OLS Anatomy. Biases and imprecisions. GLS.

Walter Sosa-Escudero

January 26, 2015

- $\hat{\beta} = (X'X)^{-1}X'Y$
- Good: unbiased, small variance
- This lecture: what makes OLS a) imprecise, b) biased.

But first we need a key result...

The Frisch-Waugh-Lovell Theorem

Two important matrices: P and M

$$Y = X\beta + u$$
$$\hat{\beta} = (X'X)^{-1}X'Y$$
$$\hat{Y} = X\hat{\beta} = X(X'X)^{-1}X'Y = PY$$
$$e = Y - \hat{Y} = Y - PY = (I - P)Y = MY$$

- $P \equiv X(X'X)^{-1}X'$.
- $M \equiv I - X(X'X)^{-1}X'$.

Properties

- M and P are *symmetric:* $M = M', P = P'$
- M and P are *idempotent*: $M'M = M$, $P'P = P$
- $M + P = I, \;\; MP = 0.$
- $PX = X, \;\; MX = 0.$
- $e = MY = M(X\beta + u) = Mu$

M 'makes' residuals out of projecting Y on X.

Prove all these results. Easy

The Frisch-Waugh-Lovell Theorem

Consider the linear model: $Y = X\beta + u$

And partition it as follows: $Y = X_1\beta_1 + X_2\beta_2 + u$

X_1, X_2 matrices of k_1 and k_2 explanatory variables. Then, $X = [X_1 \;\; X_2]$, $\beta' = (\beta_1' \;\; \beta_2')'$ and $k = k_1 + k_2$.

$M_2 \equiv I - X_2(X_2'X_2)^{-1}X_2'$, 'makes' residuals of regressing on X_2.

$Y^* \equiv M_2 Y$, $X_1^* \equiv M_2 X_1$, respectively, OLS residuals of regressing Y on X_2, and all columns of X_1 on X_2.

Suppose that we are interested in estimating β_2, and consider the following alternative methods:

- *Method 1*: Proceed as usual and regress Y on X obtaining the OLS estimator $\hat\beta = (\hat\beta_1' \;\; \hat\beta_2')' = (X'X)^{-1}X'Y$. $\hat\beta_1$ would be the desired estimate.
- *Method 2*: Regress Y^* on X_1^* and obtain as estimate $\tilde\beta_1 = (X_1^{*\prime}X_1^*)^{-1}X_1^{*\prime}\, Y^*$

Let e_1 and e_2 be the residuals vectors of the regressions in Method 1 and 2, respectively.

Theorem (Frisch and Waugh, 1933, Lovell, 1963): $\hat\beta_1 = \tilde\beta_1$ *(first part) and* $e_1 = e_2$ *(second part)*.

Proof: Davidson and MacKinnon (2002)

Intuition

- Extremely powerful idea.
- There are two ways to estimate β_1. One is direct, regressing Y in X_1 and X_2. The other, first 'eliminates' the effect of X_2.
- Gives content to the idea of 'controlling for X_2'.
- Every k-variable regression boils down to a two variable regression!
- Has innumerable applications. See Davidson and MacKinnon (1993).

Sources of imprecisions

Result:

$$V(\hat{\beta}_j) = \frac{\sigma^2}{n\,(1 - R_j^2)V(X_j)},$$

where R_j^2 is the R^2 of regressing X_j on all other explanatory variables, and $V(X_j) = n^{-1}\ \sum_{i=1}^{n}(X_{ji} - \bar{X}_j)^2$

Proof: By the FWL theorem,

$$\hat{\beta}_j = \frac{\sum_{i=1}^n X_{ji}^* Y_i}{\sum_{i=1}^n X_{ji}^{*2}}$$

and

$$V(\hat{\beta}_j) = \frac{\sigma^2}{\sum_{i=1}^n X_{ji}^{*2}} = \frac{\sigma^2}{\frac{\sum_{i=1}^n X_{ji}^{*2}}{S_{jj}} S_{jj}}$$

where $X_j^* \equiv M_{-j} X_j$ and M_{-j} is a matrix that gets residuals of regression X_j on all other explanatory variables in the model. The result follows by noting

$$R_j^2 = 1 - \frac{\sum_{i=1}^n X_{ji}^{*2}}{S_{jj}} = 1 - \frac{\sum_{i=1}^n X_{ji}^{*2}}{\sum_{i=1}^n (X_{ji} - \bar{X}_j)^2} \quad \square$$

$$V(\hat{\beta}_j) = \frac{\sigma^2}{n\,(1-R_j^2)V(X_j)},$$

This is a crucial result. There are four factors that contribute to the variance of the OLS estimate:

1. σ^2: ignorance.
2. $V(X_j)$: variability of X_j
3. R_j^2: multicollinearity.
4. n: 'micronumerosity'.

You should tattoo this in your arm. Leave space for one more.

Sources of biases

Back to the FWLT scenario

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

- Suppose we are interested in estimating β_1. We should regress Y on X_1 and X_2 (or use the FWLT trick).
- Suppose that, instead, we regress Y on just X_1 and get

$$\hat{\beta}_1^* = (X_1'X_1)^{-1}X_1'Y$$

Result (Omitted variables bias):

$$E(\hat{\beta}_1^*|X_1) = \beta_1 + \underbrace{(X_1'X_1)^{-1}X_1'\ E(X_2|X_1)\ \beta_2}_{\text{bias}}$$

- An extremely powerful result.
- If $\beta_2 \neq 0$, OLS omission of X_2 is not necessarily biased.
- It is biased if $E(X_2|X_1) \neq 0$.
- Can omit: irrelevant variables ($\beta_2 = 0$) or relevant variables but unrelated to the variables of interest.

This is the other one you should have it tattoed in your arm.

Proof:

$$\begin{aligned}
\hat{\beta}_1^* &= (X_1'X_1)^{-1}X_1'Y \\
&= (X_1'X_1)^{-1}X_1'(X_1\beta_1 + X_2\beta_2 + u) \\
E(\hat{\beta}_1^*|X_1) &= \beta_1 + (X_1'X_1)^{-1}X_1' \; E(X_2|X_1) \; \beta_2
\end{aligned}$$

Key idea: whatever is left out of the model is left in the error term.

Small and large models

Suppose we want to estimate β_1

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

What should we do with X_2?

- $\beta_2 = 0$. Omit. Why? Gauss Markov!
- $\beta_2 \neq 0$
  - $E(X_2|X_1) = 0$. Can omit X_2.
  - $E(X_2|X_1) \neq 0$. Include X_2 to avoid bias.

So, in doubt, should be include X_2 or not? What do you think?

Omitted Variable Bias: an example

Computer generated data, but based on Appleton, French and Vanderpump ("Ignoring a Covariate: an Example of Simpon's Paradox", The American Statistician, 50, 4, 1996)

- Y = risk of death.
- $SMOKE$ = consumption of cigarrettes.

```
. reg y smoke

      Source |       SS       df       MS              Number of obs =     100
-------------+------------------------------           F(  1,    98) =  194.34
       Model |  7613.25147     1  7613.25147           Prob > F      =  0.0000
    Residual |  3839.18734    98  39.1753811           R-squared     =  0.6648
-------------+------------------------------           Adj R-squared =  0.6614
       Total |  11452.4388    99    115.6812           Root MSE      =   6.259

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       smoke |  -1.819348   .1305081   -13.94   0.000    -2.078337   -1.560359
       _cons |   158.5975   4.774249    33.22   0.000     149.1231    168.0718
------------------------------------------------------------------------------
```

```
. reg y smoke age

      Source |       SS       df       MS              Number of obs =     100
-------------+------------------------------           F(  2,    97) = 5424.58
       Model |  11350.9524     2  5675.47622           Prob > F      =  0.0000
    Residual |  101.486373    97  1.04625126           R-squared     =  0.9911
-------------+------------------------------           Adj R-squared =  0.9910
       Total |  11452.4388    99    115.6812           Root MSE      =  1.0229

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       smoke |   .9431267    .050902    18.53   0.000     .8421004    1.044153
         age |   .9804631   .0164039    59.77   0.000     .9479059     1.01302
       _cons |   12.84084   2.560392     5.02   0.000     7.759169    17.92251
------------------------------------------------------------------------------

. cor y smoke age
(obs=100)

             |        y    smoke      age
-------------+---------------------------
           y |   1.0000
       smoke |  -0.8153   1.0000
         age |   0.9797  -0.9080   1.0000
```

Generalized Least Squares

The Classical Linear Model:

1. Linearity: $Y = X\beta + u$.
2. Strict exogeneity: $E(u|X) = 0$
3. No Multicollinearity: $\rho(X) = K$, w.p.1.
4. No heteroskedasticity/ serial correlation: $V(u|X) = \sigma^2 I_n$.

Gauss/Markov: $\hat{\beta} = (X'X)^{-1}X'Y$ is best linear unbiased.

Variance: $S^2(X'X)^{-1}$ is unbiased for $V(\hat{\beta}|X) = \sigma^2(X'X)^{-1}$

Valid Inference: with the normality assumption, we use t and F tests.

Now we will focus on the consequences of relaxing $V(u|X) = \sigma^2 I_n$.

The Generalized Linear Model

Suppose all classical assumptions hold, but now

- $V(u|X) = \sigma^2\Omega$ where Ω is any symmetric, positive definite $n \times n$ matrix

Allows for heteroskedasticity (diagonal terms of Ω not all equal) and/or serial correlation (off-diagonal elements may now be $\neq 0$.).

Plan

1. Explore consequences relaxing $V(u|X) = \sigma^2 I_n$.
2. Find optimal estimators and valid inference strategies.

Consequences of relaxing $V(u|X) = \sigma^2 I_n$

- $\hat{\beta}$ still linear and unbiased (why?) but the Gauss Markov Theorem does not hold anymore. We will show constructively that $\hat{\beta}$ is now inneficient by finding the BLUE for the generalized linear model.
- $V(\hat{\beta}|X)$ will now be $\sigma^2(X'X)^{-1}\Omega(X'X)^{-1}$ (check it).
- $V(\hat{\beta}|X)$ is no longer $\sigma^2(X'X)^{-1}$, and $S^2(X'X)^{-1}$ will be a biased estimator for $V(\hat{\beta})$.
- t tests no longer have the t distribution, F tests no longer valid too.

So, ignoring heteroskedasticity or serial correlation, that is, the use of $\hat{\beta}$ and $\hat{V}(\hat{\beta}|X) = S^2(X'X)^{-1}$, keeps estimation of β unbiased though inefficient, and invalidates all standard inference.

Generalized Least Squares

A simple result: if Ω is $n \times n$ symmetric and pd, there is an $n \times n$ nonsingular matrix C such that

$$\Omega^{-1} = C'C$$

What does this mean, intuitively?

Consider now the following tranformed model

$$Y^* = X^*\beta + u^*$$

where $Y^* = CY$, $X^* = CX$ and $u^* = Cu$.

Now check:

1. $Y^* = X^*\beta + u^*$, so the transformed model is trivially linear.
2. $E(u^*|X) = CE(u|X) = 0$
3. $\rho(X^*) = \rho(CX) = K$, w.p.1. (CX is a rank preserving tranformation of X!).
4. $V(u^*|X) =$

$$\begin{aligned} V(Cu|X) = E(CuuC'|X) &= CE(uu'|X)C' \\ &= C\sigma^2\Omega C' \\ &= \sigma^2 C[\Omega^{-1}]^{-1}C' \\ &= \sigma^2 C\left[(C'C)^{-1}\right]^{-1} C \\ &= \sigma^2 I_n \end{aligned}$$

So...

All classical assumption hold *for the transformed model*, hence the BLUE is:

$$\hat{\beta}_{gls} = (X^{*\prime}X^*)^{-1}X^{*\prime}Y^*$$

This the generalized least squares estimator.

- GLS is the OLS estimator of the transformed model.
- Provides the BLUE under heteroskedasticity / serial correlation.
- Now it is clear that $\hat{\beta}$ is inefficient in the generalized context (why?)
- Important: statistical properties depend on the underlying structure (they are not properties of an estimator per-se).

Note that

$$
\begin{aligned}
\hat{\beta}_{gls} &= (X^{**}X^{*})^{-1}X^{**}Y^{*} \\
&= (X'C'CX)^{-1}X'C'CY \\
&= (X'\Omega^{-1}X)X'\Omega^{-1}Y
\end{aligned}
$$

- When $\Omega = I_n$, $\hat{\beta}_{gls} = \hat{\beta}$.
- The practical implementation of $\hat{\beta}_{gls}$ requires that we know Ω (though not σ^2.)
- It is easy to check that $V(\hat{\beta}_{gls}) = \sigma^2(X^{*\prime}X^{*})^{-1}$.

Feasible GLS

Suppose there is an estimate for Ω, label it $\hat{\Omega}$. Then, replacing Ω by $\hat{\Omega}$:

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

This is the feasible GLS estimator.

Is it linear and unbiased? Efficient?