Summary Notes

M300 'Econometric Methods'

Povilas Lastauskas*

23rd November 2013

1 Theory

Research agenda for applied economics requires to address these major issues:

1. Causal relationship;
2. Ideal experiment;
3. Identification strategy;
4. Statistical inference.

Reality manifests in many ways which complicate empirical enquiry, especially when questions about *causality* are raised. Drawing from Angrist and Pischke (2008), such questions hinge on the experimentalist paradigm and potential outcomes. This new paradigm in econometrics requires some vocabulary...

1.1 Basics of evaluation problem

Suppose we analyse a variable Y_i, described by

$$\text{Potential outcome} = \begin{cases} Y_{1i} & \text{Outcome if } i \text{ is treated;} \\ Y_{0i} & \text{Outcome if } i \text{ is not treated.} \end{cases}$$

In effect, Y_{1i} (*potential* outcome *under treatment*) and Y_{0i} (*potential* outcome *without treatment*) are outcomes in alternative states of the world. Inexistence of parallel worlds

*Errors and typos to be reported to pl312@cam.ac.uk. For the most recent version, see www.lastauskas.com.

leads to inability to measure treatment effects at the individual level. Notice that the *observed* outcome Y_i can be expressed in terms of potential ones:

$$Y_i = \begin{cases} Y_{1i}, & \text{if } D_i = 1, \\ Y_{0i}, & \text{if } D_i = 0, \end{cases}$$

or $Y_i = Y_{0i} + (Y_{1i} - Y_{0i})\, D_i$. One hope is, instead of individual ones, to analyse average treatment effects. What about a comparison of the difference in means for treated and untreated?

We obtain

$$E\left[Y_i \mid D_i = 1\right] - E\left[Y_i \mid D_i = 0\right] = E\left[Y_{1i} \mid D_i = 1\right] - E\left[Y_{0i} \mid D_i = 0\right],$$

and subtracting and adding $E\left[Y_{0i} \mid D_i = 1\right]$ produces

$$\begin{array}{c} E\left[Y_i \mid D_i = 1\right] - E\left[Y_i \mid D_i = 0\right] \\ = E\left[Y_{1i} \mid D_i = 1\right] - E\left[Y_{0i} \mid D_i = 1\right] \\ + E\left[Y_{0i} \mid D_i = 1\right] - E\left[Y_{0i} \mid D_i = 0\right]. \end{array} \tag{1}$$

The result is decomposed into two components: *average treatment effect on treated (ATT)* and *selection bias*. The latter is simply the difference in average outcome under non-treatment between those who were treated and those who were not.

Problem 1. Actual treatment status D_i is not independent of potential outcomes.

One of the solutions is *randomly assign treatment* to individuals in the population since D_i becomes independent of potential outcomes, and so the selection bias disappears. Under this,

$$\begin{array}{c} E\left[Y_i \mid D_i = 1\right] - E\left[Y_i \mid D_i = 0\right] \\ = E\left[Y_{1i} \mid D_i = 1\right] - E\left[Y_{0i} \mid D_i = 1\right], \end{array}$$

since $E\left[Y_{0i} \mid D_i = 1\right] - E\left[Y_{0i} \mid D_i = 0\right] = 0$. This enables to infer ATT - which coincides with the average treatment effect (ATE) in the entire population[1] - from the difference in means.

Under random assignment, we can obtain ATT and ATE by running OLS regressions, $Y_i = \alpha + \rho D_i + \varepsilon_i$. For a moment suppose that the treatment effect is the same for everybody, such that $Y_{1i} - Y_{0i} = \rho$. Using $Y_i = Y_{0i} + (Y_{1i} - Y_{0i})\, D_i$, we can re-express it as $Y_i = EY_{0i} + (Y_{1i} - Y_{0i})\, D_i + (Y_{0i} - EY_{0i}) = \alpha + \rho D_i + \varepsilon_i$, where $\alpha \equiv EY_{0i}$ and $\varepsilon_i = Y_{0i} - EY_{0i}$. Obviously,

$$\begin{array}{c} E\left[Y_i \mid D_i = 1\right] - E\left[Y_i \mid D_i = 0\right] = \rho \\ + E\left[\varepsilon_i \mid D_i = 1\right] - E\left[\varepsilon_i \mid D_i = 0\right]. \end{array}$$

Claim 2. The *selection bias* amounts to non-zero correlation between the regression error term ε_i and the regressor D_i: what you previously called *endogeneity bias* is known as *selection bias* here.

[1] Observe that $E\left[Y_{1i} \mid D_i = 1\right] - E\left[Y_{0i} \mid D_i = 1\right] = E\left[Y_{1i}\right] - E\left[Y_{0i}\right] = \text{ATE}$.

The correlation reflects the difference in potential outcomes (under no treatment) between those who get treated and those who don't. To see this, note that $E[\varepsilon_i \mid D_i = 1] - E[\varepsilon_i \mid D_i = 0] = E[Y_{0i} \mid D_i = 1] - E[Y_{0i} \mid D_i = 0]$. Clearly, if D_i is randomly assigned, there is no selection bias so that a regression of observed outcomes Y_i on actual treatment status D_i estimates the causal effect.

1.2 Basics of Regression

The Conditional Expectation Function (CEF) for a dependent variable Y_i, given covariates X_i, is the expectation - or the population average - of Y_i with X_i held fixed. Denote by[2]

$$E[Y_i \mid X_i],$$

from which it clearly follows that CEF is random as it is a function of X_i which is random. Further, it is a *population* concept which the researcher attempts to uncover using sample CEF. There is a number of useful CEF properties:

- CEF decomposition property.

By decomposition property,

$$Y_i = E[Y_i \mid X_i] + \varepsilon_i,$$

where ε_i is mean independent of X_i, $E[\varepsilon_i \mid X_i] = 0$. Notice that this follows from

$$E[\varepsilon_i \mid X_i] = E[Y_i \mid X_i] - E[[E[Y_i \mid X_i]] \mid X_i] = 0.$$

More importantly, we can demonstrate that this property produces a result that ε_i is uncorrelated with any function of X_i. Let $h(X_i)$ be any function of X_i, then

$$E[h(X_i)\varepsilon_i] = E[E[(h(X_i)\varepsilon_i) \mid X_i]] = E[h(X_i)E(\varepsilon_i \mid X_i)] = 0.$$

In demonstrating this relationship, we employed what is known as the Law of Iterated Expectations - stating that an unconditional expectation can be written as the unconditional average of the CEF - or

$$\begin{aligned} E_x[E_y[y \mid x]] &= \int E_y[y \mid x] f(x)\,\mathrm{d}x = \int \left[\int y f(y \mid x)\,\mathrm{d}y\right] f(x)\,\mathrm{d}x = \int\int y f(y \mid x) f(x)\,\mathrm{d}y\,\mathrm{d}x \\ &= \int\int y f(y,x)\,\mathrm{d}y\,\mathrm{d}x = \int\int y f(y,x)\,\mathrm{d}x\,\mathrm{d}y = \int y \int f(y,x)\,\mathrm{d}x\,\mathrm{d}y = \int y f(y)\,\mathrm{d}y \\ &= E[y]. \end{aligned}$$

- Best predictor property.

$E[Y_i \mid X_i]$ is the Best (Minimum mean squared error MMSE) predictor of Y_i in that it minimises the function $E(Y_i - h(X_i))^2$, or

$$E[Y_i \mid X_i] = \arg\min_{h(X_i)} E\left[(Y_i - h(X_i))^2\right],$$

[2] In continuous case, CEF is (with a slight abuse of notation) $E[Y_i \mid X_i = x] = \int Y_i f_Y(Y_i \mid X_i = x)\,dY_i$,whereas in discrete case it is $E[Y_i \mid X_i = x] = \sum Y_i P(Y_i \mid X_i = x)$.

where $h(X_i)$ is any function of X_i. The proof follows immediately by subtracting and adding $E[Y_i \mid X_i]$ inside the brackets, so that

$$E(Y_i - h(X_i))^2$$
$$= E\left([Y_i - E[Y_i \mid X_i]]^2 - 2[Y_i - E[Y_i \mid X_i]][E[Y_i \mid X_i] - h(X_i)] + [E[Y_i \mid X_i] - h(X_i)]^2\right).$$

The first term does not involve $E[Y_i \mid X_i] - h(X_i)$, the second one is zero from the decomposition property, and the last one is minimised at zero, yielding a result that $h(X_i)$ is CEF.[3]

We have not specified $h(X_i)$ so far and not really linked CEF to *regression*. Define population regression parameter vector β as the solution to the following minimisation problem:[4]

$$\beta = \arg\min_{\hat{\beta}} E\left[\left(Y_i - X_i'\hat{\beta}\right)^2\right].$$

The FOCs yield

$$E\left[X_i\left(Y_i - X_i'\hat{\beta}\right)\right] = 0,$$

which can be used to produce $\beta = E[X_i X_i']^{-1} E[X_i Y_i]$. An unbiased estimator results in $E\hat{\beta} = \beta$ and its (asymptotic) variance-covariance matrix is $E[X_i X_i']^{-1} E[X_i X_i' \varepsilon_i^2] E[X_i X_i']^{-1}$. A useful property is that a linear predictor takes the form

$$E[Y_i \mid X_i] = X_i'\beta,$$

and the solution to the the Best Linear Predictor (BLP) problem, i.e., $\min E(Y_i - E[Y_i \mid X_i])^2$, yields β. *If the CEF is linear, then the population regression function is exactly it.* CEF is linear when Y and X are jointly normal, also when the model is saturated, i.e., model with a separate parameter for *every possible combination* of values that the regressors can take.

2 Examples

2.1 Linear Regression

Consider the linear regression model

$$E[y_i \mid x_i] = \alpha + \beta x_i,$$

[3] See Angrist and Pischke (2008) for the ANOVA theorem and yet another CEF property.

[4] Wooldridge (2010) approaches this by using the linear projection of Y_i on X_i. Let $E[Y_i \mid X_i]$ denote the true regression function. The linear projection of Y_i on X_i, denoted

$$L(Y_i \mid X_i) = X_i'\beta$$

is such that the parameters solve $\arg\min_{\hat{\beta}} E\left[\left(Y_i - X_i'\hat{\beta}\right)^2\right]$. In other words, the parameters in the linear projection $L(Y_i \mid X_i)$ provide the best (population) mean square error approximation to the true regression function.

where $x = 1$ if an individual belongs to group 1 and $x = 0$ if the individual is from group 2. A random sample $(y_i,\, x_i)$, $(i = 1,\, \ldots,\, n)$, of observations are available.

Find Least Squares estimators of α and β. Moreover, show that b can be written as $\bar{y}_1 - \bar{y}_2$, where $\bar{y}_j$ is the average of the observations from group j, $(j = 1,\, 2)$.

Solution

The OLS estimator for β in the model $y = \alpha + \beta x + \varepsilon$ is

$$b = \frac{\sum xy - n^{-1} \sum x \sum y}{\sum x^2 - n^{-1} \left(\sum x\right)^2} \,. \tag{2}$$

Let n_1 be the number of observations in group 1 (where $x = 1$), so there are $n_2 = n - n_1$ observations in group 2 (where $x = 0$). As x is a dummy variable we have that $\sum x = \sum_{x=1} x = \sum x^2 = n_1$. Thus, (2) becomes

$$b = \frac{\sum xy - n^{-1} n_1 \sum y}{n_1 - n^{-1}(n_1)^2} = \frac{n \sum xy - n_1 \sum y}{n_1 n_2} \,. \tag{3}$$

Since $\sum xy = \sum_{x=1} y$ and $\sum y = \sum_{x=0} y + \sum_{x=1} y$, then (2) can be simplified into

$$\begin{aligned} b &= \frac{n \sum_{x=1} y - n_1 \sum_{x=0} y - n_1 \sum_{x=1} y}{n_1 n_2} = \frac{n_2 \sum_{x=1} y - n_1 \sum_{x=0} y}{n_1 n_2} \\ &= \frac{\sum_{x=1} y}{n_1} - \frac{\sum_{x=0} y}{n_2} = \bar{y}_1 - \bar{y}_2 \,, \end{aligned} \tag{4}$$

as requested. Finally, the estimator for α is simply

$$a = \bar{y} - b\bar{x} = \frac{n_1 \bar{y}_1 + n_2 \bar{y}_2}{n} - \frac{(\bar{y}_1 - \bar{y}_2) n_1}{n} = \frac{(n_1 + n_2)\bar{y}_2}{n} = \bar{y}_2 \,. \tag{5}$$

A Method of Moments interpretation: Since x takes only two values we can derive the expectation of y conditional on these values (2 moments) to identify the two parameters α and β. We have that

$$\begin{aligned} E[y \,|\, x = 1] &= \alpha + \beta E[x \,|\, x = 1] + E[\varepsilon \,|\, x = 1] = \alpha + \beta \\ E[y \,|\, x = 0] &= \alpha + \beta E[x \,|\, x = 0] + E[\varepsilon \,|\, x = 0] = \alpha \,, \end{aligned}$$

which follows from the usual condition of no correlation between x and ε, $E[\varepsilon \,|\, x] = 0$ (for all values of x). Then, we can replace the population conditional moments by their sample counterparts and the parameters by their estimators. Note that $\bar{y}_1$ is the sample estimator of $E[y \,|\, x = 1]$ and $\bar{y}_2$ is the sample moment corresponding to $E[y \,|\, x = 0]$. Thus, the moment conditions are $\bar{y}_1 = a + b$ and $\bar{y}_2 = a$. Upon solving this system we get the same answer as before, $b = \bar{y}_1 - \bar{y}_2$.

The interpretation of the OLS estimator as a MoM estimator can be generalised to any linear model, not just to this particular problem.

2.2 Densities

The joint density function of two random variables, X and Y, is given by

$$f_{XY}(x,\, y) = \begin{cases} y^2 x e^{-xy} & \text{if } x > 0,\ 0 < y < 1, \\ 0 & \text{otherwise.} \end{cases}$$

Show that the marginal density function of X is

$$f(x) \;\; = \frac{2 - \mathrm{e}^{-x}\left(x^2 + 2x + 2\right)}{x^2},$$

and that Y is uniform distributed on $[0,\, 1]$. Also find the conditional density of X given Y. Are X and Y are independent?

Solution

The marginal density of x is defined as $f(x) = \int_0^1 f(x, y)\, \mathrm{d}y$. Using the functional form provided and integrating by parts twice we get

$$\begin{aligned} f(x) &= x \int_0^1 y^2 \mathrm{e}^{-xy} dy = x \left[-y^2 \frac{\mathrm{e}^{-xy}}{x} \right]_0^1 + 2x \int_0^1 y \frac{\mathrm{e}^{-xy}}{x}\, \mathrm{d}y \\ &= -\mathrm{e}^{-x} + 2 \left(\left[-y \frac{\mathrm{e}^{-xy}}{x} \right]_0^1 + \int_0^1 \frac{\mathrm{e}^{-xy}}{x}\, \mathrm{d}y \right) = -\mathrm{e}^{-x} + 2 \left(-\frac{\mathrm{e}^{-x}}{x} - \frac{1}{x} \left[\frac{\mathrm{e}^{-xy}}{x} \right]_0^1 \right) \\ &= -\mathrm{e}^{-x} - 2\frac{e^{-x}}{x} - 2\frac{\mathrm{e}^{-x} - 1}{x^2} = \frac{2 - \mathrm{e}^{-x}\left(x^2 + 2x + 2\right)}{x^2}, \end{aligned}$$

as requested. The marginal density of y is $f(y) = \int_0^\infty f(x, y) \mathrm{d}x$, hence

$$f(y) = y^2 \int_0^\infty x \mathrm{e}^{-xy}\, \mathrm{d}x = y^2 \left[-x \frac{\mathrm{e}^{-xy}}{y} \right]_0^\infty + y^2 \int_0^\infty \frac{\mathrm{e}^{-xy}}{y}\, \mathrm{d}x = y^2 \cdot 0 - y \left[\frac{\mathrm{e}^{-xy}}{y} \right]_0^\infty = 1,$$

so y is uniformly distributed. Therefore,

$$f(x \mid y) = \frac{f(x, y)}{f(y)} = \frac{y^2 x \mathrm{e}^{-xy}}{1} \neq f(x),$$

thus x and y are not independent.

2.3 CEF and Joint Normality

Suppose Z_1 and Z_2 are two standard normal variables and

$$\begin{aligned} X_1 &= \mu_1 + \sigma_1 Z_1 \\ X_2 &= \mu_2 + \sigma_2 \left(\rho Z_1 + \sqrt{1 - \rho^2} Z_2 \right). \end{aligned}$$

Hence, $(X_1;\, X_2)$ are bivariate normal (such that $X_i \sim N(\mu_i,\, \sigma_i^2)$). The covariance between $(X_1;\, X_2)$ is $\rho\sigma_1\sigma_2$. Recalling the OLS expressions for estimators,

$$\hat{\alpha} = \mu_2 - \mu_1\hat{\beta}$$
$$\hat{\beta} = Cov(X_1;\, X_2)/Var(X_1) = \rho\sigma_2/\sigma_1.$$

Let the linear predictor be denoted by $E^{\star}(X_2 \mid X_1) = \alpha + \beta X_1$. Then, the best linear predictor is CEF,

$$\begin{aligned} &E^{\star}(X_2 \mid X_1) \\ &= \mu_2 - \mu_1\beta + \beta X_1 \\ &= \mu_2 + \beta(X_1 - \mu_1). \end{aligned}$$

Note that employing properties of the Normal distribution,

$$E(X_2 \mid X_1) = \mu_2 + \beta(X_1 - \mu_1).$$

So the CEF is linear in this case (this is not true with other distributions.).

3 A Few Useful Results Using Matrix Algebra

This is to summarise a few important and useful results widely used in econometric theory. Let's start with a simple multiple regression

$$y = x_1\hat{\beta}_1 + x_2\hat{\beta}_2 + \ldots + x_k\hat{\beta}_k,$$

where $\boldsymbol{x}$ and $\hat{\boldsymbol{\beta}}$ are k-dimensional vectors. Further, define the product of two vectors to be $\boldsymbol{x}\cdot\hat{\boldsymbol{\beta}} = \sum x_i\hat{\beta}_i$ or a dot product (inner product). Two vectors are orthogonal if their dot product equals zero, meaning graphically that the vectors are perpendicular.

Going to more statistical interpretation, let $\boldsymbol{x}$ and $\boldsymbol{y}$ be two random variables, each with mean zero, and N elements. We can then construct a vector $\boldsymbol{x} = (x_1,\, x_2, \ldots,\, x_N)$ and a similar vector $\boldsymbol{y} = (y_1,\, y_2, \ldots,\, y_N)$. When we take their dot product, we calculate: $\boldsymbol{x}\cdot\boldsymbol{y} = \sum x_i y_i = (N-1)\,\hat{Cov}(x,\, y)$ and $\boldsymbol{x}\cdot\boldsymbol{x} = \sum x_i x_i = (N-1)\,\hat{Var}(x)$.[5]

A matrix $\boldsymbol{A}$ is defined as a collection of $n \times k$ entries arranged into n rows and k columns. Given an $n \times k$ matrix $\boldsymbol{A}$ with the entries described as above, the transpose of $\boldsymbol{A}$ is the $k \times n$ matrix $\boldsymbol{A}'$ that results from interchanging the columns and rows of $\boldsymbol{A}$. Matrix multiplication is only defined between an $n \times k$ matrix A and an $k \times n$ matrix $\boldsymbol{B}$, and *the order does matter*. Two useful types of matrices include a symmetric matrix that is the same as its transpose, $\boldsymbol{A} = \boldsymbol{A}'$ and idempotent matrices that are the same when multiplied by themselves, $\boldsymbol{A}\boldsymbol{A} = \boldsymbol{A}$. See Abadir and Magnus (2005) for excellent treatment of linear algebra and many applications useful in econometrics.

We generally deal with matrix inverses only in theory, so it's important to know some theoretical properties of inverses. I'll add some rules for transposes as well, since they mirror the others:

[5] Geometrically, two vectors in N-dimensional space, the angle θ between them must always satisfy: $\cos\theta = \frac{\boldsymbol{x}\cdot\boldsymbol{y}}{\sqrt{\boldsymbol{x}\cdot\boldsymbol{x}}\sqrt{\boldsymbol{y}\cdot\boldsymbol{y}}} = \hat{Corr}(x,\, y)$. The cosine of two rays is one if they point in exactly the same direction, zero if they are perpendicular, negative one if they point in exactly opposite directions - exactly the same as with correlations

- $\left(\boldsymbol{A}^{-1}\right)^{-1} = \boldsymbol{A}$, $(\boldsymbol{AB})^{-1} = \boldsymbol{B}^{-1}\boldsymbol{A}^{-1}$, $(\boldsymbol{A}')^{-1} = \left(\boldsymbol{A}^{-1}\right)'$,
- $(\boldsymbol{A}')' = \boldsymbol{A}$, $(\boldsymbol{AB})' = \boldsymbol{B}'\boldsymbol{A}'$, $(\boldsymbol{A}+\boldsymbol{B})' = \boldsymbol{A}'+\boldsymbol{B}'$,

where the rule $(\boldsymbol{AB})^{-1} = \boldsymbol{B}^{-1}\boldsymbol{A}^{-1}$, works only when each matrix is a square matrix. Also, matrices can be used for operations applied for linear functions. For instance, $y = ax$ yields $dy/dx = a$ whereas $\boldsymbol{y} = \boldsymbol{Ax}$ yields $\partial \boldsymbol{y}/\partial \boldsymbol{x}' = \boldsymbol{A}$.

An extension to quadratic functions is also straightforward. In a matrix representation that accounts to $\boldsymbol{y} = \boldsymbol{x}'\boldsymbol{Ax}$.[6] Then $\partial \boldsymbol{y}/\partial \boldsymbol{x}' = 2\boldsymbol{x}'\boldsymbol{A}$.[7]

Having all this basic notation and language, we can construct a system of equations with a vector $\boldsymbol{y} = (y_1, y_2, \ldots, y_N)'_{N\times 1}$ containing all of the y values, also parameter vector $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_K)'_{K\times 1}$ and matrix

$$X = \begin{bmatrix} 1 & x_{12} & \cdots & x_{1K} \\ 1 & x_{22} & \cdots & x_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{N2} & \cdots & x_{NK} \end{bmatrix}_{N\times K}$$

and a vector $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_N)'_{N\times 1}$ containing all of the "unobservable" determinants of the outcome $\boldsymbol{y}$. The system of equations can be represented as: $\boldsymbol{y}_{N\times 1} = \boldsymbol{X}_{N\times K}\boldsymbol{\beta}_{K\times 1} + \boldsymbol{\varepsilon}_{N\times 1}$. Its econometric model is $\hat{\boldsymbol{\varepsilon}}_{N\times 1} = \boldsymbol{y}_{N\times 1} - \boldsymbol{X}_{N\times K}\hat{\boldsymbol{\beta}}_{K\times 1}$, where residuals $\hat{\boldsymbol{\varepsilon}}_{N\times 1}$ are obtained once $\hat{\boldsymbol{\beta}}_{K\times 1}$ is estimated. We want residuals to be such that their size (or norm) of the vector $\hat{\boldsymbol{\varepsilon}}$ is minimised, i.e., $\min \|\hat{\boldsymbol{\varepsilon}}\| = \min \sqrt{\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}}$ or, more familiarly, $\min \|\hat{\boldsymbol{\varepsilon}}\|^2 = \min \hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}$. In other words, we want to minimise the following expression $\min_{\hat{\beta}} \hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}} = \left(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\right)'\left(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\right) = \left(\boldsymbol{y}' - \hat{\boldsymbol{\beta}}'\boldsymbol{X}'\right)\left(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\right) = \boldsymbol{y}'\boldsymbol{y} - \boldsymbol{y}'\boldsymbol{X}\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{y} + \hat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}}$. The optimal $\hat{\boldsymbol{\beta}}^{OLS}$ solves: $-\boldsymbol{y}'\boldsymbol{X} - (\boldsymbol{X}'\boldsymbol{y})' + 2\hat{\boldsymbol{\beta}}^{OLS}\boldsymbol{X}'\boldsymbol{X} = \boldsymbol{0}$ or $\hat{\boldsymbol{\beta}}^{OLS} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$, where $(\boldsymbol{X}'\boldsymbol{y})' = \boldsymbol{y}'\boldsymbol{X}$.

From this expression, notice that $\hat{\boldsymbol{\varepsilon}} = \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{y} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} = \left(\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\right)\boldsymbol{y}$. We can decompose $\boldsymbol{y}$ into two components: the orthogonal projection onto the K dimensional space spanned by $\boldsymbol{X}$, $\boldsymbol{X}\hat{\boldsymbol{\beta}}$ and the component that is the orthogonal projection onto the $n-K$ subspace that is orthogonal to the span of $\boldsymbol{X}$, $\hat{\boldsymbol{\varepsilon}}$. Since $\hat{\boldsymbol{\beta}}$ is chosen to make $\hat{\boldsymbol{\varepsilon}}$ as short as possible, $\hat{\boldsymbol{\varepsilon}}$ will be orthogonal to the space spanned by $\boldsymbol{X}$ as in this space, $\boldsymbol{X}'\hat{\boldsymbol{\varepsilon}} = 0$. The FOCs that define the least squares estimator imply that this is so.

3.1 Projection Matrices

We have that $\boldsymbol{X}\hat{\boldsymbol{\beta}}$ is the projection of $\boldsymbol{y}$ on the span of $\boldsymbol{X}$ or, $\boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} = \boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{y}$. Then, $\hat{\boldsymbol{\varepsilon}}$ is the projection of $\boldsymbol{y}$ off the space spanned by $\boldsymbol{X}$, in other words, onto the space that is orthogonal to the span of $\boldsymbol{X}$): $\hat{\boldsymbol{\varepsilon}} = \left(\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\right)\boldsymbol{y} = (\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{y} =$

[6] For a simple illustration, suppose $\boldsymbol{y} = \boldsymbol{x}'\boldsymbol{Ax} = ax_1^2 + 2bx_1x_2 + cx_2^2$. Its partial derivatives wrt to x_1 and x_2, respectively, are simply $2(ax_1 + bx_2)$ and $2(bx_1 + cx_2)$.

[7] Under symmetry, otherwise $\partial \boldsymbol{y}/\partial \boldsymbol{x}' = \boldsymbol{x}'\left(\boldsymbol{A} + \boldsymbol{A}'\right)$.

$\boldsymbol{M_X y}$. Therefore, $\boldsymbol{y} = \boldsymbol{P_X y} + \boldsymbol{M_X y} = (\boldsymbol{P_X} + \boldsymbol{M_X})\,\boldsymbol{y}$. Note that both $\boldsymbol{P_X}$ and $\boldsymbol{M_X}$ are symmetric and idempotent ($\boldsymbol{P_X P_X} = \boldsymbol{P_X}$ and $\boldsymbol{M_X M_X} = \boldsymbol{M_X}$).

To determine the goodness-of-fit, also use the expression $\boldsymbol{y'y} = \left(\hat{\boldsymbol{\beta}}'\boldsymbol{X}' + \hat{\boldsymbol{\varepsilon}}'\right)\left(\boldsymbol{X}\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{\varepsilon}}\right) = \hat{\boldsymbol{\beta}}'\boldsymbol{X'X}\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{\beta}}'\boldsymbol{X}'\hat{\boldsymbol{\varepsilon}} + \hat{\boldsymbol{\varepsilon}}'\boldsymbol{X}\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}} = \hat{\boldsymbol{\beta}}'\boldsymbol{X'X}\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}$ since $\boldsymbol{X}'\hat{\boldsymbol{\varepsilon}} = \boldsymbol{0}$. Then uncentred R_U^2 is defined as $R_U^2 = 1 - \hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}/\boldsymbol{y'y} = \hat{\boldsymbol{\beta}}'\boldsymbol{X'X}\hat{\boldsymbol{\beta}}/\boldsymbol{y'y} = \|\boldsymbol{P_X y}\|^2 / \|\boldsymbol{y}\|^2 = \cos^2\theta$, where θ is the angle between $\boldsymbol{y}$ and the span of $\boldsymbol{X}$. For a more usual centred coefficient of determination, introduce the n-vector $\boldsymbol{i} = (1, 1, \ldots, 1)'$ which we can use in forming $\boldsymbol{M_i} = I_n - \boldsymbol{i}\left(\boldsymbol{i'i}\right)^{-1}\boldsymbol{i}' = I_n - \boldsymbol{ii}'/n$. Obviously, $\boldsymbol{M_i y}$ gives the vector of deviations from the mean. Thus, $R_C^2 = 1 - \hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}/\boldsymbol{y'M_i y} = 1 - ESS/TSS$. Recalling that we construct residuals to average to zero (when a constant is included), $\boldsymbol{M_i}\hat{\boldsymbol{\varepsilon}} = \hat{\boldsymbol{\varepsilon}}$.

However, the true power of this approach is best seen in more complex environments, such as dealing with instrumental variable estimation. Suppose $E\left(\boldsymbol{X'\varepsilon}\right) \neq 0$ due to simultaneity, omitted variables or errors-in-variables. Then consider some matrix $\boldsymbol{Z}$ which is formed of variables uncorrelated with $\boldsymbol{\varepsilon}$. This matrix defines a projection matrix $\boldsymbol{P_Z} = \boldsymbol{Z}\left(\boldsymbol{Z'Z}\right)^{-1}\boldsymbol{Z}$, so that anything that is projected onto the space spanned by $\boldsymbol{Z}$ will be uncorrelated with $\boldsymbol{\varepsilon}$ by the definition of $\boldsymbol{Z}$. Then transform the original model $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$ into

$$\boldsymbol{P_Z y} = \boldsymbol{P_Z X\beta} + \boldsymbol{P_Z\varepsilon},$$

and observe that $\boldsymbol{P_Z X} = \boldsymbol{Z}\left(\boldsymbol{Z'Z}\right)^{-1}\boldsymbol{Z'X}$ is the fitted value from a regression of $\boldsymbol{X}$ on $\boldsymbol{Z}$ ("first stage") and $E\left(\left(\boldsymbol{P_Z X}\right)'\boldsymbol{P_Z\varepsilon}\right) = E\left(\boldsymbol{X'P_Z\varepsilon}\right) = 0$. This is the *generalised instrumental variables estimator*, defined as

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{\boldsymbol{IV}} &= \left(\boldsymbol{X'P_Z X}\right)^{-1}\boldsymbol{X'P_Z y}\\ &= \left(\boldsymbol{X'P_Z X}\right)^{-1}\boldsymbol{X'P_Z}\left(\boldsymbol{X\beta} + \boldsymbol{\varepsilon}\right)\\ &= \boldsymbol{\beta} + \left(\boldsymbol{X'P_Z X}\right)^{-1}\boldsymbol{X'P_Z\varepsilon},\end{aligned}$$

and the bias given by $\left(\left(\boldsymbol{X'Z}\right)\left(\boldsymbol{Z'Z}\right)^{-1}\left(\boldsymbol{Z'X}\right)\right)^{-1}\left(\boldsymbol{X'Z}\right)\left(\boldsymbol{Z'Z}\right)^{-1}\boldsymbol{Z'\varepsilon}$. However, dividing each term by N, and applying LLN, we can demonstrate that all terms go to finite matrices while $\left(\boldsymbol{Z'\varepsilon}\right)/N \xrightarrow{p} 0$, stemming from $E\left(\boldsymbol{Z'\varepsilon}\right) = 0$. Hence, $\hat{\boldsymbol{\beta}}_{\boldsymbol{IV}} \xrightarrow{p} \boldsymbol{\beta}$. Similarly, CLT can be invoked by scaling $\hat{\boldsymbol{\beta}}_{\boldsymbol{IV}} - \boldsymbol{\beta}$ by $\sqrt{N}$. Hence, IV estimator is consistent, asymptotically normally distributed but biased in general, since even though $E\left(\boldsymbol{X'P_Z\varepsilon}\right) = 0$, $E\left(\left(\boldsymbol{X'P_Z X}\right)^{-1}\boldsymbol{X'P_Z\varepsilon}\right)$ may not be zero, since $\left(\boldsymbol{X'P_Z X}\right)^{-1}$ and $\boldsymbol{X'P_Z\varepsilon}$ are not independent.

4 Miscellaneous Examples

The estimators.[8]

In the simple linear regression model, $\underline{\mathrm{X}}$ has two columns: a vector of ones and a vector containing the explanatory variable $\underline{\mathrm{x}}$. By working on the general formula for the

[8] To alleviate the notation we will use $\sum$, instead of $\sum_{i=1}^{n}$.

OLS estimator we get

$$\underline{\mathrm{b}} = \left(\underline{\mathrm{X}}'\underline{\mathrm{X}}\right)^{-1}\underline{\mathrm{X}}'\underline{\mathrm{y}} = \left(\begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_n \end{bmatrix}\begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}\right)^{-1}\begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_n \end{bmatrix}\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

$$= \begin{bmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{bmatrix}^{-1}\begin{bmatrix} \sum y_i \\ \sum x_i y_i \end{bmatrix}.$$

Take the inverse of the 2×2 matrix $\left(\underline{\mathrm{X}}'\underline{\mathrm{X}}\right)$,

$$\underline{\mathrm{b}} = \frac{1}{\left|\underline{\mathrm{X}}'\underline{\mathrm{X}}\right|}\begin{bmatrix} \sum x_i^2 & -\sum x_i \\ -\sum x_i & n \end{bmatrix}\begin{bmatrix} \sum y_i \\ \sum x_i y_i \end{bmatrix} = \frac{1}{n\sum x_i^2 - \left(\sum x_i\right)^2}\begin{bmatrix} \sum x_i^2 \sum y_i - \sum x_i \sum x_i y_i \\ -\sum x_i \sum y_i + n\sum x_i y_i \end{bmatrix}.$$

Consider the second element of vector $\underline{\mathrm{b}}$,

$$b_2 = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum x_i^2 - \left(\sum x_i\right)^2} = \frac{\sum x_i y_i - \sum x_i\left(\frac{1}{n}\sum y_i\right)}{\sum x_i^2 - \sum x_i\left(\frac{1}{n}\sum x_i\right)} = \frac{\sum x_i\left(y_i - \bar{y}\right)}{\sum x_i\left(x_i - \bar{x}\right)} = \frac{\sum\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{\sum\left(x_i - \bar{x}\right)^2}.$$

As for the other element, we have found that

$$b_1 = \frac{\sum x_i^2 \sum y_i - \sum x_i \sum x_i y_i}{n\sum x_i^2 - \left(\sum x_i\right)^2}.$$

By working on the numerator (add and subtract $\frac{1}{n}\left(\sum x_i\right)^2 \sum y_i$),

$$\sum x_i^2 \sum y_i - \sum x_i \sum x_i y_i = \left(n\sum x_i^2 - \left(\sum x_i\right)^2\right)\bar{y} - \left(n\sum x_i y_i - \sum x_i \sum y_i\right)\bar{x}.$$

Therefore,

$$b_1 = \frac{\left(n\sum x_i^2 - \left(\sum x_i\right)^2\right)\bar{y} - \left(n\sum x_i y_i - \sum x_i \sum y_i\right)\bar{x}}{n\sum x_i^2 - \left(\sum x_i\right)^2} = \bar{y} - b_2\bar{x}.$$

Standard errors.

To find the covariance matrix of $\underline{\mathrm{b}}$, note that $\underline{\mathrm{b}} = \boldsymbol{\beta} + \left(\underline{\mathrm{X}}'\underline{\mathrm{X}}\right)^{-1}\underline{\mathrm{X}}'\boldsymbol{\varepsilon}$, so

$$E\left[\left(\underline{\mathrm{b}} - \boldsymbol{\beta}\right)\left(\underline{\mathrm{b}} - \boldsymbol{\beta}\right)'\right] = E\left[\left(\underline{\mathrm{X}}'\underline{\mathrm{X}}\right)^{-1}\underline{\mathrm{X}}'\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\underline{\mathrm{X}}\left(\underline{\mathrm{X}}'\underline{\mathrm{X}}\right)^{-1}\right]$$

$$= \left(\underline{\mathrm{X}}'\underline{\mathrm{X}}\right)^{-1}\underline{\mathrm{X}}'E\left[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\right]\underline{\mathrm{X}}\left(\underline{\mathrm{X}}'\underline{\mathrm{X}}\right)^{-1} = \sigma^2\left(\underline{\mathrm{X}}'\underline{\mathrm{X}}\right)^{-1},$$

which follows from the facts that $\underline{\mathrm{X}}$ is fixed in repeated samples and $E\left[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\right] = \sigma^2\underline{\mathrm{I}}$.[9] Thus, for the simple regression model,

$$E\left[\left(\underline{\mathrm{b}} - \boldsymbol{\beta}\right)\left(\underline{\mathrm{b}} - \boldsymbol{\beta}\right)'\right] = \frac{\sigma^2}{n\sum x_i^2 - \left(\sum x_i\right)^2}\begin{bmatrix} \sum x_i^2 & -\sum x_i \\ -\sum x_i & n \end{bmatrix}.$$

[9] If $\underline{\mathrm{X}}$ is stochastic then all the moments are defined *conditional on* $\underline{\mathrm{X}}$. For example, $E\left[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}' \mid \underline{\mathrm{X}}\right] = \sigma^2\underline{\mathrm{I}}$.

The standard error of b_1 is given by the square root of the element $(1,1)$ of the covariance matrix, while the covariance between b_1 and b_2 is given by its off-diagonal element,

$$\text{se}(b_1) = \sqrt{\frac{\sigma^2 \sum x_i^2}{n \sum x_i^2 - \left(\sum x_i\right)^2}}\,, \qquad \text{and} \qquad \text{cov}(b_1, b_2) = \frac{-\sigma^2 \sum x_i}{n \sum x_i^2 - \left(\sum x_i\right)^2}\,.$$

Altering the regressors

Consider the multiple regression model $\underline{\mathrm{y}} = \underline{\mathrm{X}}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where $\boldsymbol{\beta}$ is a $k \times 1$ vector, and the linear transformation $\underline{\mathrm{Z}} = \underline{\mathrm{X}}\underline{\mathrm{A}}$ where $\underline{\mathrm{A}}$ is a $k \times k$ nonsingular matrix. In a multiple regression model the estimated parameter vector is $\underline{\mathrm{b}} = (\underline{\mathrm{X}}'\underline{\mathrm{X}})^{-1}\underline{\mathrm{X}}'\underline{\mathrm{y}}$ and the residuals can be calculated as $\underline{\mathrm{e}} = \underline{\mathrm{y}} - \hat{\boldsymbol{y}} = \underline{\mathrm{y}} - \underline{\mathrm{X}}\underline{\mathrm{b}}$. Now we regress $\underline{\mathrm{y}}$ on $\underline{\mathrm{Z}}$. The estimated parameter vector becomes

$$\boldsymbol{b_A} = (\underline{\mathrm{Z}}'\underline{\mathrm{Z}})^{-1}\underline{\mathrm{Z}}'\underline{\mathrm{y}} = ((\underline{\mathrm{X}}\underline{\mathrm{A}})'(\underline{\mathrm{X}}\underline{\mathrm{A}}))^{-1}(\underline{\mathrm{X}}\underline{\mathrm{A}})'\underline{\mathrm{y}} = (\underline{\mathrm{A}}'\underline{\mathrm{X}}'\underline{\mathrm{X}}\underline{\mathrm{A}})^{-1}\underline{\mathrm{A}}'\underline{\mathrm{X}}'\underline{\mathrm{y}}\,.$$

Note that both $\underline{\mathrm{A}}$ and $(\underline{\mathrm{X}}'\underline{\mathrm{X}})$ are square $(k \times k)$ and nonsingular. Remember that if $\underline{\mathrm{M}}, \underline{\mathrm{N}}, \underline{\mathrm{P}}$ are square and nonsingular, then $(\underline{\mathrm{M}}\underline{\mathrm{N}}\underline{\mathrm{P}})^{-1} = \underline{\mathrm{P}}^{-1}\underline{\mathrm{N}}^{-1}\underline{\mathrm{M}}^{-1}$. Thus, $\boldsymbol{b_A}$ becomes

$$\boldsymbol{b_A} = \underline{\mathrm{A}}^{-1}(\underline{\mathrm{X}}'\underline{\mathrm{X}})^{-1}\underline{\mathrm{A}}'^{-1}\underline{\mathrm{A}}'\underline{\mathrm{X}}'\underline{\mathrm{y}} = \underline{\mathrm{A}}^{-1}(\underline{\mathrm{X}}'\underline{\mathrm{X}})^{-1}\underline{\mathrm{X}}'\underline{\mathrm{y}} = \underline{\mathrm{A}}^{-1}\underline{\mathrm{b}}\,.$$

There is a linear relation between the coefficient estimated in the two regressions. Replace now the definition of $\underline{\mathrm{Z}}$ and $\boldsymbol{b_A}$ in the residuals for this new regression to see that they are the same as those in the original regression ($\underline{\mathrm{y}}$ on $\underline{\mathrm{X}}$),[10]

$$\boldsymbol{e_A} = \underline{\mathrm{y}} - \underline{\mathrm{Z}}\boldsymbol{b_A} = \underline{\mathrm{y}} - (\underline{\mathrm{X}}\underline{\mathrm{A}})(\underline{\mathrm{A}}^{-1}\underline{\mathrm{b}}) = \underline{\mathrm{y}} - \underline{\mathrm{X}}\underline{\mathrm{b}} = \underline{\mathrm{e}}\,.$$

[10] This is due to the fact that the *projections matrices* in both models are the same. We have that $\boldsymbol{P_X} = \underline{\mathrm{X}}(\underline{\mathrm{X}}'\underline{\mathrm{X}})^{-1}\underline{\mathrm{X}}'$ and $\boldsymbol{P_A} = \underline{\mathrm{Z}}(\underline{\mathrm{Z}}'\underline{\mathrm{Z}})^{-1}\underline{\mathrm{Z}}' = \underline{\mathrm{Z}}(\underline{\mathrm{Z}}'\underline{\mathrm{Z}})^{-1}\underline{\mathrm{Z}}' = \underline{\mathrm{X}}\underline{\mathrm{A}}(\underline{\mathrm{A}}'\underline{\mathrm{X}}'\underline{\mathrm{X}}\underline{\mathrm{A}})^{-1}\underline{\mathrm{A}}'\underline{\mathrm{X}}' = \underline{\mathrm{X}}\underline{\mathrm{A}}\underline{\mathrm{A}}^{-1}(\underline{\mathrm{X}}'\underline{\mathrm{X}})^{-1}(\underline{\mathrm{A}}')^{-1}\underline{\mathrm{A}}'\underline{\mathrm{X}}' = \boldsymbol{P_X}$, and therefore $\underline{\mathrm{e}} = (\underline{\mathrm{I}} - \boldsymbol{P_X})\underline{\mathrm{y}} = (\underline{\mathrm{I}} - \boldsymbol{P_A})\underline{\mathrm{y}} = \boldsymbol{e_A}$. In words, the space spanned by the columns of $\underline{\mathrm{X}}$ is the same as the span of $\underline{\mathrm{Z}}$, the only difference is the basis (which explains why $\underline{\mathrm{b}} \neq \boldsymbol{b_A}$ and in particular why $\boldsymbol{b_A}$ is a *rotation* of $\underline{\mathrm{b}}$).

References

Abadir, K., and J. Magnus (2005): *Matrix Algebra*, Econometric Exercises. Cambridge University Press.

Angrist, J. D., and J.-S. Pischke (2008): *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press.

Wooldridge, J. M. (2010): *Econometric Analysis of Cross Section and Panel Data*. The MIT Press, second edn.