Slides for the course

Statistics and econometrics

Part 5: Properties of the OLS-MM estimator

European University Institute

Andrea Ichino

September 29, 2014

Outline

Algebraic and geometric properties of the OLS estimators

Statistical properties of the OLS estimators

- Unbiasedness
- Consistency
- Efficiency
- The Gauss-Markov Theorem

Section 1

Algebraic and geometric properties of the OLS estimators

Properties concerning residuals

- The Sample Regression Function is the set of the *fitted values*

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i \quad (1)$$

- The estimated sample residuals $\hat{u} = y - \hat{y}$ satisfy:

$$\sum_{i=1}^{n} \hat{u}_i = 0 \quad (2)$$

$$\sum_{i=1}^{n} x_i \hat{u}_i = 0 \quad (3)$$

$$\sum_{i=1}^{n} (\hat{y}_i - \bar{y}) \hat{u}_i = 0 \quad (4)$$

- A geometric interpretation (see the figure drawn in class):

$$y = \hat{y} + \hat{u} \quad (5)$$

A decomposition of the total variation of y_i

The OLS-MM estimator decomposes the total variation of y_i into a component explained by x_i and a residual unexplained component.

$$SST = \text{Total Sum of Squares} = \sum_{i=1}^{n}(y_i - \bar{y})^2 \tag{6}$$

$$SSE = \text{Explained Sum of Squares} = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 \tag{7}$$

$$SSR = \text{Residual Sum of Squares} = \sum_{i=1}^{n}\hat{u}_i^2 \tag{8}$$

$$SST = SSE + SSR \tag{9}$$

The proof is easy, developping the square in SST and using (4).

Goodness of fit and the R-squared

Assuming variability in the sample ($SST \neq 0$), the R-Squared is defined as

$$R^2 = \frac{SSE}{SST} = 1 - \frac{SSR}{SST} \tag{10}$$

which takes values between 0 and 1.

The R-squared measures the proportion of the total variation of y that is explained by x.

It is also a measure of the goodness of fit of the model.

While a low R-squared may appear to be a “bad sign", we will show later that x may still be a significant determinant of y even if the R-squared is low.

Section 2

Statistical properties of the OLS estimators

Three desirable properties

One can think of several properties that an estimator (a “recipe") should have in order to produce satisfactory estimates (“cakes").

At this stage we focus on three of these possible properties.

Note that the estimate is a random variable, because it is a function of the sample observations which are random variables.

The desirable properties are:

1. Unbiasedness;
2. Consistency;
3. Efficiency.

Subsection 1

Unbiasedness

Are $\hat{\beta}_0$ and $\hat{\beta}_1$ unbiased for β_0 and β_1 ?

An estimator of population parameter is unbiased when its expected value is equal to the population parameter.

The crucial population parameter of interest is the slope of the PRF.

We want to prove that:

$$E(\hat{\beta}_1|\{x_i\}) \equiv E\left(\frac{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}|\{x_i\}\right) = \frac{Cov(y, x)}{V(x)} \equiv \beta \tag{11}$$

We need 4 assumptions, of which 3 have already been introduced.

Angrist and Pischke (2008) suggest that we should care more for consistency, which (as we will see) does not require the fourth assumption.

The necessary assumptions for unbiasedness

- SLR 1: In the population, y is related to x and u as:

$$y = \beta_0 + \beta_1 x + u \tag{12}$$

- SLR 2: The n observations y_i and x_i are a random sample of the population and the residual u_i is defined by:

$$y_i = \beta_0 + \beta_1 x_i + u_i \tag{13}$$

- SLR 3: The observations $\{x_1, ..., x_n\}$ are not all equal
- SLR 4: The residual u is mean-independent of x:

$$E(u|x) = 0 \tag{14}$$

Note that β_0 and β_1 in the PRF are defined by

$$E(ux) = 0 \qquad \text{and} \qquad E(u) = 0 \tag{15}$$

which, as we will see, imply consistency of OLS-MM for the PRF.

Proof of unbiasedness of the OLS estimator $\hat{\beta}_1$

Note first that SLR 3 is needed otherwise $\hat{\beta}_1$ would not exist.

It is then useful to consider the following general result which is easy to verify for any random variables z_i and w_i:

$$\sum_{i=1}^{n}(z_i - \bar{z})(w_i - \bar{w}) = \sum_{i=1}^{n} z_i(w_i - \bar{w}) = \sum_{i=1}^{n}(z_i - \bar{z})w_i \qquad (16)$$

Note that this holds also when $z_i = w_i$.

Using (16), the fact that $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$, and SLR 1 and SLR 2 to substitute for y_i , we can rewrite $\hat{\beta}_1$ as:

$$\begin{aligned} \hat{\beta}_1 &= \frac{\sum_{i=1}^{n} y_i(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n}(\beta_0 + \beta_1 x_i + u_i)(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \qquad (17) \\ &= \beta_1 + \frac{\sum_{i=1}^{n}(u_i)(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \end{aligned}$$

Proof of unbiasedness of the OLS estimator $\hat{\beta}_1$ (cont.)

Substituting (17) in (11) and defining the Total Sum of Squared deviation from the mean of x as

$$SST_x = \sum_{i=1}^{n} (x_i - \bar{x})^2 : \tag{18}$$

we obtain:

$$\begin{aligned} E(\hat{\beta}_1 | \{x_i\}) &= E\left(\beta_1 + \frac{\sum_{i=1}^{n}(u_i)(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} | \{x_i\}\right) \\ &= \beta_1 + \frac{1}{SST_x}\left(\sum_{i=1}^{n} E[u_i(x_i - \bar{x}) | \{x_i\}]\right) \\ &= \beta_1 + \frac{1}{SST_x}\left(\sum_{i=1}^{n}(x_i - \bar{x}) E(u_i | \{x_i\})\right) = \beta_1 \end{aligned} \tag{19}$$

The last equality holds because of SLR 4 and random sampling.

Proof of unbiasedness of the OLS estimator $\hat{\beta}_0$

The proof of unbiasedness of $\hat{\beta}_0$ is straightforward. Taking the sample average of (13) we get that

$$\bar{y} = \beta_0 + \beta_1 \bar{x} + \bar{u} \tag{20}$$

Then,

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = \beta_0 + (\beta_1 - \hat{\beta}_1)\bar{x} + \bar{u} \tag{21}$$

And therefore:

$$\begin{aligned} E(\hat{\beta}_0|x) &= \beta_0 + E(\beta_1 - \hat{\beta}_1)\bar{x}|x) + E(\bar{u}|x) \qquad (22) \\ &= \beta_0 \end{aligned}$$

because $E(\hat{\beta}_1|x) = E(\beta_1|x)$ and $E(\bar{u}|x) = 0$.

The special case in which the CEF is linear

If y and x are jointly normally distributed:

$$E(y|x) = \beta_0 + \beta_1 x \tag{23}$$

the CEF is linear and coincides with the PRF; in this case, by construction:

$$E(u|x) = E(y - \beta_0 - \beta_1 x|x) = E(y - E(y|x)|x) = E(y|x) - E(y|x) = 0 \tag{24}$$

and OLS-MM is necessarily unbiased for the PRF (and the CEF).

Galton's study of the intergenerational transmission of height h_j, that first used the word "Regression", made implicitly this assumption :

$$h_s = \alpha + \gamma h_f + \epsilon \tag{25}$$

It is also a standard in many traditional econometrics textbooks.

The general case in which the CEF is non-linear

Consider again the education-earnings example (see next two pages)

The PRF is defined (i.e. positioned in the plane) so that by construction

$$E(ux) = 0 \qquad \text{and} \qquad E(u) = 0$$

but inspection of the two figures clearly suggests that

$$E(u|x) \neq 0$$

When the CEF is non linear, the distance between y and the PRF (which is u) must necessarily change with x.

In this case assuming $E(u|x) = 0$ implies sweeping under the carpet the non linearity of the CEF.

An example of Conditional Expectation Function

Figure : The CEF of labor earnings given education in the US

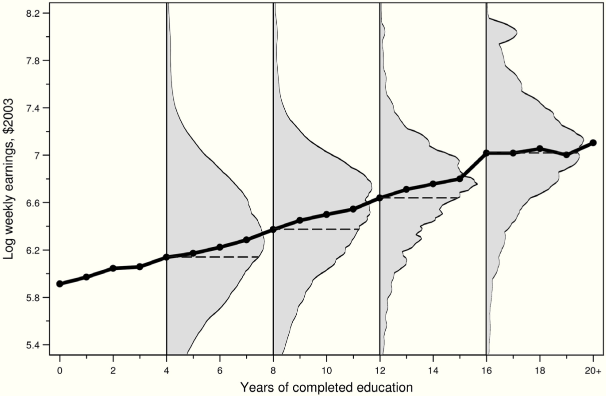

Figure 2.1.1 - CEF of Weekly earnings as a function of schooling.

The sample includes white men aged 40-49. The data are from the 1980 IPUMS 5% sample.

An example of Population Regression function

Figure : The PRF of labor earnings given education in the US

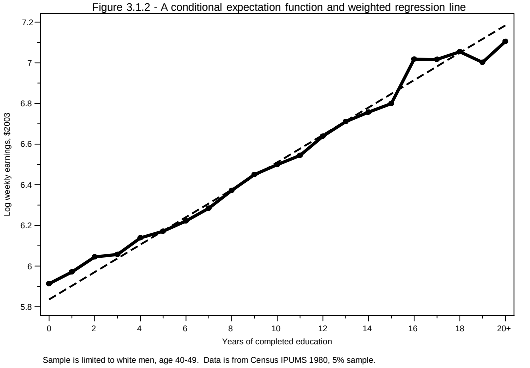

Figure 3.1.2: Regression threads the CEF of average weekly wages given schooling

An instructive case in which $E(ux) = 0$ but $E(u|x) \neq 0$

Consider a binary outcome y (college enrollment) and a regressor x (family income)

The PRF is

$$y = \beta_0 + \beta_1 x + u \tag{26}$$

and the population is described in this table:

x	y	u	ux
0	0	$0 - \beta_0$	0
0	1	$1 - \beta_0$	0
1	1	$1 - \beta_0 - \beta_1$	$1 - \beta_0 - \beta_1$
2	1	$1 - \beta_0 - 2\beta_1$	$2(1 - \beta_0 - \beta_1)$

The parameters of the PRF in this case

The parameters of the PRF are given by the solution of the two moment conditions

$$\begin{aligned} E(u) &= \frac{-\beta_0 + (1-\beta_0) + (1-\beta_0-\beta_1) + (1-\beta_0-\beta_1 2)}{4} = 0 \\ E(ux) &= \frac{1-\beta_0-\beta_1+2-2\beta_0-4\beta_1}{4} = 0 \end{aligned} \tag{27}$$

and the solutions are

$$\beta_0 = \frac{6}{11} \tag{28}$$

$$\beta_1 = \frac{3}{11} \tag{29}$$

The PRF implies that the residuals are

$$u = y - \frac{6}{11} - \frac{3}{11}x \tag{30}$$

The conditional expectation of u given x in this case

$$E(u|x=0) = 1 - 2\beta_0 = -\frac{1}{11} \tag{31}$$

$$E(u|x=1) = 1 - \beta_0 - \beta_1 = +\frac{2}{11} \tag{32}$$

$$E(u|x=2) = 1 - \beta_0 - 2\beta_1 = -\frac{1}{11} \tag{33}$$

More generally, every Limited Dependent Variable model with non binary regressors implies that:

- the CEF is non linear, but the PRF satisfies the conditions

$$E(ux) = 0 \qquad \text{and} \qquad E(u) = 0$$

- The OLS-MM estimator is biased for the PRF because

$$E(u|x) \neq 0$$

- but we now prove that is nevertheless consistent for the PRF.

Subsection 2

Consistency

Are $\hat{\beta}_0$ and $\hat{\beta}_1$ consistent for β_0 and β_1

An estimator of a population parameter is consistent when the estimates it produces can be made arbitrarily close to the population parameter by increasing the sample size.

Formally $\hat{\beta}_1$ converges in probability to β_1:

$$\lim_{n \to +\infty} Pr(|\hat{\beta}_1 - \beta_1| > \epsilon) = 0 \qquad \forall \epsilon \tag{34}$$

Equivalent notational forms to denote convergence in probability are

$$\hat{\beta}_1 \xrightarrow{p} \beta_1 \tag{35}$$

$$\operatorname*{Plim}_{n \to +\infty} \hat{\beta}_1 = \beta_1 \tag{36}$$

and similarly for $\hat{\beta}_0$.

Proof of Consistency of the OLS estimator

Using:

- the Law of Large numbers
- the Continuous Mapping Theorem for P-Convergence
- $E(UX) = E\left[x(y - \beta_0 - \beta_1 x)\right]$ which defines the PRF:

$$
\begin{aligned}
\text{plim } \hat{\beta}_1 &= \text{plim } \left(\beta_1 + \frac{\sum_{i=1}^{n}(u_i)(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \right) \\
&= \beta_1 + \frac{\text{plim } \left(\sum_{i=1}^{n}(u_i)(x_i - \bar{x})\right)}{\text{plim } \left(\sum_{i=1}^{n}(x_i - \bar{x})^2\right)} \\
&= \beta_1 + \frac{Cov(x,u)}{Var(x)} = \beta_1
\end{aligned}
\tag{37}
$$

Comment on the proof of Consistency

Note that

$$E(U|X) = 0 \qquad \Rightarrow \qquad E(UX) = 0 \tag{38}$$

but the converse is not true.

Therefore, precisely because of how we have defined the PRF,

- the OLS estimator is consistent for the PRF
- even if it may be biased for the PRF,
- and it will be biased in the likely and general case in which the CEF is non linear.

However, as we will see below in the lecture on causality:

- the fact that OLS is consistent for the PRF
- does not mean that the PRF has a causal interpretation;
- therefore OLS may be inconsistent for the causal effect X on Y.

Subsection 3

Efficiency

Are $\hat{\beta}_0$ and $\hat{\beta}_1$ efficient estimators for β_0 and β_1 ?

Remember that since the estimator is a function of random variables (the sample observations), it is itself a random variable.

We have seen that under assumptions SLR 1 - SLR 4,

$$E(\hat{\beta}_1|x) = \beta_1 \qquad \text{and} \qquad E(\hat{\beta}_0|x) = \beta_0 \tag{39}$$

We know want to find

$$V(\hat{\beta}_1|x) \qquad \text{and} \qquad V(\hat{\beta}_0|x) \tag{40}$$

The simplest context in which these variances can be computed is the one of *homoscedasticity*

A 5^{th} assumption: Homoscedasticity

SLR 5: The error u is said to be homoscedastic if it has the same variance given any value of the explanatory variable x:

$$V(u|x) = \sigma^2 \tag{41}$$

It is important to realize that SLR 5:

- is not needed to prove unbiasedness
- we introduce it just to simplify calculations, but we will later remove it because it is unlikely to hold in most applications.

What we can say at this stage is that under SLR1 - SLR5:

$$E(y|x) = \beta_0 + \beta_1 x \qquad \text{and} \qquad V(y|x) = \sigma^2 \tag{42}$$

which is the situation described in Figure 2.8 of Wooldridge.

The variance of $\hat{\beta}_1$ under homoscedasticity

Using (17) we can express the variance of $\hat{\beta}_1$ as

$$
\begin{aligned}
V(\hat{\beta}_1|x) &= V\left(\beta_1 + \frac{\sum_{i=1}^{n}(u_i)(x_i-\bar{x})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}|x\right) \text{ (β_1 is a constant)} \\
&= \left(\frac{1}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right)^2 V\left(\sum_{i=1}^{n}(u_i)(x_i-\bar{x})|x\right) \text{ (conditioning on x)} \\
&= \left(\frac{1}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right)^2 \sum_{i=1}^{n}(x_i-\bar{x})^2 V(u_i|x) \text{ (indep., random i)} \\
&= \left(\frac{1}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right)\sigma^2 \text{ (homoschedasticity)} \\
&= \frac{\sigma^2}{SST_x}
\end{aligned} \tag{43}
$$

The variance of $\hat{\beta}_1$ is smaller, the smaller is the variance of the unobserved component and the larger is the sample variance x.

How can we estimate σ^2

We have the sample SST_x, but we need an estimate of σ^2. Consider:

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_i + u_i && (44) \\ y_i &= \hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{u}_i && (45) \end{aligned}$$

$$\hat{u}_i - u_i = -(\hat{\beta}_0 - \beta_0) - (\hat{\beta}_1 - \beta_1)x_i \tag{46}$$

The estimated residual $\hat{u}_i$ is in general different from the unobservable component u_i. Taking the sample average of (46) we get:

$$\bar{u} = (\hat{\beta}_0 - \beta_0) + (\hat{\beta}_1 - \beta_1)\bar{x} \tag{47}$$

Note that the sample average of $\hat{u}_i$ is zero. Adding 47 to 46:

$$\hat{u}_i = (u_i - \bar{u}) - (\hat{\beta}_1 - \beta_1)(x_i - \bar{x}) \tag{48}$$

Since $\sigma^2 = E(u_i^2)$ it would seem natural to construct an estimator $\hat{\sigma}^2$ building around $\sum_{i=1}^{n}(\hat{u}_i^2)$.

An unbiased estimator for σ^2

Using (48):

$$
\begin{aligned}
E(\sum_{i=1}^{n} \hat{u}_i^2|X) &= E[\sum_{i=1}^{n}(u_i - \bar{u})^2|X] + E[(\hat{\beta}_1 - \beta_1)^2 \sum_{i=1}^{n}(x_i - \bar{x})^2|X] \\
&- 2E[(\hat{\beta}_1 - \beta_1) \sum_{i=1}^{n} u_i(x_i - \bar{x})|X] \\
&= (n-1)\sigma^2 + \sigma^2 - 2\sigma^2 = (n-2)\sigma^2 \qquad (49)
\end{aligned}
$$

Hence and unbiased estimator of σ^2 is:

$$\hat{\sigma} = \frac{1}{n-2} \sum_{i=1}^{n} \hat{u}_i^2 \qquad (50)$$

There are only $n-2$ degrees of freedom in the OLS residuals since

$$\sum_{i=1}^{n} \hat{u}_i = 0 \qquad \text{and} \qquad \sum_{i=1}^{n} x_i \hat{u}_i = 0 \qquad (51)$$

Steps to derive the last line in equation 49

- $E[\sum_{i=1}^{n}(u_i - \bar{u})^2] = (n-1)\sigma^2$
  where note that the RHS has $n-1$ (and not n) because otherwise the argument of the expectation would be biased (see for analogy the example of the biased estimator of the variance of a normal in part 3 of the slides).

- $E[(\hat{\beta}_1 - \beta_1)^2 \sum_{i=1}^{n}(x_i - \bar{x})^2|X] = V(\hat{\beta}_1|X)SST_X = \sigma^2$
  given equation (43).

- $E[(\hat{\beta}_1 - \beta_1)\sum_{i=1}^{n} u_i(x_i - \bar{x})|X] = E[(\hat{\beta}_1 - \beta_1)^2 SST_x|X] = V(\hat{\beta}_1|X)SST_X = \sigma^2$
  using equation (17) and again equation (43).

Asymptotic variance

Using

- the Central Limit Theorem
- The Delta Method

we can say that:

$$\sqrt{n}(\hat{\beta}_1 - \beta_1) \xrightarrow{d} \text{Normal}\left(0, \frac{\sigma^2}{Var(x)}\right) \tag{52}$$

We will come back to a proof of this result in the context of the Multiple Regression Function.

Subsection 4

The Gauss-Markov Theorem

The Gauss-Markov Theorem

Under the assumptions:

SLR 1: In the population y is a linear function of x.

SLR 2: y_i and x_i are a random sample of size n.

SLR 3: The observations $\{x_1, ..., x_n\}$ are not all equal.

SLR 4: The residual u is mean-independent of x.

SLR 5: Homoschedastic of u (needed for efficiency)

The OLS is the Best Linear Unbiased Estimators (BLUE), i.e. it has the smallest variance in the class of linear unbiased estimators for

$$y = \beta_0 + \beta_1 x + u \tag{53}$$

Proof for the more general case of multiple regression.