Robust Standard Errors

Charlie Gibbons
Political Science 239

March 11, 2009

Outline

Why robust standard errors?

- Heteroskedasticity
- Interpersonal/intragroup correlations
- Serial correlation in difference-in-difference approaches

Why robust standard errors?

Do these issues cause bias in our estimates?

Why robust standard errors?

Do these issues cause bias in our estimates?

No, they only bias our standard errors.

Why robust standard errors?

Do these issues cause bias in our estimates?

No, they only bias our standard errors.

What do we need to be true about the errors for unbiased estimates?

Why robust standard errors?

Do these issues cause bias in our estimates?

No, they only bias our standard errors.

What do we need to be true about the errors for unbiased estimates?

Conditional mean independence: $\mathbb{E}[\epsilon|X] = 0$

OLS derivations

Derivation of $\hat{\beta}$:

$$\begin{aligned} &\min_b (Y - Xb)^2 \\ \Rightarrow 2X'(Y - Xb) &= 0 \\ X'Xb &= X'Y \\ b &= \left(X'X\right)^{-1} X'Y \end{aligned}$$

OLS derivations

Derivation of bias:

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

OLS derivations

Derivation of variance:

$$
\begin{aligned}
Var(\hat{\beta}) =& \mathbb{E}\left[\left(\hat{\beta} - \bar{\beta}\right)^2\right] \\
=& \mathbb{E}\left[\left(\hat{\beta} - \beta\right)^2\right] \\
=& \mathbb{E}\left[\left(\left(X'X\right)^{-1} X'\epsilon\right)^2\right] \\
=& \mathbb{E}\left[\left(X'X\right)^{-1} X'\epsilon\epsilon' X \left(X'X\right)^{-1}\right] \\
=& \left(X'X\right)^{-1} X'\Omega X \left(X'X\right)^{-1}
\end{aligned}
$$

This gives a *sandwich estimator*; $\Omega \equiv \mathbb{E}(\epsilon\epsilon')$ is the meat and $\left(X'X\right)^{-1} X'$ is the bread.

Homoskedastic variance

The variance estimator under homoskedasticity assumes:

$$\begin{aligned}
\text{Var}\left(\epsilon_i\right) =& \sigma^2 \quad \forall i \\
\Omega =& \sigma^2 I \\
\text{Var}(\hat{\beta}) =& \left(X'X\right)^{-1} X'\sigma^2 I X \left(X'X\right)^{-1} \\
=& \sigma \left(X'X\right)^{-1}
\end{aligned}$$

We estimate $\hat{\sigma} = \frac{e'e}{N-K}$, giving our standard variance estimator.

Robust standard errors

To correct for heteroskedasticity, White expanded upon the work of Eicker and proposed that we set

$$\Omega = \mathrm{Diag}\left(e_i^2\right).$$

This gives a consistent estimate for the variance of $\hat{\beta}$, even though e_i^2 is not a consistent estimate of σ_i. Note that this still assumes that there is no interpersonal or serial correlations.

MacKinnon and White suggest "correcting" the variance given here by multiplying by $\frac{N}{N-K}$.

Summary

Under the no interobservational correlations, the standard errors under homoskedasticity and heteroskedasticity are:

$$\left[\mathrm{Diag}\left(\frac{e'e}{N-K}\left(X'X\right)^{-1}\right)\right]^{\frac{1}{2}}$$

$$\left[\mathrm{Diag}\left(\alpha\left(X'X\right)^{-1}X'\left(ee'\right)X\left(X'X\right)^{-1}\right)\right]^{\frac{1}{2}}$$

Why clustering?

Clustering attempts to eliminate intragroup correlations.
Imagine a simple fixed effects model with individuals i and groups $g \in G$:

$$Y_i = \alpha + \beta D_i + \gamma_g + \epsilon_i$$

We believe that the error can be decomposed into:

$$\epsilon_i = \nu_g + \eta_i.$$

Why clustering?

What's the intragroup correlation?

$$\begin{aligned}\mathbb{E}\left[(Y_i-\mu_i)(Y_j-\mu_j)\right] =& \mathbb{E}\left[\epsilon_i\epsilon_j\right]\\ =& \mathbb{E}\left[(\nu_g+\eta_i)(\nu_g+\eta_j)\right]\\ =& \mathbb{E}\left[\nu_g^2\right]=\sigma_\nu^2\end{aligned}$$

This gives the *intraclass correlation coefficient* $\rho = \frac{\sigma_\nu^2}{\sigma_\nu^2+\sigma_\eta^2}$

Since the ν_g are independently distributed, there is no intergroup correlation.

Why clustering?

How big is the bias when you don't cluster? Approximately

$$\frac{\mathrm{Var}_s(\hat{\beta})}{\mathrm{Var}_s(\hat{\beta})} \approx 1 + (\bar{n} - 1)\rho,$$

where $\bar{n}$ is the average group size and a fixed number of groups. The square root of this figure is called the *Moulton factor*.

Note that this bias is worse when group sizes are bigger (holding the number of groups fixed), the intraclass correlation is higher, and there is higher intraclass correlation among the regressors.

Solutions

- Parametric correction using the Moulton factor.
- Cluster-robust standard errors

Solutions

Suppose you want to model:

$$Y_{ig} = \alpha + \beta x_g + \delta x_{ig} + \epsilon_{ig}.$$

First regress:

$$Y_{ig} = \gamma_g + \delta x_{ig} + \eta_{ig}.$$

Then regress:

$$\hat{\gamma}_g = \alpha + \beta x_g + (\nu_g + (\gamma_g - \hat{\gamma}_g))$$

using (weighted) least squares.

Solutions

Block bootstrap: Draw groups, not individuals, randomly with replacement.

Solutions

Notice that, for all these methods, asymptotics kick in only as the number of groups, not individuals, gets large.

Hence, if there are a small number of groups, you may want to base your inference on a t-distribution with $G - K$ degrees of freedom.

Serial correlation

Serial correlation is about intergroup correlations—*i.e.*, group effects are correlated. Imagine the following model:

$$
\begin{aligned}
Y_{ist} =& \gamma_s + \lambda_t + \beta D_{st} + \epsilon_{ist} \\
Y_{ist} =& \gamma_s + \lambda_t + \beta D_{st} + \nu_{st} + \eta_{ist}
\end{aligned}
$$

Serial correlation

Imagine that we are doing a difference-in-difference approach with two time periods. Then we have:

$$Y_{is1} - Y_{is0} = (\lambda_1 - \lambda_0) + \beta(D_{s1} - D_{s0}) + (\nu_{s1} - \nu_{s0}) + (\eta_{is1} - \eta_{is0})$$

The last error term is average over all individuals, but the second-to-last grouping is over only the two periods. OLS treats this as being 0, but in a small sample (the number of states s), this may not be the case.

The asymptotics kick in as the number of state-years increase.

Note that this creates *inconsistency*, not just inefficiency.