictors of Y since they utimately are a function of the sample only (not the population) but they allow us to compute the residuals, which are useful for variance estimation, as we'll see later.

The residuals of the model, denoted $\hat{e}$, are defined as the difference between the sample values and the fitted values, formally,

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

They are different from the errors e which are unobservable parameters of the regression.

Definition 2.3 (R^2 and analysis-of-variance)**.** *We can measure the variance of the model with a variable called R^2. Write*

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

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*

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

Finally, we define as R^2 *the proportion of variation of* Y *that is also captured as a variation in* $\hat{Y}$ *(implying that we have a model for it):*

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

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.

2.2 Gauss-Markov Theorem

2.2.1 Assumptions of the linear model

Definition 2.4 (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, meaning that it would be the exact same if we took another sample of the population. Note that this assumption is quite strong but is not necessary. Indeed, if* X *is random, than assumption A4 will need to be conditional on* X*.*

***A3* Non-singularity of $X'X$:** *since the OLS estimator takes the inverse of $X'X$. For it to be non-singular, it must be that:*

- $n > k$*: there are more observations than explanatory variables (no over-idenfication), and*
- $\text{rank}(X) = k$ *(no multicollinearity in X)*

***A4* The errors are mean-zero and homoskedastic:** *in particular,*

$$\text{E}[e_i] = \text{E}[e] = 0 \text{ and } \text{Var}[e] = \Omega = \sigma^2 I_n$$

This property also means that there is no autocorrelation in the data $\text{Cov}(e_i, e_j) = 0$ *for all* $i \neq j$.

If the data is random, then we need that:

$$\text{E}[e_i|X_i] = \text{E}[e|X] = 0 \text{ and } \text{Var}[e|X] = \Omega = \sigma^2 I_n$$

Theorem 2.1 (Gauss-Markov Theorem)**.** *Under assumptions A1-A4, the OLS estimator $\hat{\beta}$ is the Best Linear Unbiased Estimator (BLUE). This property means that, among the class of linear unbiased estimators, the OLS estimator is the most efficient one.*

In order to prove this theorem, we will need to understand more about the general class of estimators that contain the OLS estimator.

2.2.2 Linear Unbiased Estimators

Definition 2.5 (Linear estimator)**.** *An estimator $\tilde{\beta}$ is said to be linear in the dependent variable if it can be written as a linear transformation of the dependent variable. Formally, it must be equal to a constant matrix multiplied by a random vector:*

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

Proposition 2.2 (Unbiased Linear Estimators)**.** *A linear estimator is unbiased if and only if its associated transformation matrix $\tilde{C}$ is such that*

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

Proof. Consider any linear estimator $\tilde{\beta} = \tilde{C}Y$, we have that:

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

□

Proposition 2.3 (Variance of Linear Estimators)**.** *The variance of a homoskedastic, non-autocorrelated linear estimator is given by* $\mathrm{Var}\left[\tilde{\beta}\right] = \sigma^2(\tilde{C}\tilde{C}')$.

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

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

□

We now have the tools to prove the Gauss-Markov theorem.

Consider, without loss of generality, an alternative linear estimator $\tilde{\beta} = \tilde{C}Y$ such that it is unbiased and $\tilde{C} = (X'X)^{-1}X' + D$. Since we assumed this estimator to be unbiased, we can write that:

$$\tilde{C}X = I_k \Leftrightarrow [(X'X)^{-1}X' + D]X = I_k \Leftrightarrow I_k + DX = I_k \Leftrightarrow DX = 0$$

Using this, we can find the variance of this estimator:

$$\begin{aligned}\mathrm{Var}\left[\tilde{C}Y\right] = \sigma^2 \cdot (\tilde{C}\tilde{C}') &= \sigma^2 \cdot [(X'X)^{-1}X' + D][(X'X)^{-1}X' + D]' \\ &= \sigma^2[(X'X)^{-1} + DD'] \geq \sigma^2(X'X)^{-1}\end{aligned}$$

Implying that the lowest variance achievable by a linear unbiased estimator will be equal to $\sigma^2(X'X)^{-1}$, the variance of the OLS estimator.

2.2.3 Other properties of linear unbiased estimators

Definition 2.6 (Projection matrix)**.** *Given a linear unbiased estimator $\tilde{\beta}$, we define the projection matrix, denoted P, as*

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

Proposition 2.4 (Properties of the projection matrix)**.** *The projection matrix has a few nice properties such as:*

- $PX = X$
- $P = P'$
- $PP = P$
- $\text{tr}(P) = k$
- $PY = \hat{Y}$

Proof. We have $PY = X\tilde{C}Y = X\tilde{\beta} = \hat{Y}$ □

Definition 2.7 (Orthogonal projection matrix)**.** *Given a linear unbiased estimator $\tilde{\beta}$, we define the orthogonal projection matrix, denoted M, as*

$$M = I_k - P$$

Proposition 2.5 (Properties of the projection matrix)**.** *The orthogonal projection matrix also has a few nice properties such as:*

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

Proof. □

2.3 Finite sample properties of the OLS estimator

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

Proposition 2.6 (Unbiasedness of OLS estimator)**.** *Under assumptions A1-A4, the OLS estimator $\hat{\beta}$ is unbiased.*

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

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

□

Now that we have found the expected value of $\hat{\beta}$, we will follow the previous chapter and look at its variance.

Proposition 2.7 (Variance of the OLS estimator)**.** *Under assumptions A1-A4, the variance of the OLS estimator $\hat{\beta}$ is given by:*

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

Proof. We know that $\hat{\beta} = \tilde{C}Y$ where $\tilde{C}$ is a function of X (thus a constant, or if X is random, a constant conditional on X). Therefore,

$$\begin{aligned}
\mathrm{Var}\left[\hat{\beta}\right] = \mathrm{Var}\left[\tilde{C}Y\right] = \tilde{C}\,\mathrm{Var}\left[Y\right]\tilde{C}' &= \sigma^2 \cdot \tilde{C}\tilde{C}' \\
&= \sigma^2 \cdot (X'X)^{-1}X'((X'X)^{-1}X')' \\
&= \sigma^2 \cdot (X'X)^{-1}
\end{aligned}$$

□

However, note that the variance of Y (or equivalently the variance of e) is unknown to the econometrician. Therefore, the variance of $\hat{\beta}$ cannot be computed. This might not seem to be an issue since we have only be interested in theoretical variances of estimators until now, but it will be a burden when we will try to perform inference analysis, hypothesis testing, etc. Thus, we cover how to estimate this variance in this section.

Definition 2.8 (Estimator of the error variance)**.** *Since the error term e has mean-zero, we can write its variance as $\sigma^2 = \mathrm{E}[ee']$. Using the Law of Large Numbers, we know that a consistent estimator of this object could be the sample average estimator given by:*

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

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 $\tilde{\sigma}^2 \geq \hat{\sigma}^2$.

Proposition 2.8 (Expected value of the variance estimator)**.** *Let $\hat{\sigma}^2$ be the sample moment estimator discussed in the previous definition. This estimator is biased as:*

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

Proof.

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

□

Definition 2.9 (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: $\mathrm{E}[s^2] = \sigma^2$*. 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 2.10 (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 $\hat{\beta} \sim N(\beta, \sigma^2(X'X)^{-1})$.

Proposition 2.9. *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,

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

Definition 2.11 (Interval estimation and Hypothesis testing)**.** *This last fact can be used in interval estimation as it implies that:*

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

Proposition 2.10 (Moments of the residuals)**.** *Let the residuals of the regression be $\hat{e} = Me$ as we've seen before. We have that:*

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

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

2.4 Asymptotic properties of the OLS estimator

Before going to asymptotic properties, we ignore the normality assumption of the error term. Recall that this assumption had nothing to do with unbiasedness, or it being BLUE, however, it allowed us to derive the distribution of β in finite samples. As we will see in this section, the normality assumption is not even needed to prove the consistency of the OLS estimator, nor its asymptotic distribution.

Proposition 2.11 (Consistency of $\hat{\beta}$)**.** *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. This implies that $\hat{\beta}$ is a consistent estimator of β.*

Proof. 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, $\hat{\beta} \xrightarrow{\text{ms}} \beta \Rightarrow \hat{\beta}$ is a consistent estimator of β. □

Proposition 2.12 (Root-n-consistency and asymptotic normality of the OLS estimator)**.** *If e_i is iid, then $X_i e_i$ is inid, so applying the Lindeberg-Fuller version of the Central Limit Theorem, we have that:*

$$\sqrt{n}(\hat{\beta} - \beta) \sim N\left(0, \frac{\sigma^2}{n} Q^{-1}\right)$$

We say that the OLS estimator is $\sqrt{n}$-CAN (consistent and asymptotically normal).

Proof. We have that:

$$\begin{aligned}\sqrt{n}(\hat{\beta}-\beta) = \sqrt{n}\cdot\left((X'X)^{-1}X'(X\beta+e)-\beta\right) &= \sqrt{n}\cdot\left((X'X)^{-1}X'e\right)\\ &= \sqrt{n}\cdot\left(\frac{X'X}{n}\right)^{-1}\frac{X'e}{n}\\ &= Q_n^{-1}\frac{1}{\sqrt{n}}X'e\end{aligned}$$

Now, we know that:

- $\mathrm{E}\left[X_i'e_i\right] = X_i'\,\mathrm{E}\left[e_i\right] = 0$, and
- $\mathrm{Var}\left[X_i'e_i\right] = \mathrm{E}\left[X_i'e_ie_i'X_i\right] = \sigma^2X_i'X_i < \infty$

Moreover, since:

$$\lim_{n\to\infty}\max_i\frac{\sigma^2X_i'X_i}{n\frac{1}{n}\cdot\sigma^2X'X} = 0$$

$$\text{and }\lim_{n\to\infty}\frac{1}{n}\cdot\sigma^2X'X < \infty$$

We can use the Lindeberg-Fuller Central Limit Theorem to get that:

$$\sqrt{n}\frac{\frac{X'e}{n}-0}{\sigma\frac{X'X}{n}} \xrightarrow{\mathrm{d}} N(0,1)$$

$$\left(\frac{X'X}{n}\right)^{-1}\cdot\sigma\cdot\frac{1}{\sqrt{n}}X'e \xrightarrow{\mathrm{d}} N(0,1)$$

$$Q_n^{-1}\cdot\frac{1}{\sqrt{n}}X'e \xrightarrow{\mathrm{d}} N(0,\sigma^2)$$

□

Chapter 3

Specification issues

3.1 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 $\mathrm{Var}\left[\sqrt{n}\frac{X'e}{n}\right]$:

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

3.2 Non-stationarity of X

3.3 High correlation in the error term

3.4 Collinearity

Definition 3.1 (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,*

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

Proposition 3.1 (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 3.2 (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.

3.5 Coefficient interpretation

3.5.1 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.

3.5.2 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).

3.5.3 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 ($\frac{X_t - X_{t-1}}{X_{t-1}}$).

3.5.4 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 4

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.

4.1 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 4.1 (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 4.2 (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 4.3 (Assumptions on MLE)**.** *In order to further analyze the MLE, let's describe a set of additional assumptions:*

A1. ***Random draws:*** *The sequence $\{Z_n\}$ is a sequence of n, i.i.d. random variables. The we can write the likelihood function as:*

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

A2. ***Unique true parameter:*** *There is a single "true" parameter denoted θ_0.*

A3. ***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.*

A4. Identification: *For all $\theta \in \Theta$ such that $\theta \neq \theta_0$, we have that,*

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

This assumption implies that there are no other values than θ_0 that yield the same FOC of the maximum likelihood problem.

A5. 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, for all $\theta \in \Theta$ and $Z_i \in \Omega_Z$, the support of Z.*

A6. Independence of the support: *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.*

4.2 Properties of the ML estimator

Proposition 4.1 (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:*

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

Proof. The proof of this proposition is straightforward as the natural logarithm function is a monotonic transformation. □

Definition 4.4 (Score function)**.** *The score function, denoted $s(Z|\theta)$ is defined as the gradient of the log-likelihood function of an observation Z, when differentiated*

wrt θ:

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

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

Proposition 4.2 (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 result is very important because, linked to assumption 4 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:

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

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

Definition 4.5. *The Hessian matrix of the log-likelihood function, denoted as $H(Z|\theta)$ is the derivative of the score function or equivalently, the second-order derivative of the log-likelihood function, for one observation Z.*

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

Proposition 4.3 (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,*

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

Proof. □

Definition 4.6 (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:*

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

We also define J_0 as:

$$J_0 = \frac{I_n(\theta_0)}{n}$$

Theorem 4.1 (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. In this proof, we only need to check the assumptions on consistency of extremum estimators. First, define the objective function, denoted $Q_n(\theta)$, as the average of the log-likelihood function:

$$Q_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \ln f(Z_i|\theta)$$

Recall that the ML estimator is the value $\hat{\theta}_{ML}$ that maximizes this objective function. First of all, note that the objective function has an existing plim, denoted Q_0 since, by Law of Large Numbers, $\operatorname{plim} Q_n(\theta) = \operatorname{E}\left[\ln f(Z|\theta)\right] \equiv Q_0(\theta)$. Now, we can go to the four conditions of consistency.

First, we need to satisfy identification. For that, we need that $Q_0(\theta)$ is uniquely maximized at θ_0. In this case we get:

$$\hat{\theta} = \arg\max Q_0(\theta)$$

yielding the following FOC:

$$\frac{\partial Q_0(\theta)}{\partial\theta} = 0 \Leftrightarrow \frac{\partial \operatorname{E}\left[\ln f(Z|\theta)\right]}{\partial\theta} = 0 \Leftrightarrow \operatorname{E}\left[\frac{\partial \ln f(Z|\theta)}{\partial\theta}\right] = 0 \Leftrightarrow \operatorname{E}\left[s(Z|\theta)\right] = 0$$

which is satisfied for θ_0 by Assumption 4 we made earlier. The SOC would be:

$$\frac{\partial^2 Q_0(\theta)}{\partial\theta\partial\theta'} = \operatorname{E}\left[H(Z|\theta)\right]$$

which, evaluated at θ_0 gives $\mathrm{E}\left[H(Z|\theta)\right] = -\operatorname{Var}\left[s(Z|\theta_0)\right] < 0$.

Second, the condition of compactness is satisfied by assumption.

Third, smoothness of the objective function in θ is also ensured by Assumption 5.

Finally, we can show that uniform convergence is satisfied using two facts. First, we have that:

$$|\frac{\partial Q_n(\theta)}{\partial \theta}| \xrightarrow{\mathrm{p}} \mathrm{E}\left[s(Z|\theta)\right]$$

which, in addition to the assumption that $|\frac{\partial Q_n(\theta)}{\partial \theta}|$ is bounded for any n, then we can write:

$$\sup_{\theta}|\frac{\partial Q_n(\theta)}{\partial \theta}| = C + o_p(1) = O_p(1)$$

Therefore, $\hat{\theta}_{ML} \xrightarrow{\mathrm{p}} \theta_0$. □

Theorem 4.2 (Asymptotic normality of the ML estimator)**.** *Suppose θ_{ML} is a consistent estimator of a parameter θ, following the previous theorem. Then,*

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, J_0^{-1})$$

Proof. Recall that:

$$\frac{1}{n}\sum_{i=1}^{n} s(Z_i|\hat{\theta}) = 0$$

From there, we use the mean value theorem expansion around $\tilde{\theta} \in [\theta_0, \hat{\theta}]$:

$$\frac{1}{n}\sum_{i=1}^{n}\left[s(Z_i|\theta_0) + \frac{\partial s(Z_i|\tilde{\theta})}{\partial \theta}(\hat{\theta} - \theta_0)\right] = 0$$

$$\sqrt{n}\frac{1}{n}\sum_{i=1}^{n} s(Z_i|\theta_0) + \sqrt{n}\frac{1}{n}\sum_{i=1}^{n} H(Z_i|\tilde{\theta})(\hat{\theta} - \theta_0) = 0$$

$$-\sqrt{n}\frac{1}{n}\sum_{i=1}^{n} s(Z_i|\theta_0) \cdot \left[\frac{1}{n}\sum_{i=1}^{n} H(Z_i|\tilde{\theta})\right]^{-1} = \sqrt{n}(\hat{\theta} - \theta_0)$$

And then we look at the asymptotic distributions of both elements separately.

First, using the Lindeberg-Lévy version of the Central Limit Theorem and the fact that $\mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n} s(Z_i|\theta_0)\right] = 0$, we get that:

$$\begin{aligned}\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} s(Z_i|\theta_0) - \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n} s(Z_i|\theta_0)\right]\right) &= \sqrt{n}\cdot\frac{1}{n}\sum_{i=1}^{n} s(Z_i|\theta_0) \\ &\xrightarrow{\mathrm{d}} N(0, \mathrm{Var}\left[s(Z|\theta_0)\right]) \\ &\xrightarrow{\mathrm{d}} N(0, J_0)\end{aligned}$$

Then, take the first-degree Taylor expansion for the term inside the bracket around $\bar{\theta} \in (\theta_0, \tilde{\theta})$:

$$\frac{1}{n}\sum_{i=1}^{n} H(Z_i|\tilde{\theta}) = \frac{1}{n}\sum_{i=1}^{n} H(Z_i|\theta_0) + \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{k} \frac{\partial H(Z_i|\bar{\theta})}{\partial \theta_j}(\bar{\theta}_j - \theta_0)$$

From the Law of Large Numbers, we know that:

$$\frac{1}{n}\sum_{i=1}^{n} H(Z_i|\theta_0) \xrightarrow{\mathrm{p}} \mathrm{E}\left[H(Z|\theta_0)\right]$$

And using the fact that $\tilde{\theta} \xrightarrow{\mathrm{p}} \theta_0$ (since it is inside $[\theta_0, \hat{\theta}]$), we can also write that $\bar{\theta} \xrightarrow{\mathrm{p}} \theta_0$ so that everything left is know to be $o_p(1)$. Thus, we have that:

$$\frac{1}{n}\sum_{i=1}^{n} H(Z_i|\tilde{\theta}) \xrightarrow{\mathrm{p}} \mathrm{E}\left[H(Z|\theta_0)\right] + o_p(1) \xrightarrow{\mathrm{p}} -J_0$$

Now, combining the two elements (using Slutsky's identities) we have that:

$$\begin{aligned}\sqrt{n}(\hat{\theta} - \theta_0) &\xrightarrow{\mathrm{d}} J_0^{-1}\cdot N(0, J_0) \xrightarrow{\mathrm{d}} N(0, J_0^{-1}) \\ \Leftrightarrow \hat{\theta} &\xrightarrow{\mathrm{d}} N\left(\theta_0, \frac{J_0^{-1}}{n}\right)\end{aligned}$$

□

However, as should be expected by now, there is no way to compute the variance of the ML estimator using only the data, since J_0 is a function of the true parameter θ_0: we will need to use our ML estimate to compute an estimator of J_0. In order to do this, one could use any of three equivalent methods:

- $\hat{J}_0 = -\bar{H} = -\frac{1}{n}\sum_{i=1}^{n} H(Z_i|\hat{\theta})$
- $\hat{J}_0 = \widehat{\mathrm{Var}}(s(Z|\hat{\theta})) = \frac{1}{n}\sum_{i=1}^{n} s(Z_i|\hat{\theta})s(Z_i|\hat{\theta})'$
- $\hat{J}_0 = \bar{H}\left(\widehat{\mathrm{Var}}(s(Z|\hat{\theta}))\right)\bar{H}$

4.3 Application of MLE to Binary Choice models

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:

$$\begin{aligned} s(X_i, \beta) = 0 &\Leftrightarrow \frac{\partial \ln f(Y_i|\beta)}{\partial \beta} = 0 \\ &\Leftrightarrow \left[\frac{Y_i}{F(X_i, \beta)} - \frac{(1 - Y_i)}{1 - F(X_i, \beta)}\right]\frac{\partial F(X_i, \beta)}{\partial \beta} = 0 \end{aligned}$$

We can also compute the information matrix

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

Chapter 5

Inference and Hypothesis Testing

5.1 Review

In the case of a linear regression model with iid 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 $\hat{e}_i$, even in finite samples (recall that this normality assumption is not need for asymptotic properties).

First, recall that $\hat{\beta} - \beta = (X'X)^{-1}X'e$, which is 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, using $\hat{e} = Me$, we have that

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

These two results can also give us the joint distribution of $\hat{\beta}$ and $\hat{e}$, in fact:

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

which, again, is a linear projection of e, thus we can guess its mean ($\mathrm{E}[Ae] = \mathrm{E}[e] = 0$) for any constant A and variance matrix ($\mathrm{Var}[Ae] = A\,\mathrm{Var}[e]\,A'$). And indeed, using the variance formulas, we find that $\hat{\beta} - \beta$ and $\hat{e}$ are uncorrelated (therefore $\hat{\beta}$ also is uncorrelated to $\hat{e}$):

$$\mathrm{Var}[Ae] = A\,\mathrm{Var}[e]\,A' = \sigma^2 AA' = \sigma^2 \cdot \begin{bmatrix} (X'X)^{-1} & 0 \\ 0 & M \end{bmatrix}$$

Finally, consider the adjusted sample variance estimator $s^2 = (n-k)^{-1}\sum_{i=1}^{n}\hat{e}$. We can write that:

$$(n-k)s^2 = \hat{e}'\hat{e} = (Me)'Me = e'M'Me = e'Me$$

Then, using the spectral decomposition of M, namely $M = H\Lambda H'$ 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 derived in this section (that will help us in the next) are:

- $\hat{\beta} \sim N(\beta, \sigma^2(X'X)^{-1})$
- $\hat{e} \sim N(0, \sigma^2 M)$
- $\hat{\beta}$ and $\hat{e}$ are independent
- $\frac{(n-k)s^2}{\sigma^2} \sim \chi^2_{n-k}$
- $\hat{\beta}$ and s^2 are independent

5.2 Univariate tests

In this section, we cover tests and inference that can be applied to a particular estimator, say the coefficient on a single covariate.

5.2.1 T-statistic

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

Definition 5.1 (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 5.2 (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 5.1 (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 $t_{n-k}(c) = 1 - \alpha/2$.

5.2.2 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 5.3 (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 5.4 (Coverage probability)**.** *The coverage probability is defined as* $\Pr\left[\beta \in \hat{C}\right] = 1 - \alpha$

Proposition 5.2 (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,*

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

5.3 Multivariate tests

Multivariate tests are useful compared to univariate in case the restrictions we want to test apply to multiple variables. For example, it could be that one would want to make sure that a set of multiple variables have their place in the model. For that purpose, we cover three test procedures:

5.3.1 Wald tests

Wald tests are all based on a simple result that states that, if W is a q-dimensional random vector following a normal $N(0, \Omega)$, then

$$W'\Omega^{-1}W \sim \chi^2_q$$

Linear Restrictions: F-statistic

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:

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

Note that in this case, β is a vector of q parameters to be tested at the same time. Using the result described in the introduction to Wald tests, we have:

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

However, σ^2 is unknown so we have to use the adjusted sample variance s^2, and derive the so-called F-statistic (which is really a multivariate version of the t-statistic):

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

This test statistic only requires estimation of $\hat{\beta}$, the unrestricted model estimate. When the value of the statistic is on the far right of the distribution, one can safely assume that the restriction is not valid, thus rejecting the test.

In particular, for a regression model with N observation, q linear restrictions and k regressors, the estimator for the F-statistic can be reduced to:

$$\hat{F} = \frac{\sum_{i=1} n \left[\hat{e}_{Ri}^2 - \hat{e}_{Ui}^2\right] / q}{\sum_{i=1} n \hat{e}_{Ui}^2 / (n-k)} \sim F_{q,n-k}$$

In this case, one would estimate both the restricted and unrestricted model, recover the residuals and perform the test. This test is one-sided.

Nonlinear Restrictions: Wald statistic

In the more general case in which we have an unrestricted estimator that is $\sqrt{n}$-CAN but we want to test a nonlinear restriction such as: $H_0 : g(\theta) = 0$ with $g(\cdot)$ being any differentiable function, we need another testing procedure. Based on the same result as before, we can now write that:

$$g(\hat{\theta}_U)' \operatorname{Var}\left[g(\hat{\theta}_U)\right]^{-1} g(\hat{\theta}_U) \xrightarrow{\text{d}} \chi_q^2$$

Note that we get a convergence in distribution result instead of a the usual result because we are using an estimate of θ rather than its true value under H_0. Then, using the delta method, we have that:

$$\operatorname{Var}\left[g(\hat{\theta}_U)\right] = \frac{\partial g}{\partial \theta} \operatorname{Var}\left[\hat{\theta}_U\right] \frac{\partial g'}{\partial \theta}$$

which allows us to write the final ideal Wald statistic as:

$$g(\hat{\theta}_U)' \left[\frac{\partial g}{\partial \theta} \operatorname{Var}\left[\hat{\theta}_U\right] \frac{\partial g'}{\partial \theta}\right]^{-1} g(\hat{\theta}_U) \xrightarrow{\text{d}} \chi_q^2$$

However, and as is usual now, we do not know the exact form of $\operatorname{Var}\left[\hat{\theta}_U\right]$ since we do not know σ^2, the variance of the error term. Using s^2 can nonetheless get us somewhere, since $s^2 \xrightarrow{\text{p}} \sigma^2$, then using Slutsky's theorem, we have $\widehat{\operatorname{Var}}\left[\hat{\theta}_U\right] \xrightarrow{\text{p}} \operatorname{Var}\left[\hat{\theta}_U\right]$, and finally:

$$g(\hat{\theta}_U)' \left[\frac{\partial g}{\partial \theta} \widehat{\operatorname{Var}}\left[\hat{\theta}_U\right] \frac{\partial g'}{\partial \theta}\right]^{-1} g(\hat{\theta}_U) \xrightarrow{\text{d}} \chi_q^2$$

5.4 Likelihood Ratio tests

The Likelihood Ratio (LR) test discussed in this section is another way to test for single or multiple, linear or nonlinear restrictions on a model. To perform this test, 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 partitioned regression model can be written as:

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

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 5.3 (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.

5.5 Lagrange Multiplier tests

Finally, the last test we cover in this section is called the Lagrange Multiplier test. Like the Wald test, this test can be used to test any restriction on the parameters, such that $H_0 : g(\hat{\theta}_R) = 0$ where g is differentiable and $\hat{\theta}_R$ solves the MLE problem. Then, following the same result as in the Wald test, we have that:

$$\frac{\partial \ln L(\hat{\theta}_R)}{\partial \theta} \cdot (I(\hat{\theta}_R))^{-1} \frac{\partial \ln L(\hat{\theta}_R)}{\partial \theta}' \sim \chi^2_q$$

Contrary to the Wald test, this test requires only the restricted estimation.

Chapter 6

Generalized Least-Squares and non-iid errors

In this chapter, the goal is to let go of two main assumptions that we made about the variance of the error term. Respectively, we will cover both issues of heteroskedasticity (when the error term does not have identical variance over observations) and autocorrelation (when error terms of different observations are correlated).

6.1 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. But does that mean that that our OLS estimator is completely useless? Next we will see how does this violation affects our OLS estimates.

6.1.1 OLS estimator

Recall that the OLS estimator is defined as:

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

First, we want to look at the bias of this estimator under heteroskedasticity. Very easily, we get:

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

since the violation of homoskedasticity does not change the mean-zero assumption.

Second, we want to look at its consistency. For that, we look at the limit of its variance:

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

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. This result is very important because it means that the OLS estimator will be consistent for well-behaved models, even if the Gauss-Markov assumptions are not satisfied.

6.1.2 Generalized Least-Squares estimator

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 behind "building" a new estimator follows from two elements: first, we want an estimator that takes into account the new form of the variance matrix (can use the extra information); second, since we know how to deal with homoskedastic models, we could transform the variance matrix into an identity matrix and somehow make our OLS estimator work. The Generalized Least-Squares (GLS) estimator does exactly those two things.

Let P be a matrix such that:

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

This implies

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

This is what we call the spectral decomposition of Ω. Now, this very simple procedure made the term Pe homoskedastic, thus by transforming the whole model by P, we get an easy-to-deal-with model that satisfies all Gauss-Markov assumptions. But what are the implications of transforming the whole model?

Let $Y^* = X^*\beta + e^*$ where starred variables are the true variables projected by matrix P (i.e. $Y^* = PY$). Using the OLS estimator on the modified model, we get:

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

As we did with the OLS estimator, let's look at the properties of this new estimator. Note that consistency follows directly from the transformation we made, so we only look at bias. We get that:

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

and:

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

Which will go to zero as the sample size increases, thus yielding a consistent estimator.

However, the limitation to this method is that we might not know the variance matrix Ω. Indeed, it might be that we only suspect heteroskedasticity but we do not know the form of it. In these cases, one would need to estimate Ω in order to get compute the GLS estimator. Formally, we say that this GLS estimator is not feasible, however, its functional form might give us indications on how to get a feasible GLS estimator in practice.

Weighted Least Squares

Suppose the true model is

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

where $\text{Var}[e_i] = \sigma^2 w_i^2$. In this context we can guess that $\text{Var}\left[\frac{e_i}{w_i}\right] = \sigma^2$ 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).

6.1.3 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. While rejection in this test will definitely imply heteroskedasticity, keep in mind that failing to reject the null in this test does not tell us any meaningful information about the error term.

6.1.4 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.

6.2 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 $j > 0$.

6.2.1 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 6.1 (Autocorrelation at lag s)**.** *For a given lag s, we write the autocorrelation in the error term as:*

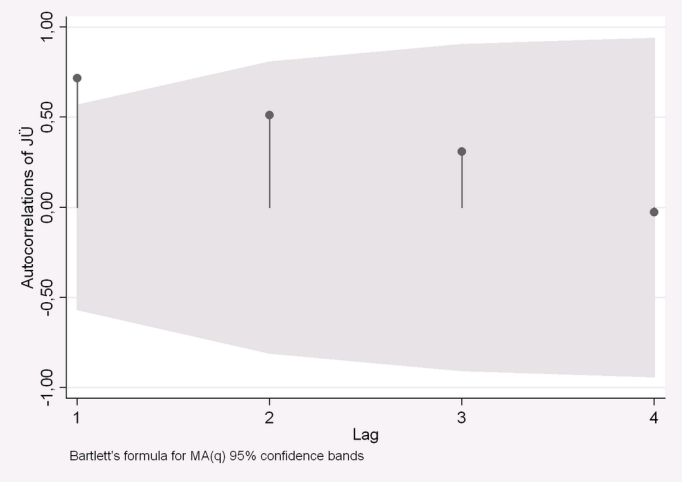
$$r_s = \frac{\operatorname{Cov}\left(e_t, e_{t-s}\right)}{\operatorname{Var}\left[e_t\right]}$$

Definition 6.2 (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 6.3 (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 $Q \xrightarrow{d} \chi^2_L$. 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.

6.2.2 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. Formally, we say that the error term follows an $\mathrm{AR}(1)$ process. 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 the $\mathrm{MA}(\infty)$ representation of the error term as:

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

and therefore, we can compute the first two moment of the error term:

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

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

The two last equations imply that the error term e_t is a homoskedastic mean-zero process, with autocorrelation being the only issue.

Using the two previous result, we get:

$$\mathrm{E}\left[e_t e_{t-s}\right] = \mathrm{E}\left[\left(\rho^s e_{t-s} + \sum_{k=1}^{\infty} \rho^{s+k} v_{t-s-k}\right) e_{t-s}\right] = \rho^s \sigma_e^2$$

In matrix form,

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

6.2.3 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.

6.2.4 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.

6.2.5 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:

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

Chapter 7

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).

7.1 Dynamic Regression Models

7.1.1 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 7.1 (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 7.2 (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 7.3 (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 7.4 (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:

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

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

7.1.2 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:

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

7.2 Simple Distributed Lag Models

7.3 Autoregressive Distributed Lag Models

7.4 Issues with Dynamic Models

Chapter 8

Instrumental Variables, 2SLS, Endogeneity and Simultaneity

8.1 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,

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

8.2 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,

$$\mathrm{plim}\, \hat{\beta} = \beta + \frac{-\beta \mathrm{Var}[v]}{\mathrm{Var}[X^*] + \mathrm{Var}[v]} = \beta \cdot \left(1 - \frac{\mathrm{Var}[v]}{\mathrm{Var}[X^*] + \mathrm{Var}[v]}\right)$$

There are two important issues about this result: first, it shows an asymptotic bias of our OLS estimator, in the sense that, even when we take the limit, the estimator is biased ; second, the bias is a downward bias (decreasing the value of β) and is positively correlated with β (the bigger β is, the bigger the bias).

For now, this problem seems manageable as we know the direction of the bias and could keep that in mind with interpretation, however, 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.

8.3 Instrumental variables

8.3.1 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.

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

8.4 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.

8.5 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).

8.5.1 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.

8.5.2 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.

8.6 Simultaneity

8.6.1 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).

8.6.2 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 9

Non-linear models, GMM and extremum estimators

9.1 Nonlinear Least Squares

9.1.1 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).

9.1.2 Estimation

Estimation of the model relies on finding the parameter $\hat{\theta}$ that reduces the MSE of the model. As we've seen, the analytic solution to the problem might be very difficult to compute and solve, thus we need to turn to numerical methods. We cover three types of numerical estimation methods here.

Lewbel's method (better name?)

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-based methods

The previous method relied only on the functional form of the model, using $g(\cdot)$ and $g'(\cdot)$, and used a known estimation procedure in OLS. Other methods can be used to directly solve the objective function numerically (instead of approximating a linear equation). In particular, gradient based methods of optimization will use information on the gradient of $g(\cdot)$ to find the solution. While this method will be very efficient is the model is well-behaved, it could be attracted to trivial solutions or local minima when the model is not smooth enough. When this happens, we will turn to global optimization methods.

Global methods

Global optimization methods relate to gradient-based ones in the sense that they take on the problem of finding the solution to the objective function, rather than

working on the model analytically. However, global optimization methods do not use any information on the functional forms of the function and try to get to the optimum point by evaluating the function at many points, based on an algorithm (i.e. Nelder-Mead) or naively (i.e. grid search). While this method will not be as efficient as gradient-based methods (since it does not use any information on the function), it will perform better when the functional form might trick the gradient-based methods.

9.2 Extremum Estimators

Extremum estimators are a class of estimators that solve an optimization problem of the form:

$$\hat{\theta} = \arg\max_{\theta} Q_n(\theta)$$

We can derive the asymptotic distribution of this class of estimators under four assumptions:

1. The estimator is consistent (i.e. $\hat{\theta} \xrightarrow{\text{p}} \theta_0$).
2. The true value θ_0 is not on the boundary of the parameter space Θ.
3. The objective function $Q_n(\theta)$ is twice continuously differentiable.
4. The derivative of the objective function is asymptotically linear, such that

$$\sqrt{n}\left(\frac{\partial Q_n(\theta)}{\partial \theta} - \bar{S}_n\right) \xrightarrow{\text{p}} 0$$

   where $\bar{S}_n = \frac{1}{n}\sum_{i=1}^{n} S_i$ converges to a zero-mean normal distribution at rate $\sqrt{n}$, with variance matrix Σ_0.

Let $H(\theta) = \operatorname{plim} \frac{\partial^2 Q_n(\theta)}{\partial \theta \partial \theta'}$, then, if $H(\theta)$ is bounded, continuous and nonsingular in the neighborhood of θ_0, we have that:

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{\text{d}} N(0, H_0^{-1}\Sigma_0 H_0^{-1})$$

9.3 Generalized Method of Moments

9.3.1 Moment Equation Model

Let $g_i(\theta)$ be a vector of l moments as a function of the data within the i-th observation and θ a k-dimensional unknown parameter. The moment equation model is defined as a system of l equations (also called moment conditions) such that:

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

In this system, we have l equations with which we are trying to identify k parameters (inside θ). This implies that we will not always be able to find a unique solution to the system. In particular, if $l < k$, we have more unknown parameters than equations, it will not be possible to find a solution: the model is underidentified. If $l \geq k$, then a unique solution must exist. Moreover, if $l > k$, then we have more equations than unknowns and excessive information (which can be used for other means than identification): the model is over-identified. In this chapter, we will only discuss the case when the model is just-identified or over-identified.

9.3.2 Method of Moments Estimator

As we've seen in the previous section, in order to identify the parameters in θ, you need to solve the moment equation model. However, the expectation of the moment conditions $g_i(\theta)$ is never observed and thus the solution cannot be computed as is. In order to go around this issue, we will use the sample analog of the expectation term: the sample average. Define $\bar{g}_n(\theta)$ as the sample average of the vector of moment conditions over n observations. Formally,

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

Following this, define the method of moments estimator (MME) as the value $\hat{\theta}$

that solves the moment equation model using the sample average:

$$\bar{g}_n(\hat{\theta}) = \frac{1}{n}\sum_{i=1}^{n} g_i(\hat{\theta}) = 0$$

Solutions to the system might be found analytically (OLS for example) or numerically. Note that this method works only for just-identified moment equation models, i.e. models in which $l = k$. For overidentified models, this method will be generally impossible.

9.3.3 Generalized Method of Moments Estimator

For the particular case of over-identified moment equation models, we cannot find a an estimator θ that would set the sample average to 0 exactly. The second-best solution is therefore to set $\bar{g}_n(\theta)$ as close to zero as possible. Again, an obvious way to do that is to use Least-Squares by squaring $\bar{g}_n(\theta)$ and finding $\hat{\theta}$ to minimize it. Before doing that, we will define W a weighting matrix that will help solving the model by assigning weights to moment conditions. This weighting matrix does not alter the interpretation of the problem; we are still doing least-squares but with weights. In particular, if $W = I_l$, then we are doing exactly least-squares. Hence, the GMM estimator can be defined as:

$$\hat{\theta} \in \arg\min_{\theta} J(\theta) \equiv n \cdot \bar{g}_n(\theta)' W \bar{g}_n(\theta)$$

The presence of n in the equation does not change the solution (as it is a scalar). On the contrary, the estimator value does depend on W and because of that, choosing the right W is crucial to estimating the model correctly. Note that even though different W can yield different estimator values, in the limit, the GMM estimator is consistent for any W. This means that choosing the best W is important for small samples and efficiency purposes only.

9.3.4 Which weighting matrix to choose?

As stated earlier, for any weighting matrix W, the GMM estimator will be consistent and converge in distribution to a normal distribution at rate $\sqrt{n}$. However,

the variance of the estimator is dependent on W since it is given by:

$$\text{Var}\left[\hat{\theta}\right] = (Q'WQ)^{-1}(Q'W\Omega WQ)(Q'WQ)^{-1}$$

where $\Omega = \text{E}\left[g_i g_i'\right]$ and $Q = \text{E}\left[\frac{\partial g_i(\theta)}{\partial \theta}'\right]$. Using this, we can find the optimal weighting matrix, which makes the GMM estimator efficient (achieves the lowest variance) as $W = \Omega^{-1}$. But as we are used to, this term is not observed so we will also need to estimate it somehow. There are multiple ways to do so.

First, one could not make such an effort and just go with a user-specified weighting matrix, such as $W = I_l$ for example. This will still achieve a consistent, although not efficient estimator. We call this estimator the one-step GMM.

Another way would be to try and estimate Ω using its sample average (or a sligthly modified version of it):

$$\hat{\Omega} = \frac{1}{n}\sum_{i=1}^{n}\left(g_i(\hat{\theta})\right)\left(g_i(\hat{\theta})\right)'$$

$$\text{or } \hat{\Omega}^* = \frac{1}{n}\sum_{i=1}^{n}\left(g_i(\hat{\theta}) - \bar{g}_n(\hat{\theta})\right)\left(g_i(\hat{\theta}) - \bar{g}_n(\hat{\theta})\right)'$$

As we can see, these estimates rely on an already estimated parameter $\hat{\theta}$ meaning one needs to perform a preliminary estimation of θ. This also suggests multiple ways to do it.

Two-step GMM

As the name suggests, this procedure is composed of a first estimation of the model using GMM and a user-specified weighting matrix (usually $W = I_l$), then a second estimation using the information obtained in the first stage. In particular, the steps are detailed below:

1. Run a GMM estimation using $W = I_l$ (or any other weighting matrix) and recover an estimated parameter $\hat{\theta}$
2. Compute an estimate of Ω using either $\hat{\Omega}$ or $\hat{\Omega}^*$. Invert it to obtain $\hat{W} = \hat{\Omega}^{-1}$.

3. Run a second GMM estimation using $W = \hat{W}$ and recover $\hat{\theta}$ as your final estimated parameter.

Iterated GMM

After reading the previous procedure, you might wonder why we should stop at two steps? Why not more? There is no particularly good reason to stop at two steps and you could go further by repeating the previous process until some convergence criterion is met. This would be called the iterated GMM estimator. All in all, while it requires more steps, this estimator is generally as efficient as the two-step version.

Continuously-updated GMM

Another question that might have popped up looking at the two-step procedure is why would we need two steps, if the only unknown in computing Ω is the object of our problem. Then the Continuously-Updated GMM estimator (CU-GMM) would be for you. It relies on plugging the estimate for $\hat{\Omega}$ directly into the first-stage optimization problem such that:

$$\hat{\theta} \in \arg\min_{\theta} J(\theta) \equiv n \cdot \bar{g}_n(\theta)' \left(\frac{1}{n} \sum_{i=1}^{n} \left(g_i(\hat{\theta}) \right) \left(g_i(\hat{\theta}) \right)' \right)^{-1} \bar{g}_n(\theta)$$

The CU-GMM estimator is not a quadratic problem in θ anymore and thus will require more advanced numerical techniques to solve. In exchange, it delivers a lower bias, although fatter tails in the distribution of θ. It is not very common in application.

9.3.5 Computing the variance

As always in econometrics, one will be interested in computing the variance of estimators in order to perform further analysis such as hypothesis testing, constructing confidence intervals, etc. Recall the theoretical formula for the

variance:

$$\operatorname{Var}\left[\hat{\theta}\right] = (Q'WQ)^{-1}(Q'W\Omega WQ)(Q'WQ)^{-1}$$

The issue here is that both $Q = \mathrm{E}\left[\frac{\partial g_i(\theta)}{\partial \theta'}\right]$ and $W = \mathrm{E}\left[g_i g_i'\right]$ are unknown, and as always, the solution is that we will have to estimate them. There are two main ways to do this: one is the "classical way" using previous estimates; the other is using bootstrapping.

Variance estimation

As we've seen in the previous section, we already have two estimators for the matrix Ω, relying on the law of large numbers (i.e. using the sample average) as:

$$\hat{\Omega} = \frac{1}{n}\sum_{i=1}^{n}\left(g_i(\hat{\theta})\right)\left(g_i(\hat{\theta})\right)'$$

$$\text{or } \hat{\Omega}^* = \frac{1}{n}\sum_{i=1}^{n}\left(g_i(\hat{\theta}) - \bar{g}_n(\hat{\theta})\right)\left(g_i(\hat{\theta}) - \bar{g}_n(\hat{\theta})\right)'$$

And using the same intuition, we can estimate Q using its sample average:

$$\hat{Q} = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial g_i(\hat{\theta})}{\partial \beta}$$

Bootstrap for GMM

The standard bootstrap algorithm generates boostrap samples by drawing observations from the data, with replacement, until some size is met. Then, the GMM estimator is computed over this particular sample. Repeating this process B times will give B estimators, from which we can compute the variance, confidence intervals, etc.

Chapter 10

Non-parametric estimators

10.1 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.

10.2 Estimation of the EDF

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 10.1 (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.

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:

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

10.3 Kernel Density Estimation

Density estimation might be interesting in its own right, when you need to identify the particular distribution of a random variable. Nevertheless, it is mostly studied as a fundamental building block for more complicated semi-/nonparametric models. Following the example in the previous section, suppose we want to estimate how Y is related to X where

$$Y = m_Y(X) + U$$

Then we recovered that, using the assumption that $m_Y(\cdot)$ is twice differentiable and bounded in its second-order derivative, as well as the assumption that $\mathrm{E}[U|X] = 0$, we have:

$$\mathrm{E}[Y|X = x] = m_Y(x) = \int_{\mathcal{X}} y \cdot f_{Y|X}(y, x)\mathrm{d}x$$

Moreover, from probability theory (Bayes' theorem):

$$\int_{\mathcal{X}} y \cdot f_{Y|X}(y, x)\mathrm{d}x = \int_{\mathcal{X}} y \cdot \frac{f_{YX}(y, x)}{f_Y(x)}\mathrm{d}x$$

where you have two density functions to estimate.

10.3.1 Introductory examples

Let X be a random variable that can take the value of 1 with true probability p^0 or 0 else. Think of how you would estimate the probability p^0.

One answer is to draw the random variable many times and get a series $\{x_1, x_2, ...\}$ then estimating $\hat{p}$ as the number of times we actually observed 1 divided by the number of draws. Formally, if we perform n random draws,

$$\hat{p} = \frac{\sum_{i=1}^{n} \mathbb{I}\{x_i = 1\}}{n}$$

where $\mathbb{I}\{\cdot\}$ is a function that takes a value of 1 if the condition inside is true, 0 if not. For example, if one million draws are made and 333 333 of them have turned out to be ones, then: $\hat{p} = \frac{333333}{1000000} \approx 1/3$.

Now, let's assume X is actually a continuous variable that can take any real value on its support. Thinking about the previous example, how would you estimate the probability that the realization of X falls in a given interval of length h around a given x, or more formally, falls in $[x - h/2, x + h/2]$. This value h is called the bandwidth.

Again, we could use the same strategy and draw the random variable n times, counting the times x_i falls in the ball around x and compare it with the total

number of draws:

$$\hat{\Pr}\left[X \in B_{h/2}(x)\right] = \frac{\sum_{i=1}^{n} \mathbb{I}\{x_i \in B_{h/2}(x)\}}{n} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\{x-h/2 \leq x_i \leq x+h/2\}$$

Is this type of estimator unbiased? We can check by looking at:

$$\mathrm{E}\left[\hat{\Pr}\left[X \in B_{h/2}(x)\right]\right] = \mathrm{E}\left[\mathbb{I}\{x - h/2 \leq X \leq x + h/2\}\right] = \Pr\left[X \in B_{h/2}(x)\right]$$

which shows that it is indeed an unbiased estimator.

10.3.2 Density estimation

We have just seen how to estimate probabilities without making assumptions on any structure; in this subsection, we will see how it relates to estimating a density function.

First, think of what the pdf of X, denoted $f_X(x)$, actually is. It is the probability that X takes the exact value x. In a sense, this is close to what we just did, however, we're looking for X to be a point rather than in a set. The probability of being in a set is given by the cdf $F_X(x)$. It turns out that as we reduce the size of the set more and more, the two concept become closer and closer. Formally, as h tends to 0, the set around $B_{h/2}(x)$ will only contain x. Since $f_X(x)$ is the derivative of $F_X(x)$, we can write:

$$f_X(x) = \lim_{h\to 0} \frac{F_X(x + h/2) - F_X(x - h/2)}{h} = \lim_{h\to 0} \frac{\Pr\left[X \in B_{h/2}(x)\right]}{h}$$

where you should recognize the last term from the previous subsection.

And in fact, you could estimate the pdf by using the estimator for the probability as seen above:

$$\hat{f}_X(x) = \frac{\hat{\Pr}\left[X \in B_{h/2}(x)\right]}{h} = \frac{1}{nh}\sum_{i=1}^{n} \mathbb{I}\{x - h/2 \leq x_i \leq x + h/2\}$$

for a given h that is relatively small (more about this later). We now have our first own density estimator, let's look at it in more detail.

The basic idea behind the estimator is to count how many observations fall in the neighborhood x, relative to the total number of observations, and the size of the neighborhood. Here we use "count" since our indicator function is rather naïve and only does that: setting a weight of one for observations in the neighborhood, and 0 for observations out of the neighborhood. The weight assignment function is called a kernel (hence the name of kernel density estimator). In particular, the one used above is called a uniform kernel because it assigns a uniform weight to all observations within the neighborhood. In practice, this is a very bad kernel and it should rarely be used. The parameter h that defines the size of the neighborhood is called the bandwidth.

10.3.3 Properties of Kernel Density Estimators

Definition 10.2 (Standard Kernel)**.** *A standard kernel $K : \mathbb{R} \to \mathbb{R}_+$ is a non-negative function such that:*

- $\int K(\psi)\mathrm{d}\psi = 1$*: the cdf of the kernel goes to one.*
- $\int \psi K(\psi)\mathrm{d}\psi = 0$*: the kernel is symmetric around 0.*
- $\int K^2(\psi)\mathrm{d}\psi = \kappa_2 < \infty$*:*
- $\int \psi^2 K(\psi)\mathrm{d}\psi = \mu_2 < \infty$*:*

You should view these properties through the lens of what we actually use a kernel for. Since a kernel is essentially a "weight-assigning" function, it must make sense that it is symmetric (equally off observations in either direction should be equally bad), that it is non-negative (although it might be interesting to assign negative weights to observations we really don't want) and that it stops assigning weights after a certain distance.

Using this definition, we can then define a kernel density estimator.

Definition 10.3 (Rosenblatt-Parzen Kernel density estimator)**.** *A Kernel density estimator for a given pdf $f_X(x)$ is defined as:*

$$\hat{f}_X(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x_i - x}{h}\right)$$

where $K(\cdot) : \mathbb{R} \to \mathbb{R}_+$ is a standard kernel.

Interesting examples of kernels include:

- Uniform kernel: $K(\psi) = \mathbb{I}\{|\psi| \leq 1/2\}$
- Gaussian kernel: $K(\psi) = \frac{1}{\sqrt{2}} \exp\{-0.5 \cdot \psi^2\}$
- Epanechnikov kernel: $K(\psi) = \mathbb{I}\{|\psi| \leq 1\} \cdot (1 - \psi^2) \cdot (3/4)$

As we did in the parametric econometrics classes, we now look for the kernel density estimator properties such as bias and variance.

Bias of the KDE

Assume a random sampling over iid data. Then,

$$\begin{aligned}\mathrm{E}\left[\hat{f}_X(x)\right] &= \frac{1}{nh}\mathrm{E}\left[\sum_{i=1}^{n} K\left(\frac{x_i - x}{h}\right)\right] = \frac{1}{nh} \cdot n \cdot \mathrm{E}\left[K\left(\frac{X - x}{h}\right)\right] \\ &= \frac{1}{h}\int K\left(\frac{\xi - x}{h}\right) \cdot f_X(\xi)\mathrm{d}\xi\end{aligned}$$

Then, we perform a change of variables such that the term inside the kernel is ψ, meaning $\xi = \psi h + x$ and $\mathrm{d}\xi = h \cdot \mathrm{d}\psi$. Replacing it in the bias formula we get:

$$\mathrm{E}\left[\hat{f}_X(x)\right] = \frac{1}{h}\int K(\psi) \cdot f_X(\psi h + x) \cdot h \cdot \mathrm{d}\psi = \int K(\psi) \cdot f_X(\psi h + x)\mathrm{d}\psi$$

Further, let's use a second order mean value expansion to recover $f_X(x)$:

$$f_X(\psi h + x) = f_X(x) + \psi h f'_X(x) + \frac{(\psi h)^2}{2} f''_X(x_r)$$

where x_r includes a remainder term such that: $x_r = x + \lambda\psi h$. This yields us:

$$\mathrm{E}\left[\hat{f}_X(x)\right] = \int K(\psi) \cdot \left[f_X(x) + \psi h f'_X(x) + \frac{(\psi h)^2}{2} f''_X(x_r)\right] \mathrm{d}\psi$$

$$= \int K(\psi) \cdot f_X(x) \mathrm{d}\psi + \int K(\psi) \cdot \psi h f'_X(x) \mathrm{d}\psi + \int K(\psi) \cdot \frac{(\psi h)^2}{2} f''_X(x_r) \mathrm{d}\psi$$
$$= f_X(x) \underbrace{\int K(\psi) \cdot \mathrm{d}\psi}_{=1} + h f'_X(x) \underbrace{\int K(\psi) \cdot \psi \mathrm{d}\psi}_{=0} + \frac{h^2}{2} \int K(\psi) \cdot \psi^2 f''_X(x_r) \mathrm{d}\psi$$

The last term is quite problematic since it cannot be simplified out of the integral. However, we know that $f''_X(x)$ could be, so we can naively subtract it and we'll see later that the remainder is actually not very relevant.

$$\frac{h^2}{2} \int K(\psi) \cdot \psi^2 f''_X(x_r) \mathrm{d}\psi = \frac{h^2}{2} \int K(\psi) \cdot \psi^2 (f''_X(x_r) - f''_X(x)) \mathrm{d}\psi$$
$$+ \frac{h^2}{2} \int K(\psi) \cdot \psi^2 f''_X(x) \mathrm{d}\psi$$
$$= R + \frac{h^2}{2} f''_X(x) \mu_2$$

where R is bounded by $o(h^2)$. Finally, we can write the expectation of our kernel density estimator as:

$$\mathrm{E}\left[\hat{f}_X(x)\right] = f_X(x) + \underbrace{\frac{h^2}{2} f''_X(x) \mu_2 + o(h^2)}_{\mathrm{Bias}\left[\hat{f}_X(x)\right]}$$

and the bias is given by the last two terms. From this equation, you can see that the bias is increasing with the bandwidth. This is intuitive since a greater bandwidth also implies more observations that are not related to x (global information) relative to the observations actually close to x (local information). Global information being more likely to introduce bias in the estimator, h is positively correlated with bias. In the opposite direction, the bias seems to disappear as h goes to 0. This means that the estimator is more efficient when the bandwidth is very small, then why not make the bandwidth as small as possible? One could show by similar equation work that the variance of the estimator is given by:

$$\mathrm{Var}\left[\hat{f}_X(x)\right] = \frac{1}{nh} f_X(x) \kappa_2 + o((nh)^{-1})$$

which this time is actually increasing as h tends to 0. Again, intuitively this makes sense as reducing the size of the bandwidth will eventually reduce the number of observations and thus increase the variance. This phenomenon is called the bias-variance trade-off.

Bias-variance trade-off

In order to have a sense of what the bias and variance look like over the whole distribution, we integrate them with respect to x:

$$\int \left(\text{Bias}\left[\hat{f}_X(x)\right]\right)^2 \mathrm{d}x = c_1 \cdot h^4 \quad \int \text{Var}\left[\hat{f}_X(x)\right] \mathrm{d}x = c_2 \cdot (nh)^{-1}$$

This allows us to design an optimal measure of the trade-off, analogous to the mean squared error in the parametric case, defined as the Mean Integrated Squared Error: $\text{MISE}(h) \equiv c_1 \cdot h^4 + c_2 \cdot (nh)^{-1}$. Now suppose we want to find the best bandwidth to minimize MISE:

$$\frac{\partial MISE}{\partial h} = 0 \Leftrightarrow 4 \cdot c_1 \cdot h^3 - c_2 \cdot n^{-1} h^{-2} = 0 \Leftrightarrow h \sim n^{-1/5}$$

meaning that h must be proportional to $n^{-1/5}$. Again, this makes a lot of sense since it implies that increasing the number of observations allows you to reduce the size of the bands: the more observations you have, the more likely it is that they will fall around x, and thus the less need you have to keep wide bands.

Asymptotics

The rate of convergence of the KDE is $\sqrt{nh}$ where n is the number of observations and h the bandwidth. For an optimal bandwidth, we had $h = n^{-1/5}$, yielding a convergence rate of $\sqrt{n \cdot n^{-1/5}} = \sqrt{n^{4/5}} = n^{2/5}$. Therefore, the nonparametric estimator has a slower rate of convergence than its parametric counterparts of OLS and ML estimators.

10.3.4 Going beyond the univariate, first-order KDE

As we've seen, the KDE method is very interesting in how it gets around the lack of structure, but it creates a new trade-off between bias and variance. In order to reduce further the bias, one might be interested in increasing the dimensions of the kernels, to allow for capturing more data points.

Density derivatives

If $f_X(x)$ is a differentiable function of x, one could also use the derivative of the kernel to estimate the object. In practice, to estimate a r-th order derivative, a r-th order kernel would set all moments up to the r-th one to 0, and the r-th one as a finite moment μ_r. This technique displays the advantage of having convergence at a rate closer to $\sqrt{n}$. However, you would get potentially negative tails (meaning it would not be a proper density), and the estimator would be very efficient in small samples.

Multivariate Density estimation

Another way of achieving bias reduction would be to use multivariate density estimation, meaning X would now be a random vector in $\mathbb{R}^k$ space. The kernel density estimator would then be a product of all univariate kernels. Obviously, this method adds additional variation in the function, but you would need way more observations to get an interesting result. To see that, recall the intuition behind the kernel function: it assigns weights to observations based on the mean distance of these observations to a point of interest, given a bandwidth. It turns out that increasing the dimension of the neighborhood (from a line, to a square, to hypercubes) will also increase the volume of the object, thus reducing the probability of observations being near the point of interest, and increasing the need for observations. This problem is called the curse of dimensionality.

10.4 Kernel Regression Estimation

In the previous section, we were interested in estimating the distribution of one variable X. However, in most economics applications, a more interesting element to estimate is the distribution of a variable Y, conditional on X. This is the mean regression model that we are going to study here.

Recall our definition of a kernel density estimator for a true distribution $f_X(x)$:

$$\hat{f}_X(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x_i - x}{h}\right)$$

where K is a standard kernel (refer to ...). Also recall the mean regression model of

$$\mathrm{E}\left[Y|X = x\right] = m_Y(x) = \int y \cdot \frac{f_{XY}(x, y)}{f_X(x)} \mathrm{d}y$$

Our goal is to use kernel density estimators for both the distribution of X and the joint distribution of X and Y. Formally, we look for:

$$\hat{m}_Y(x) = \int y \cdot \frac{\hat{f}_{XY}(x, y)}{\hat{f}_X(x)} \mathrm{d}y$$

10.4.1 Nadaraya-Watson Estimator

Intuitively, we turn directly to our kernel density estimators. We already wrote the definition of our estimator in the univariate case of estimating $\hat{f}_X(x)$, but what is the KDE for the joint distribution of X and Y? For that we use a multivariate KDE including a product kernel. In particular, we get:

$$\hat{f}_{XY}(x, y) = \frac{1}{nh^2} \sum_{i=1}^{n} \left(K\left(\frac{x_i - x}{h}\right) \cdot K\left(\frac{y_i - y}{h}\right) \right)$$

The two KDE can be used in the mean regression estimator to get:

$$\begin{aligned} \hat{m}_Y(x) = \int y \cdot \frac{\hat{f}_{XY}(x, y)}{\hat{f}_X(x)} \mathrm{d}y &= \int y \cdot \frac{(nh^2)^{-1} \sum_{i=1}^{n} \left(K\left(\frac{x_i - x}{h}\right) \cdot K\left(\frac{y_i - y}{h}\right) \right)}{(nh)^{-1} \sum_{i=1}^{n} K\left(\frac{x_i - x}{h}\right)} \mathrm{d}y \\ &= \frac{1}{h} \cdot \frac{\sum_{i=1}^{n} K\left(\frac{x_i - x}{h}\right) \cdot \int y K\left(\frac{y_i - y}{h}\right) \mathrm{d}y}{\sum_{i=1}^{n} K\left(\frac{x_i - x}{h}\right)} \end{aligned}$$

The only term that is not obvious here is the last term of the numerator. Let's look at it in detail.

Apply a change of variable so that ψ is the term inside the kernel. We get $y = \psi h + y_i$ (recall that since the kernel is symmetric $K(y_i - y) = K(y - y_i)$).

We also have $\mathrm{d}y = h\mathrm{d}\psi$. Then we can write:

$$\int yK\left(\frac{y_i - y}{h}\right)\mathrm{d}y = \int (\psi h + y_i)K(\psi)\, h\mathrm{d}\psi$$

and separating we have:

$$h^2 \int \psi K(\psi)\,\mathrm{d}\psi + hy_i \int K(\psi)\,\mathrm{d}\psi = h \cdot y_i$$

from the properties of the kernel. Finally, plugging this expression back into the mean regression estimator we get:

$$\hat{m}_Y(x) = \frac{\sum_{i=1}^n K\left(\frac{x_i - x}{h}\right) \cdot y_i}{\sum_{i=1}^n K\left(\frac{x_i - x}{h}\right)}$$

Definition 10.4 (Nadaraya-Watson estimator)**.** *For a given model of two variables Y and X such that $Y = m_Y(X) + U$, the Nadaraya-Watson estimator for the function $m_Y(x)$ is defined as:*

$$\hat{m}_Y(x) = \frac{\sum_{i=1}^n K\left(\frac{x_i - x}{h}\right) \cdot y_i}{\sum_{i=1}^n K\left(\frac{x_i - x}{h}\right)}$$

where $K(\cdot) : \mathbb{R} \to \mathbb{R}$ is a standard kernel.

Note that a kernel regression estimator is only a valid estimator for $m(\cdot)$ in a local neighborhood of size h.

Nadaraya-Watson and OLS

Consider a model of Y and X such that $Y = \alpha + U$ or in matrix form:

$$\begin{bmatrix} Y_1 \\ \vdots \\ Y_N \end{bmatrix} = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} \cdot \alpha + \begin{bmatrix} U_1 \\ \vdots \\ U_N \end{bmatrix}$$

By OLS, the estimation of α is now straightforward:

$$\hat{\alpha} = (\iota'\iota)^{-1}\iota' Y = \frac{\sum_i Y_i}{n} = \bar{Y}$$

where ι is a n-dimensional vector of 1. This means that the OLS estimation is equivalent to fitting a constant (the average of Y) globally on the model. Now, if you consider the NW estimator, you should see a type of relation between both. In fact, around the neighborhood of x, the two estimators are exactly the same. Hence, intuitively, you could see the NW estimator as fitting a constant locally for all x. To see that, reweight the the data by $K\left(\frac{x-X_i}{h}\right)^{1/2}$ so that observations in the neighborhood are 1, while others are 0 (in the case of the uniform kernel), the NW estimator is in fact the average of Y for values of Y inside of the neighborhood.

10.4.2 Local OLS estimator

We have seen that the intuition behind the NW estimator was about fitting a constant locally on the model. Naturally, one could think about extending this line of reasoning and fit more complex models inside the kernel. In particular, a well-studied extension is to fit a line, as a local linear model. This type of models is usually called local OLS models. They are represented by the following model:

$$Y = m(x)\tilde{\iota} + h \cdot m'(x)\frac{X_i - x}{h} + U = \tilde{X}\beta(x) + U$$

Note that adding dimensions to the polynomial used to fit the model locally does not change the value of the $m(x)$ function, but rather adds information about higher-order derivatives of the $m(x)$ function at the point x. For example, estimating a simple line locally gives the value of the point at x as well as the slope of m at x.

Definition 10.5 (Local OLS estimator)**.** *For a given model of two variables Y and X such that $Y = m(x)\tilde{\iota}+h\cdot m'(x)\frac{X_i-x}{h}+U = \tilde{X}\beta(x)+U$, the local OLS estimator for the function $m_Y(x)$ is defined as:*

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

10.4.3 Bias and Variance

The bias of the kernel regression estimation is given by:

$$\mathrm{E}\left[\hat{\beta}_0|X\right] - \beta_0 = \mathrm{E}\left[\hat{m}(x)|X\right] - m(x) = \frac{h^2}{2}\cdot\mu_2 m''_Y(x) + o_p(h^2)$$

which is very similar to the formula derived for the bias of the kernel density estimator. If we also look at the variance, we get:

$$\text{Var}\left[\hat{\beta}_0|X\right] = \text{Var}\left[\hat{m}(x)|X\right] = \frac{1}{nh} \cdot \kappa^2 \frac{\sigma_U^2}{f_X(x)} + o_p((nh)^{-1})$$

which is slightly different than the kernel density counterpart. In fact, in the regression case, $f_X(x)$ enters in the denominator, meaning that increasing $f(x)$ (the probability of finding observations at the point x) will decrease the variance while in the density estimation, increasing the number of observations increased the variance.

10.4.4 General considerations

Asymptotic normality

Under similar regularity conditions as the density estimator, the regression estimation will tend in distribution to a standard normal as the number of observations increase to ∞. The rate of convergence is also $\sqrt{nh}$ in this setting.

Curse of dimensionality

Again, in a similar way as the KDE, the kernel regression estimator faces the curse of dimensionality as the number of regressors k increases. The rate of convergence is then $\sqrt{nh^k}$.

Higher-order bias reduction

Same as KDE.

Order of local polynomial

As it is discussed in the section, instead of fitting a constant or a line, one could go further and fit higher order polynomials in the data in order to get more information on the shape of the fitted line. It turns out that asymptotically, there is no cost to move to the next-odd order when estimating a given object of interest. For example, say you want to estimate $m(\cdot)$ up to its p-th derivative, then you could use a $p+1$ or $p+3$ order local polynomial estimation. You should remember that when it comes to local OLS, it is an odd world.

Moreover, estimating polynomials will actually achieve bias reduction in the same way that high order kernel do. This bias reduction comes without the cost of putting negative weight on some observations like the higher order kernels do. This is why it is generally thought that high order polynomial is more interesting than high order kernels.

Selection of bandwidth

There are two schools of thoughts when it comes to bandwidth selection in the case of kernel regression.

We have seen in the kernel density estimation section that we chose the bandwidth in order to minimize the mean integrated squared error. In the context of kernel regression, the MISE does not have an analytic expression, and we thus have to approximate it using the following expression:

$$AMISE = \int \frac{h^2}{2} \cdot \mu_2 m_Y''(x) + \frac{1}{nh} \cdot \kappa^2 \frac{\sigma_U^2}{f_X(x)} \mathrm{d}x$$

and then minimize over h.

The second way to select the adequate bandwidth is to use cross-validations. Define the leave-one-out estimator $\hat{m}_{-j}$ as the the standard local OLS estimator, without the j-th observation. Now let the average prediction error (APE) be:

$$APE = \frac{1}{n} \sum_j (Y_j - \hat{m}_{-j}(X_j))^2$$

It turns out that choosing h to minimize the APE is equivalent to minimizing the MISE.

Choice of kernel

Use Epanechnikov.

10.4.5 Series/sieve regression

10.4.6 Testing

In nonparametrics, hypothesis testing is separated from the regression estimation. It is also difficult to perform as generally, hypothesis testing is meant to look at interesting features on the whole dataset (the whole function), while nonparametrics focuses on local features of the data.

Omission of variables test

10.4.7 Applications

Binary Dependent Variable

Consider the case where Y, the dependent variable, only takes values of 1 or 0, such that:

$$Y = \begin{cases} 1 & \text{if } X\beta + U > 0 \\ 0 & \text{else.} \end{cases}$$

where U is assumed to be independent of X, $U \perp X$. As outlined in the beginning of this section, we look for an estimator for the expectation of Y given $X = x$.

Since Y is now a discrete random variable, we can write:

$$\begin{aligned}\mathrm{E}\left[Y|X=x\right] &= 1 \cdot \Pr\left[X\beta + U > 0 | X = x\right] + 0 \cdot \Pr\left[X\beta + U \leq 0 | X = x\right] \\ &= \Pr\left[X\beta + U > 0 | X = x\right] = \Pr\left[U > -x\beta\right] \\ &= 1 - G(-x\beta)\end{aligned}$$

This setting is problematic since it does not allow for "point-identification" of β. To see that, note that we have two objects to estimate here: $\theta = \{G(-X\beta), \beta\}$. However, we could also define $\tilde{\beta} = \beta/c$ and $\tilde{G}(z) = G(cz)$, which would yield that $\tilde{G}(X\tilde{\beta}) = G(X\beta)$. This means that observing the same data, one could estimate $\tilde{G}$ or G, leaving β unidentified, or set identified (all vectors $\tilde{\beta}$). We say that β is identified up to scale c.

In order to solve this issue, we can impose a restriction on the size of β such that we can single out a parameter from all proportional parameters. We call this restriction a normalization.

This normalization turns out not to affect the economic meaningfulness of the model. In fact, we have just seen that $G(\cdot)$ is perfectly identified, but identification of β, although an advantage, is not necessary. To see that, consider another object of interest in this model:

$$\nabla_x \mathrm{E}\left[Y|X=x\right] = \beta \cdot g(-x\beta)$$

where g denotes the pdf of the distribution of U. Then, define the set of parameters we want to estimate as $\theta = \{G(X\beta), \beta g(X\beta)\}$. Now let $\tilde{\beta} = \beta/c$ and $\tilde{G}(x) = G(cx)$. From this we get that $G(-X\beta) = \tilde{G}(-X\tilde{\beta})$ and $\beta g(-X\beta) = \tilde{\beta}\tilde{g}(-X\tilde{\beta})$. Therefore, we can write that $\tilde{\theta} = \theta$, meaning that whatever the value of c is, our set of objects of interest will not change.

Chapter 11

Program Evaluation and Treatment Effects

11.1 Intuition

Suppose the data follows the model:

$$Y = \phi(D, A)$$

where $\phi(\cdot)$ is a very general function of the data, such that it is not differentiable; D is the discrete (binary) variable indicating if yes (1) or no (0) the treatment was administered; and A is a potentially infinite dimensional error.

We denote Y_1 and Y_0 as respectively the values of the outcome for each different treatment:

$$Y_1 = \phi(1, A); \quad Y_0 = \phi(0, A)$$

Note that both variables are not (never) directly observed. The observations in the data are realized outcomes depending on the realization of the random variable A. Thus, the function ϕ that transforms the data can never be observed.

Ideally, we want to be able to recover the effect of the treatment, or how the outcome changes when $D = 0$ increases to $D = 1$, for a given $A = a$ (an

individual). We call this the individual treatment effect:

$$Y_1 - Y_0 = \phi(1, A) - \phi(0, A)$$

which varies for any A, across the population. However, knowing the effect for any individual might not be that useful in practical terms. In fact, when designing a policy or evaluating programs, you might be interested only in a subgroup of people, or the population as a whole, but rarely about each individual. This is why we might be more interested in the Average Treatment Effect (or ATE), defined as:

$$ATE \equiv \mathrm{E}\left[Y_1 - Y_0\right] = \mathrm{E}\left[\phi(1, A) - \phi(0, A)\right]$$

the average of the individual treatment effect across all individuals. One could also be directly looking at the subgroup of interest, say the average treatment effect on the treated (ATT), i.e.

$$ATT \equiv \mathrm{E}\left[Y_1 - Y_0 | D = 1\right] = \mathrm{E}\left[\phi(1, A) - \phi(0, A) | D = 1\right]$$

Going further, one might be interested in identifying a subgroup on other characteristics X, using the average treatment effect conditional on X (or CATE):

$$CATE \equiv \mathrm{E}\left[Y_1 - Y_0 | X = x\right] = \mathrm{E}\left[\phi(1, A) - \phi(0, A) | X = x\right]$$

Finally, in the same line of reasoning, one could separate subpopulations in terms of endogeneity of their response to the treatment, using estimators we'll study later like LATE, MTE, etc. All these estimators take the form:

$$\mathrm{E}\left[Y_1 - Y_0 | \textit{Subpop.}\right] = \mathrm{E}\left[\phi(1, A) - \phi(0, A) | \textit{Subpop.}\right]$$

To go back to our first object of interest, an alternative interpretation of the average treatment effect can be found by rewriting the equation in terms of the binary variable:

$$Y = \phi(D, A) = Y_0 + (Y_1 - Y_0) \cdot D = \alpha(A) + \beta(A) \cdot D$$

where Y_0 becomes a random intercept and $Y_1 - Y_0$ a random slope. Then, the ATE is the average random slope of the model: $ATE = \mathrm{E}\left[\beta(A)\right]$.

11.2 Identification

If Y_1 and Y_0 were known for the whole population under study, there would not be a whole field dedicated to compute the ATE. In fact, averaging over a simple substraction would be quite easy. However, for any individual i, only one of the outcomes can be observed at the same time. In fact, either an individual received the treatment (Y_{1i} is observed) or he did not (Y_{0i} is observed). Because of that fact, we will have to make some assumptions on the unobservables to make progress.

11.2.1 Joint full independence

In particular, the first assumption we ought to make is the so-called "joint full independence" of outcomes with respect to treatment. Formally, we write: $(Y_1, Y_0) \perp D$, meaning that jointly, Y_1 and Y_0 are fully independent from D. This also implies that $A \perp D$.

Intuitively, this assumption (denoted $A2$) means two things. First, that everything not observed by the econometrician (A) is independent of the treatment D, i.e. receiving the treatment or not does not change the unobserved variables that might affect the outcome of the treatment. Second, the unobserved variable A has no effect on the treatment being delivered or not, i.e. the treatment is purely random, even on unobserved characteristics.

This assumption has an interesting implication on the regression of Y on D. Consider the regression separated for each group of treatment. For the treated:

$$\begin{aligned}
\mathrm{E}[Y|D=1] &= \mathrm{E}[Y_0 + (Y_1 - Y_0) \cdot D|D=1] \\
&= \mathrm{E}[Y_0 + (Y_1 - Y_0)|D=1] \\
&= \mathrm{E}[Y_1|D=1] \\
&= \mathrm{E}[Y_1] \text{ by assumption of joint full independence.}
\end{aligned}$$

Similarly, for the non-treated, you get: $\mathrm{E}[Y|D=0] = \mathrm{E}[Y_0]$. And thus,

$$ATE = \mathrm{E}[Y_1 - Y_0] = \mathrm{E}[Y_1] - \mathrm{E}[Y_0] = \mathrm{E}[Y|D=1] - \mathrm{E}[Y|D=0]$$

In words, this assumption allows the econometrician to compute the ATE as the difference between the average effect across the treatment group ($\mathrm{E}[Y|D=1]$)

and the average effect across the control group ($\mathrm{E}[Y|D=0]$). Remember that this can only be true if unobservables across the whole population are independent of the treatment.

Under the same assumption, we also get that $\mathrm{E}[Y|D=0] = \mathrm{E}[Y_0] = \mathrm{E}[Y_0|D=1]$ and thus $ATE = ATT$.

11.2.2 Unconfoundedness

Although the previous assumption allows for some very interesting results, it requires a lot of effort to ensure. In fact, the assumption requires perfect randomization of the treatment assignment. This setting is called a perfect experiment, but it is not so common in research, as it is hard to randomize, and/or make sure that everyone follows the instructions. Nevertheless, we could study a more realistic setting where, conditional on some observables X, we would have independence. Assuming the following model:

$$Y = \phi(D, X, A) = \alpha(A, X) + \beta(A, X) \cdot D$$

the unconfoundedness assumption ($A2'$) requires that $(Y_1, Y_0) \perp D|X$, implying that $A \perp D|X$, instead of the previous $A \perp D$. A weaker assumption ($A2''$) would be that only the expectations of Y would be the same regardless of the treatment once conditioned on X (more formally, $\mathrm{E}[Y_j|D, X] = \mathrm{E}[Y_j|X]$ for both $j = 0, 1$).

Using this assumption and following the same reasoning as with joint full independence, we can come up with the Conditional Average Treatment Effect (CATE):

$$\begin{aligned} CATE(x) = \mathrm{E}[Y_1 - Y_0|X = x] &= \mathrm{E}[Y_1|X = x] - \mathrm{E}[Y_0|X = x] \\ &= \mathrm{E}[Y|D = 1, X = x] - \mathrm{E}[Y|D = 0, X = x] \end{aligned}$$

and thus, the average treatment effect as:

$$ATE = \int CATE(x)\mathrm{d}F(x) = \int \left(\mathrm{E}[Y|D = 1, X = x] - \mathrm{E}[Y|D = 0, X = x]\right) \mathrm{d}F(x)$$

which in words is the expectation of the CATE over X.

Estimation

This equation for the ATE should really ring a bell if you have followed the last chapter. In fact, both elements within the integral can be estimated with nonparametric (kernel) regression. However, this type of regression applied directly to the problem will throw you directly under the curse of dimensionality (having expectations conditional on both D and X, the latter potentially being multidimensional as well).

A solution to this problem could be to implement a variable linking the treatment D to the observables X. Define a propensity score $p(x) \equiv \Pr[D = 1|X = x]$. Along uncertainty in X, you get uncertainty in p, which can be summarized as a random variable P. Then, using the previous assumptions $A2'$, we get:

$$CATE(p) = \mathrm{E}[Y|D = 1, P = p] - \mathrm{E}[Y|D = 0, P = p]$$

where P only has a single dimension.

In practice, this propensity score p could be estimated nonparametrically or not. The issue with nonparametric estimation is that you are just displacing the dimensionality curse situation. Using a parametric structure such as the probit, logit or the kind would help in reducing the dimensionality issue.

Now, using the definition of the ATE, we have:

$$ATE = \mathrm{E}[\mathrm{E}[Y|D = 1, P = p] - \mathrm{E}[Y|D = 0, P = p]]$$

$$\widehat{ATE} = \frac{1}{n}\sum_i \hat{m}_1(P_i) - \hat{m}_0(P_i)$$

Practical issues

Using last chapter, we know how to estimate $m(\cdot)$. Nevertheless, the setting derived just above is slightly different than before in the sense that the object of interest, $\widehat{ATE}$, is now an average over kernel regression estimators. Among other things, this changes how we interpret the optimal bandwidth. In fact, since we are now averaging, we could deal with smaller bandwidths without being

too scared of the effect it would have on variance (averages reduce variances). Because a smaller h is not that costly anymore, the cross-validation approach does not deliver the best bandwidth anymore, so we'll have to use different approaches. In particular, the field has come up with two interesting approaches: (1) the propensity score matching and (2) direct averages.

The propensity score matching is very intuitive and maps to a sort of nearest neighbor estimator. The idea is that for any individual i in the control group (with propensity score p_i), you find the individual i' in the treatment group such that $i' \in \arg\min_{j \in I_1} |p_i - p_j|$ where I_1 is the set of individual who received the treatment. In words, you "match" every individual in the control group with at least one individual in the treatment, based on the proximity of their proximity score. Then, for each pair you compute the difference between their CATE, and finally average over all pairs to get the ATE. The advantage in that estimator is that as n increase, you will find more and more individuals in the matching pairs. However, one disadvantage is that even with an infinite number of individuals, the bias of this estimator will not vanish.

The second approach of direct averages uses a clever rewriting of the problem such that the ATE is defined as:

$$ATE \equiv \mathrm{E}\left[\frac{(D - p(X)) \cdot Y}{p(X) \cdot (1 - p(X)}\right]$$

which suggests the simple following sample counterpart:

$$\widehat{ATE} \equiv \frac{1}{n} \sum_i \frac{(D_i - \hat{p}(X_i)) \cdot Y_i}{\hat{p}(X_i) \cdot (1 - \hat{p}(X_i)}$$

where $\hat{p}(\cdot)$ can be any first-stage estimator of the propensity score (non-parametric, probit, etc.).

11.2.3 Regression Discontinuity Design (RDD)

The RDD is another setup used to analyze treatment effects conditional on covariates. The idea is quite simple and intuitive since it relies on an existing discontinuity in the treatment selection (who gets it and who do not) to study

the effect of the treatment. In simpler words, if along a dataset the only discontinuity is whether a treatment was received or not (all other variables are continuous), then by studying the response for people around the discontinuity, you can identify the effect of the treatment.

For example, consider a situation in which students in a high-school are selected to go in an "honors" class based on their grade in some exam. The threshold is set at 800 points, such that everyone (this is important) above the threshold goes to the "honors" program, and everyone below does not. Then, assuming that people close to the threshold (in both directions) are similar in ability, we could study the effect of the "honors" program by looking at the average difference in effect between people on both sides of the threshold.

Model

Consider the following structural model:

$$Y = Y_0(X, A) + [Y_1(X, A) - Y_0(X, A)] \cdot D \text{ where } D = \mathbb{I}\{X \geq c\}$$

or in words, the total outcome Y is equal to the control outcome (Y_0) plus the difference between the treatment and control outcomes ($Y_1 - Y_0$), in case the treatment was administered, which is the case if and only if $X \geq c$. Then, we get:

- On the right side of the discontinuity:

$$\begin{aligned}\lim_{x \to c^+} \mathrm{E}\left[Y | D = 1, X = x\right] &= \lim_{x \to c^+} \mathrm{E}\left[Y_1(x, A) | D = 1, X = x\right] \\ &= \lim_{x \to c^+} \mathrm{E}\left[Y_1(x, A) | X = x\right] \text{ (by A2')}\end{aligned}$$

- On the left side of the discontinuity:

$$\begin{aligned}\lim_{x \to c^-} \mathrm{E}\left[Y | D = 0, X = x\right] &= \lim_{x \to c^-} \mathrm{E}\left[Y_0(x, A) | D = 0, X = x\right] \\ &= \lim_{x \to c^-} \mathrm{E}\left[Y_0(x, A) | X = x\right] \text{ (by A2')}\end{aligned}$$

Assume that the distribution of the unobservables A, conditional on observables X is exactly the same within the infinitesimal neighborhood of the threshold c. Formally, assume:

$$\lim_{x \to c^+} f_{A|X}(a; x) = \lim_{x \to c^-} f_{A|X}(a; x) = f_{A|X}(a; c)$$

Moreover, assume that the outcome Y_j, conditional on both X and A, is the same within the infinitesimal neighborhood of the threshold. Formally,

$$\lim_{x \to c^+} Y_j(x, a) = \lim_{x \to c^-} Y_j(x, a) = Y_j(c, a)$$

for both Y_0 and Y_1.

Then, we have that:

$$\begin{aligned}
&\lim_{x \to c^+} \mathrm{E}\left[Y | D = 1, X = x\right] - \lim_{x \to c^-} \mathrm{E}\left[Y | D = 0, X = x\right] \\
&= \mathrm{E}\left[Y_1(c, A) | X = c\right] - \mathrm{E}\left[Y_0(c, A) | X = c\right] \\
&= \mathrm{E}\left[Y_1(c, A) - Y_0(c, A) | X = c\right] = CATE(c)
\end{aligned}$$

This technique gives us the conditional average treatment effect based on being around the threshold. For that reason, it cannot be used to recover the global average treatment effect (the ATE), even using the techniques developed above. One should always keep in mind that the RDD model only applies for the neighborhood of the discontinuity.

11.2.4 Endogeneity

All three of the previous methods to compute the average treatment effect or the conditional average treatment effect rely on some version of assumption 2 ($A2$) which is correct only in the case of conditional or unconditional exogeneity of the treatment. However, in most applications, while selection of the treatment could be perfectly random, individual compliance with the selected treatment is not guaranteed. In fact, if you consider the effect of a training program for unemployed individuals, some people could be randomly selected to participate in a program, but decide not to do it. In order to control for that, we need a model that allows for endogenous selection.

Model

This model relies on two stages:

$$\begin{aligned}
\text{(2nd stage): } & Y = Y_0 + \Delta \cdot D \\
\text{(1st stage): } & D = \mathbb{I}\{\psi(Z, V) > 0\}
\end{aligned}$$

where $\Delta \equiv Y_1 - Y_0$ as in previous models, Z are instruments, V are first-stage unobservables and $\psi(\cdot)$ is a function that maps the space defined by (Z, V) to the decision space (where a positive number means the treatment is accepted, and a negative that is refused).

The first-stage equation describes the choice of participation in the treatment: given some exogenous stimulus Z and unobservables V (that the individual observes, but not the econometrician), if $\psi(Z, V) > 0$, then the individual participates in the program, else, he does not.

In the second stage, as we did in the previous sections, an outcome is realized based on the individual's decision D. If $D = 1$, $Y = Y_1$, else, $Y = Y_0$. Recall that Y_i are also functions of observables X and unobservables A, as in the previous sections. Moreover, both unobservables V and A might be correlated in some way if for example individuals have private information (inside V) about the potential success of the program (within A).

The instrument Z can have one or more dimensions, but a major question in this literature is whether Z should include at least one discrete variable or at least one continuous. In Angrist and Imbens point of view, the most convincing instrument is a single binary IV. In Heckman's point of view, a continuous IV does the job well enough.

Binary IV

The application of binary IVs come with four definitions that should be understood perfectly before going on. The graph below as well as the definitions should include enough information to understand.

Definition 11.1 (Classification of individuals)**.** *There are four classes of individuals in a given program evaluation framework. This classification relies on the individuals' participation behavior (D), based on the binary instrument (Z).*

- *If an individual will participate in the program regardless of Z, then he is an "always-taker".*
- *If an individual will not participate in the program regardless of Z, then he is a "never-taker".*

- *If an individual will participate in the program if he does not get Z, but he refuses to participate if he gets Z, then he is a "defier".*
- *If an individual will not participate in the program if he does not get Z, but he accepts to participate if he gets Z, then he is a "complier".*

Now, define $D_0 = \mathbb{I}\{\psi(0, V) > 0\}$ and $D_1 = \mathbb{I}\{\psi(1, V) > 0\}$. We can write the first-stage equation as:

$$D = (1 - Z)D_0 + ZD_1 = D_0 + (D_1 - D_0) \cdot Z$$

and thus the second-stage equation as:

$$Y = Y_0 + D_0 \cdot \Delta + (D_1 - D_0) \cdot \Delta \cdot Z$$

Assuming the binary instrumental variable Z is jointly independent of participation and outcome ($A2'''$), or formally, $Z \perp (Y_1, Y_0, D_1, D_0)$. Then,

$$\begin{aligned}
\mathrm{E}[Y|Z = 1] &= \mathrm{E}[Y_0 + D_0 \cdot \Delta + (D_1 - D_0) \cdot \Delta \cdot Z|Z = 1] \\
&= \mathrm{E}[Y_0 + D_1 \cdot \Delta|Z = 1] \\
&= \mathrm{E}[Y_0 + D_1 \cdot \Delta] \text{ (by } A2''') \\
&= \mathrm{E}[Y_0] + \mathrm{E}[D_1 \cdot \Delta]
\end{aligned}$$

and also,

$$\begin{aligned}
\mathrm{E}[Y|Z = 0] &= \mathrm{E}[Y_0 + D_0 \cdot \Delta + (D_1 - D_0) \cdot \Delta \cdot Z|Z = 0] \\
&= \mathrm{E}[Y_0 + D_0 \cdot \Delta|Z = 0] \\
&= \mathrm{E}[Y_0 + D_0 \cdot \Delta] \text{ (by } A2''') \\
&= \mathrm{E}[Y_0] + \mathrm{E}[D_0 \cdot \Delta]
\end{aligned}$$

which implies that:

$$\mathrm{E}[Y|Z = 1] - \mathrm{E}[Y|Z = 0] = \mathrm{E}[(D_1 - D_0) \cdot \Delta]$$

This last term can be simplified with assumptions about the presence of some types of individuals in the sample. In fact, consider dividing the last term in the groups defined above:

$$\begin{aligned}
\mathrm{E}[(D_1 - D_0) \cdot \Delta] = {} & 1 \cdot \mathrm{E}[\Delta|(D_1 - D_0) = 1] \cdot \Pr[(D_1 - D_0) = 1] \quad \text{(compliers)} \\
& - 1 \cdot \mathrm{E}[\Delta|(D_1 - D_0) = -1] \cdot \Pr[(D_1 - D_0) = -1] \quad \text{(defiers)} \\
& + 0 \cdot \mathrm{E}[\Delta|(D_1 - D_0) = 0] \cdot \Pr[(D_1 - D_0) = 0] \quad \text{(others)}
\end{aligned}$$

and assume that there are no defiers in the sample, formally, that $\Pr[D_1 - D_0 = -1]$ is equal to 0. Then, we get:

$$\mathrm{E}[(D_1 - D_0) \cdot \Delta] = \mathrm{E}[\Delta | (D_1 - D_0) = 1] \cdot \Pr[(D_1 - D_0) = 1]$$

$$\Leftrightarrow \frac{\mathrm{E}[(D_1 - D_0) \cdot \Delta]}{\Pr[(D_1 - D_0) = 1]} = \mathrm{E}[\Delta | (D_1 - D_0) = 1]$$

where the last term is the average treatment effect conditional on being a complier. In order to compute it, we need to know the probability of being a complier. Using the $A2'''$ assumption, one can show that:

$$\Pr[(D_1 - D_0) = 1] = \mathrm{E}[D|Z = 1] - \mathrm{E}[D|Z = 0]$$

Finally, using the implication above, we have:

$$LATE \equiv \mathrm{E}[\Delta | (D_1 - D_0) = 1] = \frac{\mathrm{E}[Y|Z = 1] - \mathrm{E}[Y|Z = 0]}{\mathrm{E}[D|Z = 1] - \mathrm{E}[D|Z = 0]}$$

which is called the Local Average Treatment Effect but actually only means the ATE for compliers.

This estimator has been heavily criticized due to the fact that it depends on the instruments chosen. In fact, the subpopulation of interest (compliers) can change if Z is different. For example, consider the unemployment training program, where the instrument would be a coupon of 500$ to selected individuals. Then, for a higher coupon value, say of 1000$, the number of compliers would change for sure, making the estimator very different.

Continuous IV

Now, assume that the instrument is continuous. We have the following two-stage model:

$$\text{(2nd stage): } Y = Y_0 + \Delta \cdot D$$
$$\text{(1st stage): } D = \mathbb{I}\{p(Z) > V\}$$

where $p(\cdot)$ is the propensity score as used in the previous sections. First, note that in this context, the no defiers condition in the binary IV case is equivalent to

the threshold structure in the first-stage of this model. Second, one could assume wlog that $V \sim U[0,1]$.

As in the previous subsection, start by looking at:

$$\begin{aligned}
\mathrm{E}[Y|Z=z] &= \mathrm{E}[Y_0 + \Delta \cdot D|Z=z] \\
&= \mathrm{E}[Y_0] + \mathrm{E}[\Delta \cdot D|Z=z] \\
&= \mathrm{E}[Y_0] + \mathrm{E}[\mathrm{E}[\Delta|Z,V] \cdot D|Z=z] \\
&= \mathrm{E}[Y_0] + \mathrm{E}[\mathrm{E}[\Delta|V] \cdot \mathbb{I}\{p(z) > V\}] \text{ (by } A2''') \\
&= \mathrm{E}[Y_0] + \int_0^{p(z)} \mathrm{E}[\Delta|V=v]\,\mathrm{d}v
\end{aligned}$$

Then, by Leibniz' rule:

$$\partial_z \mathrm{E}[Y|Z=z] = \mathrm{E}[\Delta|V=p(z)] \cdot \partial_z p(z)$$

$$\Leftrightarrow \frac{\partial_z \mathrm{E}[Y|Z=z]}{\partial_z p(z)} = \mathrm{E}[\Delta|V=p(z)]$$

which is the analog result to the LATE estimator in Angrist and Imbens' work. In words, the right-hand side term is the marginal treatment effect for the population that is indifferent between participating in the program or not for a given z. The left-hand side term is the instrumental variable at the point z. Using p in place of $p(z)$ we get:

$$\mathrm{E}[\Delta|V=p] = \partial_z \mathrm{E}[Y|P=p]$$

which we can use to get the global average treatment effect, as the integral over the marginal treatment effect for individuals that are indifferent at each level of p:

$$ATE = \int_0^1 \mathrm{E}[\Delta|V=p]\,\mathrm{d}p = \int_0^1 \partial_z \mathrm{E}[Y|P=p]\,\mathrm{d}p = \mathrm{E}[Y|P=1] - \mathrm{E}[Y|P=0]$$

This strategy has also been heavily criticized, this time based on the fact that it should be impossible to observe propensity score of exactly 1 and 0. In fact, if one uses a parametric model to estimate $p(\cdot)$, then identification would only come for $Z = \pm\infty$. We call this issue identification at infinity.

Chapter 12

Regression Discontinuity Design

to be continued...