Econometrics II

Nonstandard Standard Error Issues: A Guide for the Practitioner

Måns Söderbom[*]

10 May 2011

[*] Department of Economics, University of Gothenburg. Email: mans.soderbom@economics.gu.se. Web: www.economics.gu.se/soderbom, www.soderbom.net.

1. Introduction

In this lecture I will discuss different methods for computing **standard errors**.

I will focus on practical issues that we need to keep in mind when doing applied research. I will highlight the fact that robust standard errors, despite their popularity, are biased and can be quite misleading in small samples. I will also discuss 'clustered' standard errors.

The lecture is based on Angrist & Pischke, Chapter 8. As you can see if you read this chapter, the material gets quite technical at times - especially when the authors set out to derive the bias of robust standard errors analytically in Section 8.1. Read the technical details if you are interested! A good understanding of the theoretical mechanisms will probably help you grasp the general issues.

However, my view is that you can appreciate the problems relevant for applied economists without penetrating all the maths. I will adopt this principle in this lecture and skip most of the technical details.

2. The Bias of Robust Standard Error Estimates

2.1. Background

In the old days (early 1990s, and earlier), it was very common for empirical researchers to based inference on conventional standard errors, e.g. those obtained from the conventional variance formula $\sigma^2 \left(X'X\right)^{-1}$.

As you know, the conventional formula for standard errors will not be appropriate under heteroskedasticity; it will also be inappropriate if observations are not independent. For these and other related reasons, alternative ways of computing standard errors, that are supposedly 'robust' to such problems, have become more popular.

Today, I would say it is standard to show robust standard errors in applied work.

However, robust standard errors are not free from problems and it's important to understand how these methods work and especially the situations in which they won't work well.

I will now try to shed some light on this particular issue.

2.2. Important equations

Recall the definition of the OLS estimator:

$$\hat{\beta} = \left[\sum_i X_i X_i'\right]^{-1} \sum_i X_i Y_i,$$

or, in matrix notation,

$$\hat{\beta} = \left[X'X\right]^{-1} X'y,$$

where X is the $N \times K$ matrix with rows X_i'; K is the number of regressors; and y is the $N \times 1$ vector of Y_i's.

The vector $\hat{\beta}$ has an asymptotic normal distribution (see Section 3.1.1):

$$\sqrt{N}\left(\hat{\beta} - \beta\right) \sim N\left(0, \Omega\right),$$

where Ω is the asymptotic covariance matrix, and β is defined in terms of population moments as

$$\beta = E\left[X_i X_i'\right]^{-1} E\left[X_i Y_i\right].$$

Define the robust asymptotic covariance matrix as

$$\Omega_r = E\left[X_i X_i'\right]^{-1} E\left(X_i X_i' e_i^2\right) E\left[X_i X_i'\right]^{-1}. \tag{2.1}$$

Recall that, if residuals are homoskedastic (constant variance), we have $E\left(e_i^2\right) = \sigma^2$, and we obtain the conventional variance matrix

$$\Omega_c = \sigma^2 E\left[X_i X_i'\right]^{-1}. \tag{2.2}$$

Note that (2.1) and (2.2) are defined in terms of population moments. As usual, when we are working with a finite sample, we compute the variance by replacing population moments by sample moments:

$$\hat{\Omega}_r = N\left[X'X\right]^{-1}\left(\sum \frac{X_i X_i' \hat{e}_i^2}{N}\right)\left[X'X\right]^{-1}, \tag{2.3}$$

$$\hat{\Omega}_c = \left[X'X\right]^{-1}\hat{\sigma}^2 = \left[X'X\right]^{-1-1}\left(\sum \frac{\hat{e}_i^2}{N}\right), \tag{2.4}$$

where $\hat{e}_i = Y_i - X_i\hat{\beta}$ is the estimated regression residual.

Asymptotically, $N\hat{\Omega}_c$ converges in probability to Ω_c and $N\hat{\Omega}_r$ converges to Ω_r. However, in finite samples, the variance estimators (2.3) and (2.4) will be **biased**.

Moreover, if the residuals are homoskedastic, the robust estimator is **more biased** than the conventional estimator!

Hence, robust standard errors can be more misleading than conventional standard errors.

To illustrate this, consider the results from a Monte Carlo experiment, shown in Table 8.1.1 in AP (page 304).

Summary of the experiment:

- The model is

  $$Y_i = \beta_0 + \beta_1 D_i + \varepsilon_i,$$

  where D_i is a dummy variable.

- The true value of the parameter on D_i is zero: $\beta_1 = 0$. Key question: *Will we get the inference right if we use variance estimators* (2.3) *and* (2.4)?

- The sample size is very small: $N = 30$.

- The sample mean of D_i is equal to 0.10.

- Residuals are drawn from the distributions

  $$\varepsilon_i \sim \left\{ \begin{array}{c} N\left(0,\sigma^2\right) \text{ if } D_i = 0 \\ N\left(0,1\right) \text{ if } D_i = 1 \end{array} \right\},$$

and there are three designs: A, $\sigma^2 = 0.5$ (lots of heteroskedasticity); B, $\sigma^2 = 0.85$ (little heteroskedasticity); C, $\sigma^2 = 1$ (no heteroskedasticity).

- The results in Table 8.1.1 were generated using 25,000 replications.

[Table 8.1.1 here]

TABLE 8.1.1
Monte Carlo results for robust standard error estimates

			Empirical 5% Rejection Rates	
	Mean	Standard Deviation	Normal	t
Parameter Estimate	(1)	(2)	(3)	(4)
A. Lots of heteroskedasticity			You will reject Ho too often!	
$\hat{\beta}_1$	−.001	.586 ①		
Standard Errors				
Conventional	.331 ②	.052	.278	.257
→ HC_0	.417	.203	.247	.231
HC_1	.447	.218	.223	.208
HC_2	.523 ③	.260	.177	.164
HC_3	.636	.321	.130	.120
max(HC_0, Conventional)	.448	.172	.188	.171
max(HC_1, Conventional)	.473	.190	.173	.157
max(HC_2, Conventional)	.542	.238	.141	.128
max(HC_3, Conventional)	.649	.305	.107	.097
B. Little heteroskedasticity			Better now	
$\hat{\beta}_1$	.004	.600 ①		
Standard Errors	②			
Conventional	.520	.070	.098	.084
→ HC_0	.441	.193	.217	.202 ← But this is not good
HC_1	.473	.207	.194	.179
HC_2	.546	.250	.156	.143
HC_3	.657	.312	.114	.104
max(HC_0, Conventional)	.562	.121	.083	.070
max(HC_1, Conventional)	.578	.138	.078	.067
max(HC_2, Conventional)	.627	.186	.067	.057
max(HC_3, Conventional)	.713	.259	.053	.045
C. No heteroskedasticity				
$\hat{\beta}_1$	−.003	.611		
Standard Errors				
Conventional	.604	.081	.061	.050
→ HC_0 too low →	.453	.190	.209	.193 ← Not good!
HC_1	.486	.203	.185	.171
HC_2	.557	.247	.150	.136
HC_3	.667	.309	.110	.100
max(HC_0, Conventional)	.629	.109	.055	.045
max(HC_1, Conventional)	.640	.122	.053	.044
max(HC_2, Conventional)	.679	.166	.047	.039
max(HC_3, Conventional)	.754	.237	.039	.031

Notes: The table reports results from a sampling experiment with 25,000 replications. Columns 1 and 2 shows the mean and standard deviation of estimated *standard errors*, except for the first row in each panel which shows the mean and standard deviation of $\hat{\beta}_1$. The model is as described by (8.1.9), with $\beta_1 = 0$, $r = .1$, $N = 30$, and heteroskedasticity as indicated in the panel headings.

3. Clustering and Serial Correlation in Panels

Using clustered standard errors can radically change your inference. Until maybe 15 years ago, you would rarely see clustered standard errors in applied work. That has all changed now - partly, I guess, because computing them has become straightforward, and partly because microeconometricians have realized the importance of the issue.

To illustrate, suppose we're interested in the bivariate regression

$$Y_{ig} = \beta_0 + \beta_1 x_g + e_{ig},$$

where Y_{ig} is the dependent variable for individual i in cluster g. Note the absence of an i-subscript on the x-variable. This is not a typo. Instead, this notation implies that the explanatory variable only varies across clusters (e.g. towns, schools, etc.); there is no variation within clusters. For example, i might denote student and g class, and Y_{ig} might be student performance and x_g class size; if there are on average 25 students per class, there will be $25 \times G$ observations in your sample.

Equipped with $25 \times G$ observations you are likely to have a large sample. This should result in low standard errors and high t-values, right? Well, maybe not.

One obvious suspicion is that the performance of students within the same class will be correlated (perhaps because they share the same teacher):

$$E\left[e_{ig}e_{jg}\right] = \rho_e \sigma_e^2 > 0, \tag{3.1}$$

where ρ_e is the intraclass correlation of the residual and σ_e^2 is the residual variance.

Note the similarity of (3.1) to the expression for the serial correlation in the POLS residual for a panel data model in which there is unobserved heterogeneity

$$corr\left(v_{it}^{OLS}, v_{i,t-s}^{OLS}\right) = \frac{\sigma_\alpha^2}{\sigma_\alpha^2 + \sigma_u^2}$$

(see eq. 3.7 in the panel data lecture).

Indeed, correlation within groups is often modelled using an additive random effects model, where

$$e_{ig} = v_g + \eta_{ig},$$

where v_g is a random component specific to class g and η_{ig} is student specific, assumed uncorrelated across students.

In this setting, failing to acknowledge the group structure in the data can lead to radically downward biased standard errors. Intuitively, the mistake we're making if ignoring the group structure is to assume the residuals are uncorrelated across all observations; while we are prepared to believe residuals are uncorrelated **across** clusters, we do not want to impose zero correlation **within** clusters.

In the special case where regressors are nonstochastic and the clusters are of equal size n, one can show that

$$\frac{V\left(\hat{\beta}_1\right)}{V_c\left(\beta_1\right)} = 1 + (n-1)\,\rho_e,$$

where $V\left(\hat{\beta}_1\right)$ is the correct sampling variance and $V_c\left(\beta_1\right)$ denotes the estimate based on the conventional variance formula (2.4). So you see, using the conventional formula gets more and more problematic as the intraclass correlation and the cluster size increase. The square root of the $V\left(\hat{\beta}_1\right)/V_c\left(\beta_1\right)$ formula is known as the Moulton factor.

So what's an applied economist to do? One solution is to estimate the regression based on group averages instead of microdata (e.g. treat each class as a datapoint, rather than each observation). This is a fairly unusual approach. By far the most common approach is to use the **clustered covariance matrix:**

$$\hat{\Omega}_{cl} = \left(X'X\right)^{-1}\left(\sum_g X_g'\hat{\Psi}_g X_g\right)\left(X'X\right)^{-1},$$

where

$$\begin{aligned}
\hat{\Psi}_g &= a\hat{e}_g\hat{e}_g' \\
&= a\begin{bmatrix}
\hat{e}_{1g}^2 & \hat{e}_{1g}\hat{e}_{2g} & ... & \hat{e}_{1g}\hat{e}_{n_g g} \\
\hat{e}_{1g}\hat{e}_{2g} & \hat{e}_{2g}^2 & & \\
... & & & \\
\hat{e}_{1g}\hat{e}_{n_g g} & ... & & \hat{e}_{n_g g}^2
\end{bmatrix}
\end{aligned}$$

and a is a degrees of freedom adjustment factor. Note that, if there is no intracluster correlation, the off-diagonal elements will be small, and so the clustered standard errors should be similar to robust standard errors.

This sounds good. However, the formula for clustered standard errors will suffer from similar finite sample bias as that discussed earlier for the robust estimator. What matters in this context is the number of clusters - *not* the sample size. That is, clustered standard errors may be very misleading if the number of clusters is small (see section 8.2.3 in AP).

For very similar reasons to those just discussed, we will probably want to use clustered standard errors when estimating panel data models - see Section 8.2.2 in AP for a nice discussion.