OLS as an accurate model

Sunday, February 12, 2012
12:30 PM

Up to this point, all the properties of OLS regression have been agnostic about the underlying DGP -- that is, they hold regardless of the underlying DGP. To recall, these properties include:

1) OLS fits a line that minimizes the sum of squared estimated errors $\hat{u}'\hat{u}$
2) $X\hat{\beta}$ provides the best linear error-minimizing approximation to $E[y|X]$

$E[y_0|X]$

But if we are willing to assume that the world is a linear model:

$$y = X\beta + u$$

Then OLS has some very attractive properties when applied to data from this truly linear DGP.

Each of these results relies on unproved *assumptions* that we make about the world; the results are, in fact, derived from a combination of these assumptions and the logical rules of mathematics.

Five of these assumptions are generally thought to be the most important, because they are the minimal set of assumptions from which the best-known results flow.

The Classical Linear Regression Model (CLRM)

Sunday, February 12, 2012
12:34 PM

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + u$$

1) $y = X\beta + u$

- Correct specification of X
- linear specification
- constant β values

2) $E|u_i| = 0$

$E[\hat{u}] = 0$ no matter the DGP!

These properties are properties of u, not of $\hat{u}$!

i and j being any 2 obs

3) $E|u_i u_j| = cov(u_i u_j) = 0$
$E|u_i^2| = \sigma^2 = var(u_i)$

homoskedasticity – variance of the error term is constant across observations.

4) X is non-stochastic (fixed)

5) rank $X_{n \times k} = k$

no perfect collinearity among the independent variables

the columns of X (all independent variables) are linearly independent from each other.

Dummy variable trap. dv = y

Reg 1 x = 1 lm(y ~ x+z+w)

Reg 2 z = 1 x + z + w = 1

Reg 3 w = 1

x + z + w = c

1 6 0 1

1	6	0	1
1	0	0	1
1	0	0	1
0	1	0	1
0	1	0	1
0	0	1	1
0	0	1	1
0	0	1	1

Use these assumptions as tools to demonstrate certain properties of OLS.

Not all properties depend upon all the assumptions.

Some properties can be sustained even under a relaxation of some of these assumption.

OLS may have different (unknown) properties under violations of these assumptions.

β is an unbiased estimate of β

Sunday, February 12, 2012
12:31 PM

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

Theorem: β is an unbiased estimate of β; that is, $E[\beta] = \beta$.

Pf: $\hat{\beta} = (X'X)^{-1}X'y$

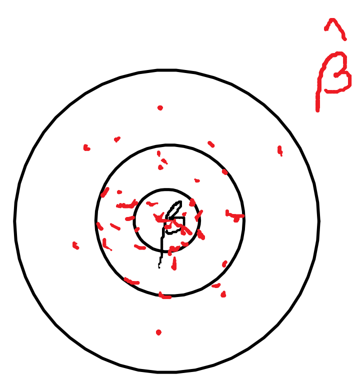
$$E[\hat{\beta}] = E[\underbrace{(X'X)^{-1}X'}_{\text{fixed}}\underbrace{y}_{\text{random}}]$$

$$= (X'X)^{-1}X'E[y] \qquad A4.$$

$$= (X'X)^{-1}X'E[\underbrace{X\beta}_{\text{fixed}} + \underbrace{u}_{\text{random}}] \qquad A1.$$

$$= \cancel{(X'X)^{-1}X'X}\beta + (X'X)^{-1}X'E[u]$$

$$E[\hat{\beta}] = \beta \qquad A2$$



What assumptions did we need for this proof? Which assumptions did we NOT need for this proof?

This property holds when u is heteroskedastic (does not have constant variance.)

Properties of expectations

$$E[Ax] = A\,E[x]$$

A : fixed

x : random

$$E[Ax] = \int Ax f(x)\,dx = A\int x f(x)\,dx.$$

where $f(x) = $ p.d.f. of x.

$x, y = $ random

$$E[x+y] = E[x] + E[y]$$

$$E[A+x] = A + E[x]$$

A fixed

x random

Relaxing assumptions: Stochastic X

Sunday, February 12, 2012
12:40 PM

Suppose that X is not fixed/non-stochastic. We can still demonstrate that $E[\hat{\beta}] = \beta$. We will, however, need to make a different assumption...

$$\text{Pf:} \quad \hat{\beta} = (X'X)^{-1}X'y$$

$$E[\hat{\beta}] = E[\underbrace{(X'X)^{-1}X'}_{\text{fixed}}\underbrace{y}_{\text{random}}]$$

$$= (X'X)^{-1}X'E[y] \qquad \text{A4.}$$

$$= (X'X)^{-1}X'E[\underbrace{X\beta}_{\text{fixed}} + \underbrace{u}_{\text{random}}] \qquad \text{A1.}$$

$$= (X'X)^{-1}X'X\beta + E[(X'X)^{-1}X'u]$$

$$E[\hat{\beta}] = \beta.$$

Xs are random.

New assumption 4: $E[(X'X)^{-1}X'u|X] = 0$.

X is not correlated with u.

$\hat{\beta} = (X'X)^{-1}X'y \equiv$ running a regression of x on y.

$$\hat{\alpha} = (X'X)^{-1}X'u \longrightarrow \hat{\alpha} = 0$$

unknown conditional mean

$$E[y|x] \qquad E[u|x] = 0$$

Can a biased model be a useful model?

Sunday, February 12, 2012
12:43 PM

N = units (countries, people, etc.)

T = time

The autodistributed lag model: frequently estimated, always biased.

$$y_t = \beta_0 + \beta_1 \underline{y_{t-1}} + \underline{u_t}$$

Let's figure out what an estimate of β, or β, will look like for a properly specified model of this type.

$$\hat{\beta}_1 = \text{regression of } \underline{M_i}\, y_t \text{ on } \underline{M_i}\, y_{t-1}$$

M_i = residual matrix from a regression on the constant term (de-meaning)

$$\hat{\beta}_1 = \left[(M_i y_{t-1})' M_i y_{t-1}\right]^{-1} (M_i y_{t-1})' M_i y_t$$

$$= \left[y_{t-1}' M_i' M_i y_{t-1}\right]^{-1} y_{t-1}' M_i' M_i y_t$$

$$= \left[y_{t-1}' M_i M_i y_{t-1}\right]^{-1} y_{t-1}' M_i M_i y_t$$

$$= \left[y_{t-1}' M_i y_{t-1}\right]^{-1} y_{t-1}' M_i \underline{y_t}$$

$$= \left[y_{t-1}' M_i y_{t-1}\right]^{-1} y_{t-1}' M_i \left[y_{t-1} \beta_1 + u_t\right] \quad *$$

$$= \underline{\cancel{(y_{t-1}' M_i y_{t-1})^{-1} y_{t-1}' M_i y_{t-1}}}\, \beta_1 +$$

$$\boxed{(y_{t-1}' M_i y_{t-1})' y_{t-1}' M_i u_t}$$

$$E[\hat{\beta}_1] = \beta_1 + E\left[(y_{t-1}' M_i y_{t-1})' y_{t-1}' M_i u_t\right]$$

① y_{t-1} is stochastic.

② clearly y_{t-1} is correlated with u_{t-1}

$$y_t = \underline{\beta_0 + \beta_1 y_{t-1} + u_t}$$

not necessarily (and in fact typically not) correlated.

$$y_t = \beta_0 + \beta_1 (\underbrace{\hat{y}_{t-1}}_{X\beta} + \underbrace{\hat{u}_{t-1}}_{\text{noise}}) + u_t$$

We can also show that ADL models are biased in R. (Show this.)

ADL models are biased, but the might be *consistent*. What does it mean for a model to be consistent?

$$\lim_{n \to \infty} E[\hat{\beta}] = \beta.$$

unbiasedness is a property of $E[\hat{\beta}]$ alone regardless of the sample size.

Consistency is a property of large (technically of infinite) samples.

$$E[\hat{\beta}] = \beta + \underbrace{(X'X)^{-1}X'u}_{\neq 0}$$

For a stochastic quantity $a(y)$ we will say that the probability limit as $n \to \infty$ is equal to a_0 if

$$\lim_{n \to \infty} \Pr\left(\| a(y) - a_0 \| < \varepsilon \right) = 1, \quad \varepsilon \sim 0.$$

the distance between some function of the random quantity y — $a(y)$ — to the point $a_0 \longrightarrow 0$ as $n \longrightarrow \infty$.

$$\operatorname*{plim}_{n \to \infty} a(y) = a_0$$

Sample mean

Ex: $\operatorname*{plim}_{n \to \infty} \bar{y} =$

true population mean.

$$\operatorname*{plim}_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} y_i = \mu_y.$$

①

$$E[\bar{y}] = E\left[\frac{1}{n}\sum_{i=1}^{n} y_i\right] = \frac{1}{n} E\left[\sum_{i=1}^{n} y_i\right]$$

$$= \frac{1}{n}\sum_{i=1}^{n} \mu_y = \frac{1}{n}\cdot n \cdot \mu_y = \mu_y .$$

(2) $$\mathrm{var}(\bar{y}) = \mathrm{var}\left[\frac{1}{n}\sum_{i=1}^{n} y_i\right] = \left(\frac{1}{n}\right)^2 \boxed{\sum_{i=1}^{n} \sigma^2}$$

$y_i = \mu_y + u_i$, $u_i \sim \text{iid var } \sigma^2$

$$\mathrm{var}(x+y) = \mathrm{var}(x) + \mathrm{var}(y) + \underbrace{2\,\mathrm{cov}(x, y)}_{=0.}$$

x, y random

$\mathrm{var}(ax) = a^2 \,\mathrm{var}\, x$ where x = random
a = constant.

$$\mathrm{var}(\bar{y}) = \left(\frac{1}{n}\right)^2 \cdot \sum_{i=1}^{n} \sigma^2 = \frac{1}{n^2}\cdot n\sigma^2$$

$$= \frac{1}{n}\sigma^2$$

$$\lim \frac{1}{n}\sigma^2 \rightarrow \frac{1}{\infty}\sigma^2 = 0.$$

(1) est on target

(2) variability in estimate diminishes to nil in large samples.

Note that probability limits (or plim) are not the same as expectations (or $E[\cdot]$).

$$E[f(y)] \neq f(E(y)) \qquad f(\cdot) \text{ any function}$$

but

$$\operatorname*{plim}_{n\to\infty} f(y) = f\left(\operatorname*{plim}_{n\to\infty} y\right).$$

We can show that ADLs are, in fact, consistent by making one assumption: $E[u_t|X_t] = 0$.

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

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

$$\operatorname*{plim}_{n\to\infty}(\beta - \hat{\beta}) = \operatorname*{plim}_{n\to\infty}\left[(X'X)^{-1}X'u\right]$$

$$\frac{1}{n}^{-1} \frac{1}{n} =$$

$$n/n = 1.$$

$$= \operatorname*{plim}_{n \to \infty} \left(\frac{1}{n} X'X\right)^{-1} \frac{1}{n} X'u$$

$$= \operatorname*{plim}_{n \to \infty} \left(\frac{1}{n} X'X\right)^{-1} \operatorname*{plim}_{n \to \infty} \left(\frac{1}{n} X'u\right)$$

$$= S_{X'X} \cdot \operatorname*{plim}_{n \to \infty} \frac{1}{n} \sum_{t=1} x_t u_t$$

it is reasonable to assume that

$$X'u = \sum_{t=1}^{n} x_t u_t = 0$$

we CAN'T say all x and all u for all t are uncorrelated.

$$\operatorname*{plim}_{n \to \infty} (\hat{\beta} - \beta) = S_{X'X} \cdot 0$$

$$= 0. \checkmark$$

As long as u_t and u_{t-1} are uncorrelated $\forall$ $t: 1 \ldots T$, then ADL model is consistent. //

Quantifying uncertainty about β

Sunday, February 12, 2012
12:50 PM

Just because $E|\beta| = \beta$ doesn't mean that any particular β is necessarily close to the true value of β. Uncertainty remains in the estimate of $\hat{\beta}$ -- we can't be sure that any particular estimate is just right. Is there any way to quantify this uncertainty? The answer is yes: we calculate $var(\beta)$.

$$var(\hat{\beta}) = \left[\hat{\beta} - E[\hat{\beta}]\right]\left[\hat{\beta} - E[\hat{\beta}]\right]'$$

$$\hat{\beta} = \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} \qquad E[\hat{\beta}] = \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}$$

$$\underset{2 \times 1}{\begin{bmatrix} \hat{\beta}_1 - \beta_1 \\ \hat{\beta}_2 - \beta_2 \end{bmatrix}} \underset{1 \times 2}{\begin{bmatrix} \hat{\beta}_1 - \beta_1 & \hat{\beta}_2 - \beta_2 \end{bmatrix}}$$

$$= \underset{2 \times 2}{\begin{bmatrix} (\hat{\beta}_1 - \beta_1)^2 & (\hat{\beta}_1 - \beta_1)(\hat{\beta}_2 - \beta_2) \\ (\hat{\beta}_2 - \beta_2)(\hat{\beta}_1 - \beta_1) & (\hat{\beta}_2 - \beta_2)^2 \end{bmatrix}}$$

$$= \begin{bmatrix} var(\hat{\beta}_1) & cov(\hat{\beta}_1, \hat{\beta}_2) \\ \dots & \dots \end{bmatrix}$$

$$= \begin{bmatrix} & \\ \operatorname{Cov}(\hat{\beta}_2, \hat{\beta}_1) & \operatorname{var}(\hat{\beta}_2) \end{bmatrix}$$

VCV – variance-covariance matrix of $\hat{\beta}$.
$k \times k$ where $k = \#$ of $\hat{\beta}$.

What is $\hat{\beta} - \beta$? $= (X'X)^{-1} X' u$

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

$$y = X\beta + u$$

$$(X'X)^{-1} X' y = \beta + (X'X)^{-1} X' u$$

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

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

$$\operatorname{var}(\hat{\beta}) = \left[\hat{\beta} - E[\hat{\beta}]\right]\left[\hat{\beta} - E[\hat{\beta}]\right]'$$

$$= \left[\hat{\beta} - \beta\right]\left[\hat{\beta} - \beta\right]'$$

via unbiasedness of OLS

$$= \left[(X'X)^{-1}X'u\right]\left[(X'X)^{-1}X'u\right]'$$

$$= (X'X)^{-1}X'uu'\left[(X'X)^{-1}X'\right]'$$

$$= (X'X)^{-1}X'uu'X\left[(X'X)^{-1}\right]'$$

$$= (X'X)^{-1}X'uu'X(X'X)^{-1}$$

$\underset{n\times 1}{u}$ $\underset{1\times n}{u'}$

$\underset{n\times n}{uu'}$

A3 $E[u_i^2] = \sigma^2$

$Var(u_i) = \sigma^2$

$E[u_i u_j] = cov(u_i, u_j) = 0$

$$\begin{bmatrix} \sigma^2 & 0 & 0 & \cdots \\ & \sigma^2 & 0 & \cdots \\ 0 & & \sigma^2 & \cdots \\ 0 & 0 & 0 & \cdots \\ \vdots & \vdots & \vdots & \end{bmatrix}$$

$$\begin{bmatrix} u_1^2 & \boxed{u_1u_2} & \boxed{u_1u_3} & \cdots \\ u_1u_2 & u_2^2 & u_2u_3 & \cdots \\ u_1u_3 & u_2u_3 & u_3^2 & \cdots \\ \vdots & & & \end{bmatrix} = uu'$$

$\sigma^2 I_{n\times n}$

$$= (X'X)^{-1}X'\sigma^2 I X (X'X)^{-1}$$

$$= \sigma^2 (X'X)^{-1}X' I X (X'X)^{-1}$$

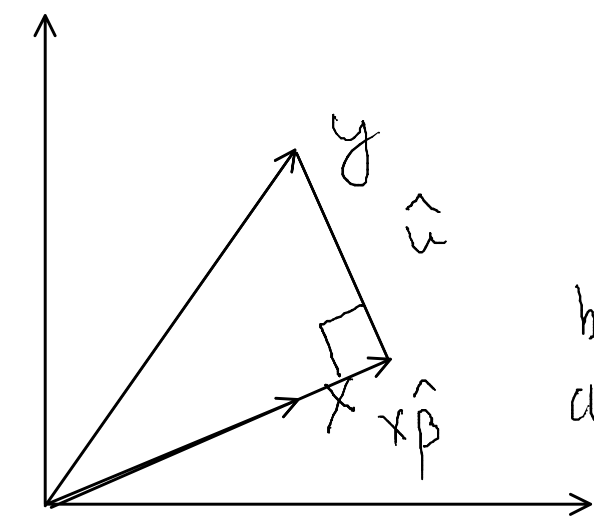
$$= \sigma^2 \cancel{(X'X)^{-1} X'X} (X'X)^{-1}$$

$$\boxed{= \sigma^2 (X'X)^{-1}}$$ *

formula for var$(\hat{\beta})$, a.k.a, the VCV matrix for your regression.

Now $\sigma^2 = E|u_i^2|$, but estimating this is trickier than you might think.

We need a way to estimate σ^2 — $E[u_i^2]$ — but tricky!



unless $\hat{\beta} = \beta$ exactly, $\|u\| > \|\hat{u}\|$ because $\hat{u}$ is chosen as a minimum.

$$\|u\| \geq \|\hat{u}\|$$

OLS underestimates u.

So, $\frac{1}{n}\hat{u}'\hat{u}$ or $var(\hat{u})$ underestimates $var(u)$.

It can be shown that

$$E\left[\frac{1}{n}\hat{u}'\hat{u}\right] = \frac{n-k}{n}\sigma^2 \quad \text{or}$$

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

where k is the rank of X.

Pf: pp. 107-110 of D&M.

Hence,

upweighting.

$$\hat{\sigma}^2 = \frac{1}{n-k}\hat{\sigma}_0^2 = \boxed{\frac{1}{n-k}}\,\underline{\underline{\hat{u}'\hat{u}}}$$

$$var(y) = \boxed{\frac{1}{n}}\,y'y \longrightarrow \boxed{\frac{1}{n-k}}\,\hat{u}'\hat{u}$$

larger

This matrix is called the variance-covariance matrix, or VCV matrix, of β. Let's construct a VCV matrix in R.

$$VCV(\hat{\beta}) = (X'X)^{-1}\left[\frac{1}{n-k}\hat{u}'\hat{u}\right]$$

Properties of the VCV

Sunday, February 12, 2012
1:15 PM

It turns out that estimates of the VCV that come out of OLS are the "most efficient" estimates possible with a linear model--that is, they are the smallest possible accurate estimates of $var(\hat{\beta})$. This doesn't mean that they're the smallest possible estimates...

$var(\hat{\beta}) = \Sigma$ → the 95% CI for the $\hat{\beta}$ that came out of this estimate would not cover the true β 95% of the time.

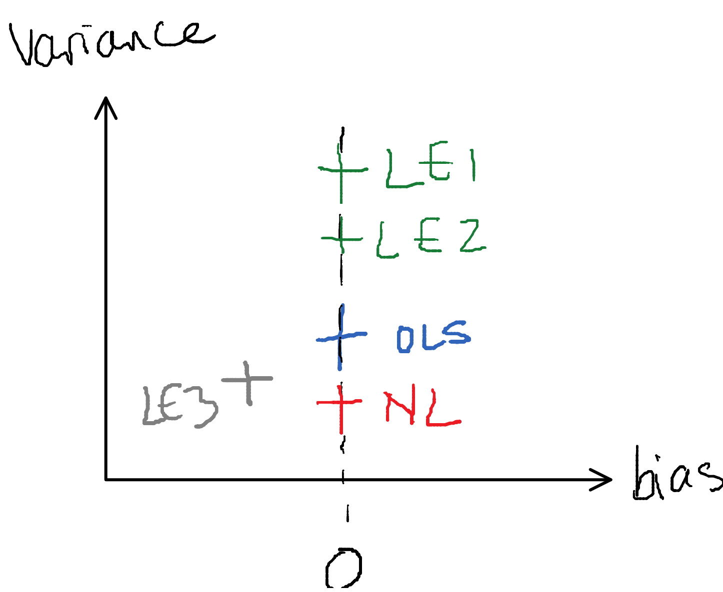
Theorem 3.1 (in Davidson and MacKinnon) -- the Gauss-Markov Theorem: If $E[u|X] = 0$ and $E[uu'] = \sigma^2 I$ and $y = X\beta + u$, then the OLS estimator $\hat{\beta}$ is more efficient than any other linear unbiased estimator.

Pf: see book.

$\boxed{A1, A2, A3}$

What Gauss-Markov says is that OLS is the Best Linear Unbiased Estimator (BLUE) under the CLRM assumptions -- IF those assumptions are correct. **Other estimators might be more efficient if they are (a) non-linear, or (b) biased.**

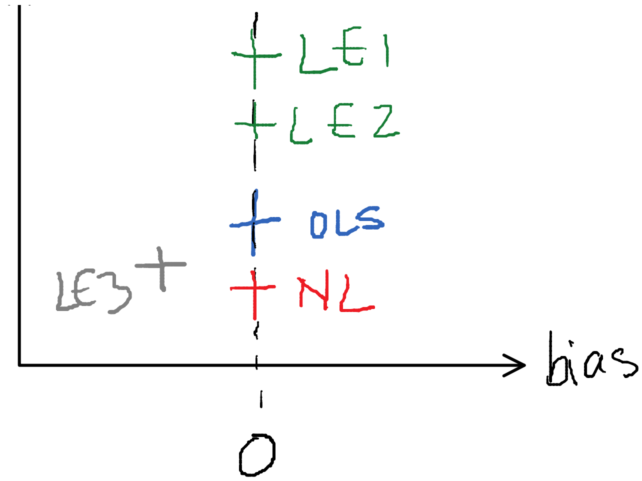
bias-variance tradeoff



$\left.\begin{array}{l} LE1 \\ LE2 \end{array}\right\}$ unbiased.

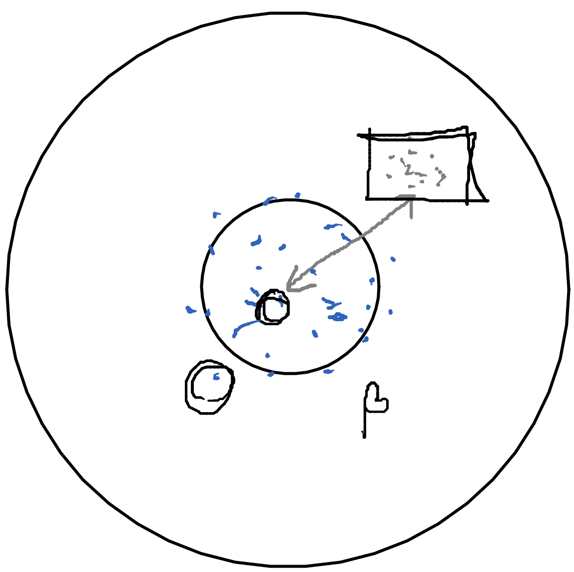
Nonlinear

LE: biased but lower variance



Nonlinear

LE : biased but lower variance



$\hat{\beta}$ = OLS estimates

$\hat{\beta}$ = Linear biased estimates