Fall 2018
University of California, Berkeley

Christina Romer
David Romer

# ECONOMICS 210C / ECONOMICS 236A
# MONETARY HISTORY

## A Little about GLS, Heteroskedasticity-Consistent Standard Errors, Clustering, and All That

## OCTOBER 24

## I. THE BIG PICTURE

- We spend much of the course worrying about the possibility that coefficient estimates (or other estimates of economic relationships) may be biased. But standard errors can also be biased – sometimes greatly.

- Just as there is no mechanical way to solve the problem of potential bias in point estimates, there is no mechanical way to solve the problem of potential bias in standard errors. As with obtaining reliable coefficient estimates, obtaining reliable standard errors requires a mix of good econometric technique and good judgment.

## II. INTRODUCTION AND GENERAL CONSIDERATIONS

**1.** Recall:

Consider $Y = X\beta + \varepsilon$. The OLS estimate of $\beta$ is $\hat{\beta} = (X'X)^{-1}X'Y$.

Since $Y = X\beta + \varepsilon$, it follows that $\hat{\beta} = (X'X)^{-1}X'(X\beta + \varepsilon)$, which equals $\beta + (X'X)^{-1}X'\varepsilon$. Thus, $\hat{\beta} - \beta = (X'X)^{-1}X'\varepsilon$.

Thus, the variance-covariance matrix of $\hat{\beta} - \beta$ is $\mathrm{E}[(X'X)^{-1}X'\varepsilon\varepsilon'X(X'X)^{-1}]$, or $(X'X)^{-1}X'\Omega X(X'X)^{-1}$, where $\Omega$ is the variance-covariance matrix of $\varepsilon$. The standard errors are the square roots of the diagonal elements of $(X'X)^{-1}X'\Omega X(X'X)^{-1}$.

**2.** If we know $\Omega$, the first best is to do GLS.

And if we have information about the functional form of $\Omega$, we can usually do some type of two-step procedure ("feasible GLS"). For example, if we have panel data for, say, states, and believe the residuals are likely to heteroskedastic by state, we can first do OLS, estimate the variance of the residuals by state, and then do weighted least squares.

One can make a case that GLS (and feasible GLS) is not done often enough – people are often too quick to do OLS and focus on correcting the standard errors.

**3.** If one does OLS, standard errors can be computed using $(X'X)^{-1}X'\hat{\Omega}X(X'X)^{-1}$, where $\hat{\Omega}$ is an estimate of $\Omega$.

For example, conventional OLS standard errors are computed under the assumption that $\Omega$ is the identity matrix times a constant, $\sigma_\varepsilon^2$. In this case, $(X'X)^{-1}X'\hat{\Omega}X(X'X)^{-1}$ simplifies to $(X'X)^{-1}\sigma_\varepsilon^2$.

**4.** The basic idea of corrected standard errors is to use information from the estimated residuals to construct $\hat{\Omega}$.

The "original sin" of corrected standard errors is the following. Since $\Omega \equiv \mathrm{E}[\varepsilon\varepsilon']$, it is tempting to estimate $\Omega$ as $\hat{\Omega} = \hat{\varepsilon}\hat{\varepsilon}'$, where $\hat{\varepsilon}$ is the vector of regression residuals. With this choice of a $\hat{\Omega}$, our estimated variance-covariance matrix for $\hat{\beta} - \beta$ is $(X'X)^{-1}X'\hat{\varepsilon}\hat{\varepsilon}'X(X'X)^{-1}$.

But notice that we can write this as s $(X'X)^{-1}(X'\hat{\varepsilon})(X'\hat{\varepsilon})'(X'X)^{-1}$. Since $X'\hat{\varepsilon} = 0$, this gives us standard errors of zero. Oops!

This is the fundamental reason that there is no mechanical way to address all possible sources of bias in standard errors.

**5.** Most approaches to correcting standard errors involve using the entries of $\hat{\varepsilon}\hat{\varepsilon}'$, for some entries of $\hat{\Omega}$ , but using zero for some of the off-diagonal entries. (An exception is Newey-West and similar procedures, which use downweighted values of the entries of $\hat{\varepsilon}\hat{\varepsilon}'$ for some entries of $\hat{\Omega}$.)

**6.** At an informal level, this suggests that corrected standard errors will suffer from two biases:

- They will be biased toward zero (since we know that if we used all the entries of $\hat{\varepsilon}\hat{\varepsilon}'$ to construct $\hat{\Omega}$, the standard errors would be exactly zero).
- They will be biased toward the conventional OLS standard errors (since those standard errors set all the off-diagonal elements of $\hat{\Omega}$ to zero).

Note that if the true standard errors are larger than the conventional standard errors, both biases operate in the direction of understating the standard errors.

**7.** In practice, there is a third danger with corrected standard errors: they often behave "peculiarly" and in unexpected ways.

**8.** One implication of this discussion is that standard errors can be too low: standard errors that are implausibly small are a red flag that something may be wrong. Suspiciously low standard errors should make you concerned that they may be very downward biased. (Alternatively, they could reflect a problem with the specification – that is, that what is driving the estimates is not what you think it is.)

As a concrete example, suppose you are trying to estimate the markup (the ratio of price to marginal cost), and you obtain a point estimate of 1.40 with a standard error of 0.01. Taken at face value, this is utterly overwhelming evidence against any value of the markup less than 1.35 or greater than 1.45. This should make you extremely suspicious: the markup is notoriously difficult to estimate, and previous researchers have found a wide range of values. So it is very likely that your standard errors are very biased (or that there is something about your specification that is causing what you think of as an estimate of the markup to actually be driven by something else, with the result that a value of the "markup" of 1.40 provides the best fit to the data for reason that have little to do with the true markup).

**9.** In the examples that follow, we will assume that there is only one right-hand side variable. Thus, $X$ is a vector and $\beta$ is a scalar. We will also assume that $X$ has mean zero.

## III. HUBER-WHITE STANDARD ERRORS

The simplest approach to correcting standard errors is to only address heteroskedasticity. That is, the off-diagonal entries of $\hat{\Omega}$ are set to zero, and the $\hat{\varepsilon}^2$'s are used for the diagonal entries.

With this approach (and only one right-hand side variable), $X\hat{\Omega}X$ is $\sum X_i^2 \hat{\varepsilon}_i^2$ (here $i$ indexes the observations). With conventional standard errors, it is $\hat{\sigma}_\varepsilon^2 \sum X_i^2$, where $\hat{\sigma}_\varepsilon^2$ is the average of the $\hat{\varepsilon}_i^{2\prime}$s.

Thus, the heteroskedasticity-corrected standard error is larger than the conventional standard error if the residuals are larger (in absolute value) when the *x*'s are larger in absolute value. In the opposite case, they are smaller.

(A corollary: The heteroskedasticity-corrected standard error for the <u>mean</u> of a variable is numerically identical to the conventional standard error.)

## IV. SERIAL CORRELATION

### A. Conventional (Uncorrected) Standard Errors

Suppose that $\Omega$ takes the form

$$
\begin{array}{l}
v\ u\ 0\ 0\ 0\ 0\ \ldots \\
u\ v\ u\ 0\ 0\ 0\ \ldots \\
0\ u\ v\ u\ 0\ 0\ \ldots \\
0\ 0\ u\ v\ u\ 0\ \ldots \\
\ldots
\end{array}
$$

Then $X'\Omega X$ equals $v\sum x_t^2$ plus $2u\sum x_t x_{t-1}$. Since the *x*'s are assumed to have mean zero, we can rewrite $2u\sum x_i x_{i-1}$ as $2u(T-1) * \mathrm{Cov}(x_t, x_{t-1})$, where $T$ is the number of observations. If we just compute conventional standard errors (and for simplicity, we assume that we have a perfect estimate of $v$, so that $\hat{\sigma}_\varepsilon^2 = v$), our estimate of $X'\Omega X$ is $v\sum x_t^2$.

For concreteness, assume that u is positive – that is, that the residuals are positively serially correlated. Then our estimate of $X'\Omega X$ is too low if $\mathrm{Cov}(x_t, x_{t-1})$ is positive (that is, the *x*'s are positively serially correlated) and too high if $\mathrm{Cov}(x_t, x_{t-1})$ is negative.

A corollary is that if the x's are serially uncorrelated (for example, if they are the innovations to some variable), our estimate of $X'\Omega X$ will be fine.

### B. Hansen-Hodrick Standard Errors

A corrected-standard-errors approach to addressing the possibility of correlation between consecutive values of $\varepsilon$ is to use the elements of $\hat{\varepsilon}\hat{\varepsilon}'$ for the diagonal elements of $\hat{\Omega}$ and for the immediate off-diagonal elements.

Likewise, if one is worried that there may be correlation not just between $\varepsilon_t$ and $\varepsilon_{t-1}$ but also between $\varepsilon_t$ and $\varepsilon_{t-2}$, one can also use the elements of $\hat{\varepsilon}\hat{\varepsilon}'$ for the elements of $\hat{\Omega}$ two away from the diagonal. And so on.

These are "Hansen-Hodrick" standard errors.

An example of how corrected standard errors can behave "weirdly" is that there is no guarantee that the diagonal elements of the resulting estimate of the variance-covariance matrix of $\hat{\beta} - \beta$ are positive. And if an entry isn't positive, the corresponding standard error – which is the square root of that element – isn't defined. What's more surprising is that this problem often arises in practice even with what don't appear to be unusual data.

## V. A LITTLE BIT ABOUT CLUSTERING

Suppose your observations have some type of natural structure to then. For example, you might have data on a number of countries over a number of years. Or you might have data on counties, with the counties being in different states.

Then there are sometimes reasons that there might be correlation across the residuals in a category. (For example, there might correlation among the residuals for all the countries in a given time period, or correlation among the residuals for all the counties in a given state.) In that case, one can compute clustered standard errors – that is, using the elements of $\hat{\varepsilon}\hat{\varepsilon}'$ for any of the elements of $\hat{\Omega}$ involving residuals from the same category (that is, two countries in the same year in our time-series/cross-section example, two counties in the same state in the state-county example). The non-zero values of the resulting $\hat{\Omega}$ then consist of a bunch of blocks along the diagonal.

Here's a case where you can explicitly work out why clustering can be valuable. Suppose we accidentally have $n$ copies of each observation. The conventional estimate of the variance of $\beta$ will be too low – it is computed under the assumption that we have $n$ times as many observations as we have genuinely independent ones. But you can work out that if we cluster among each of these groups of $n$, we get exactly the same estimate of the variance of $\beta$ as we would if we'd had only one copy of each observation (and had used heteroskedasticity-corrected standard errors).

Although clustering can be valuable, many researchers seem too quick to cluster. Many concerns presented as arguments for clustering are actually arguments for including fixed effects. For example, if your concern is that there might be a state-level policy that affects all counties in a state, why not include state fixed effects? Other concerns presented as arguments for clustering in fact point to other strategies for dealing with $\hat{\Omega}$. For example, if you think that the residuals of counties that are near one another may be positively correlated, this implies not just correlation among counties within in state, but also correlation of two counties that are adjacent to one another but in different states.

In addition, clustering can easily produce weird standard errors, especially if the number of clusters is low. Here's an attempt at intuition. Suppose that the estimate of $\beta$ we'd get from estimating $\beta$ only using the observations within a cluster is close to the estimate we get from the full sample. That means that $X'\varepsilon$ within than cluster is close to zero, and so the elements of $X'\varepsilon\varepsilon'X$ for the diagonal block corresponding to that cluster are all close to zero. So now we've

basically set a big chunk of $\hat{\Omega}$ to zero. The end result is that our estimate of $X'\hat{\Omega}X$ may be driven by a fairly small fraction of the sample.

## VI. FINAL COMMENT

It is a good idea to always also compute the standard errors the conventional way, so you can see how any correction you are doing is affecting them. And trying any other simple approach you can think of is also a good idea. For example, in our county-state example, you could aggregate the data up to the state level and then run the regression only using the state-level observations.