PLSC 503: "Multivariate Analysis for Political Research"

"Variances"
February 28, 2017

Introduction

Weisberg has an interesting approach to the subject of variance in regression models. Rather than talk about heteroscedasticity at length, "econometrics"-style, he has this chapter (Chapter 7) instead. It covers a range of things.

Variances Proportional to X and WLS

Weisberg starts with a standard linear regression model, but one where we relax the variance assumption slightly:

$$\text{Var}(u_i) = \sigma^2 / w_i \tag{1}$$

where w_i is a *known* quantity. We'll talk in a minute about when this might actually be a good idea; for now, just note that it means that we are now minimizing the sum of the weighted squared differences:

$$\text{RSS} = \sum_{i=1}^{N} w_i (Y_i - \mathbf{X}_i \boldsymbol{\beta}).$$

This yields the *weighted least squares* estimator, which in the multivariate context can be written as:

$$\hat{\boldsymbol{\beta}}_{WLS} = [\mathbf{X}'(\sigma^2 \boldsymbol{\Omega})^{-1} \mathbf{X}]^{-1} \mathbf{X}'(\sigma^2 \boldsymbol{\Omega})^{-1} \mathbf{Y}$$

which is more compactly written

$$\hat{\boldsymbol{\beta}}_{WLS} = [\mathbf{X}' \mathbf{W}^{-1} \mathbf{X}]^{-1} \mathbf{X}' \mathbf{W}^{-1} \mathbf{Y} \tag{2}$$

where

$$\mathbf{W} = \begin{bmatrix} \frac{\sigma^2}{w_1} & 0 & \cdots & 0 \\ 0 & \frac{\sigma^2}{w_2} & \cdots & \vdots \\ \vdots & 0 & \ddots & 0 \\ 0 & \cdots & 0 & \frac{\sigma^2}{w_N} \end{bmatrix}$$

is the product of σ^2 and

$$\mathbf{\Omega} = \begin{bmatrix} \frac{1}{w_1} & 0 & \cdots & 0 \\ 0 & \frac{1}{w_2} & \cdots & \vdots \\ \vdots & 0 & \ddots & 0 \\ 0 & \cdots & 0 & \frac{1}{w_N} \end{bmatrix}$$

Note that both "terms" ($\mathbf{X'X}$ and $\mathbf{X'Y}$) are multiplied by the inverse of the diagonal matrix of "weights" associated with each observation. This is identical to what we're doing above in the two-variable case, but extends it to cases with K variables.

Not surprisingly, the variance-covariance of $\hat{\boldsymbol{\beta}}_{WLS}$ is straightforward:

$$\begin{aligned} \text{Var}(\hat{\boldsymbol{\beta}}_{WLS}) &= \sigma^2(\mathbf{X'\Omega}^{-1}\mathbf{X})^{-1} \\ &\equiv (\mathbf{X'W}^{-1}\mathbf{X})^{-1} \end{aligned} \tag{3}$$

Weisberg also notes that the prediction variances for a "new" observation i_o depend on its weight w_o.

Practical WLS

The WLS approach is very flexible, and has some nice properties. In particular, if the nature of the nonconstant variance in your data is related to one of the $\mathbf{X}$ variables in a particular way, we can "weight" the observations accordingly. So, for example,

- If $\text{Var}(u_i) = \sigma^2 X_i^2$ (that is, the variance of u_i is proportional to X_i^2), then
  - We can divide each side of the equation by X_i:

$$\frac{Y_i}{X_i} = \beta_0 \left(\frac{1}{X_i}\right) + \beta_1 \left(\frac{X_i}{X_i}\right) + \frac{u_i}{X_i}$$

  - Now, $\text{Var}\left(\frac{u_i}{X_i}\right) = \text{E}\left(\frac{u_i}{X_i}\right)^2 = \sigma^2$, which is homoscedastic.
  - We can then simply estimate this transformed equation using OLS.
  - Once we've done that, we can "get back" to the original coefficients by multiplying them by X_i.
- A more common example is where we have aggregated data across the various is, in which the variance of u_i is inversely proportional to N_i, the number of observations on which Y_i is based:

$$\text{Var}(u_i) = \sigma^2 \frac{1}{N_i}$$

There are at least two very common circumstances in which this happens with observational social science data:

1. *Aggregating Data.*
    - Aggregate data combine measurements on subunits.
    - If the number of subunits varies, aggregation can yield values that have different levels of reliability.
    - Think about it this way: Are you more sure of Ruth Bader Ginsburg's liberalism than of Elena Kagan's? Why?
    - Or, consider a state-level measure of partisanship that is calculated by taking all the individuals in the National Election Study from that state, and calculating their average party ID score.
        - Now, if the data are a national probability sample, there will be a lot more people from Texas in there than from (say) Delaware (in fact, probably about 30 times more).
        - That means that our measurement of party identification for Texas will be based on more data, and thus will be more reliably accurate than the one for Delaware.
        - If we then use that aggregate ($N = 51$) partisanship measure as our Y variable, it is likely that we'll do a better job "explaining" Texas with our $\mathbf{X}$ variables than we do explaining Delaware.
        - In other words, the variability of our errors $\hat{u}_i$ will be a decreasing function of state population.
2. *"Pooling" Data.*
    - This simply means combining data across a number of (say) units, or time points.
    - So, we often "pool" data on different countries, or different states, or different candidates, etc.
    - We also pool data temporally, when we examine (e.g.) U.S. presidential election results across numerous election years.
    - Implicit in the act of pooling is that the data are *exchangeable*: that, conditional on the values of $\mathbf{X}_i$, the process governing $\mathbf{Y}$ is the same across units (however defined).
    - If exchangeability doesn't hold, pooling can lead to difference variances across units.
    - So, for example, consider studying UN votes by multiple countries observed over multiple years.
        - If one nation is more politically unstable than another, they may have greater variability in their votes.

- Unless our model (that is, the $\mathbf{X}$s) accounts for this variation, we can wind up being less able to predict those votes accurately for the less stable countries.

- A common place to find such problems is in *panel data*, which is why it can be so hard to deal with.

For example, consider the state averages we mentioned previously, where there are large differences in state populations. If this is the case, we want to weight the observations by $\frac{1}{\sqrt{N_i}}$, just as we weighted them by $\frac{1}{X_i}$ above. Weisberg notes that this simple formula for the variance in the presence of "aggregated" observations requires the subunits of measurement to be independent both within and across units; if they are not, things get more complex (see below).

It's straightforward to implement weights in R by using the `weights` option of the `lm` command:

```
> fit<-lm(y~x1+x2+x3+x4,weights=pop)
```

One can estimate a weighted least-squares model in Stata using `-regress-` with the `[aweight]` option:

```
. reg y x1 x2 x3 x4 [aweight=pop]
```

Stata treats `aweight`s as inversely scaling the variance of that observation; so, the error variance of the jth observation is $\frac{\sigma^2}{w_j}$.[1]

Misspecified Variance and "Robust" Standard Errors

If you've read much quantitative political science written in the past two decades or so, you've probably encountered the use of "robust standard errors." These are often attributed to an article by White (1980), though they were actually first derived by Huber (1967).

The Math

Recall that the formula for the variance-covariance of the parameters under heteroscedasticity is:

$$\begin{aligned} \text{Var}(\boldsymbol{\beta}_{\text{Het.}}) &= (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{W}^{-1}\mathbf{X})(\mathbf{X}'\mathbf{X})^{-1} \\ &= (\mathbf{X}'\mathbf{X})^{-1}\,\mathbf{Q}\,(\mathbf{X}'\mathbf{X})^{-1} \end{aligned} \tag{4}$$

[1] Note that Stata also rescales the `aweight`s to sum to the number of observations. This underscores the fact that, when it comes to weights, it is their relative – rather than their actual – magnitude that is important.

where $\mathbf{Q} = (\mathbf{X}'\mathbf{W}^{-1}\mathbf{X})$ and $\mathbf{W} = \sigma^2\mathbf{\Omega}$. Note that we can rewrite $\mathbf{Q}$ as

$$\begin{aligned} \mathbf{Q} &= \sigma^2(\mathbf{X}'\mathbf{\Omega}^{-1}\mathbf{X}) \\ &= \sum_{i=1}^{N} \sigma_i^2 \mathbf{X}_i \mathbf{X}_i' \end{aligned} \tag{5}$$

where $\mathbf{X}_i$ is the $1 \times K$ column vector of covariate values for observation i. The terms in (5) indicate that the unit-specific error variability is "scaled" by the variation in that observation's covariate values, and vice-versa.

Normally, one would have to know the values of $\mathbf{\Omega}$ (and therefore $\mathbf{W}$) to use (4). Huber's (and White's) insight was to figure out that it's only necessary to consistently estimate the "middle bit," $\mathbf{Q}$. White's idea in particular was to derive an estimate of $\mathbf{Q}$ using the OLS residuals $\hat{u}_i$, and then plug that estimate in to (4). Thus, an estimate of $\mathbf{Q}$ is:

$$\widehat{\mathbf{Q}} = \sum_{i=1}^{N} \hat{u}_i^2 \mathbf{X}_i \mathbf{X}_i' \tag{6}$$

White demonstrated that, under very general conditions, $\widehat{\mathbf{Q}}$ is a consistent estimator of $\mathbf{Q}$. That, in turn, means that a variance-covariance estimate based on $\widehat{\mathbf{Q}}$ is consistent for $\text{Var}(\boldsymbol{\beta}_{\text{Het.}})$. The specific estimator White proposed is thus:

$$\begin{aligned} \widehat{\text{Var}(\boldsymbol{\beta})}_{\text{Robust}} &= (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\widehat{\mathbf{Q}}^{-1}\mathbf{X})(\mathbf{X}'\mathbf{X})^{-1} \\ &= (\mathbf{X}'\mathbf{X})^{-1}\left[\mathbf{X}'\left(\sum_{i=1}^{N} \hat{u}_i^2 \mathbf{X}_i \mathbf{X}_i'\right)^{-1}\mathbf{X}\right](\mathbf{X}'\mathbf{X})^{-1} \end{aligned} \tag{7}$$

The variance-covariance estimator in (7) is also often known as the "sandwich" estimator, since the middle bit $\widehat{\mathbf{Q}}$ is "sandwiched" between more conventional terms for the variance/covariance of the covariates.

Intuitively, the idea is that one calculates the variances using the observation-specific, OLS-estimated residual variances as "weights." Including the variability of the covariates in "weights" also takes care of much of any heteroskedasticity that might be related to those variables.

Practical Issues

"Robust" standard errors are attractive, in large measure because they offer a heteroscedasticity-consistent alternative even when the investigator does not know the form of the heteroscedasticity. However, one never gets something for nothing:

- The fact that it is only a consistent estimator means that, unlike conventional OLS estimates, the usual t and F tests are now only asymptotically valid. That means that there are potential issues of bias in small samples (see, e.g., Chesher and Jewitt 1987 *Econometrica*).
- Also, robust variance estimates are slightly worse (in efficiency terms) than OLS estimates if the errors are truly homoscedastic (a la Kauermann and Carroll 2001 *JASA*).

At the same time, despite these limitations, it also bears noting that "robust" standard errors are significantly "better" than conventional OLS estimates if the data are heteroskedastic in any way. Moreover, the fact that they are consistent means that they become more accurate in larger sample sizes. That means that, in general, its not a bad idea to use these, especially if you have reason to suspect heteroscedasticity and/or have large amounts of data.

"Clustered" Robust Variance Estimates

One way to think of "robust" standard errors is as one end of a continuum:

- If you have a *lot* of information about the nature of the heteroscedasticity in your data, you can probably use GLS or FGLS and get the best results.
- On the other hand, if you have *no* information about the heteroscedasticity, then "robust" standard errors will give you good, consistent variance estimates.

As it happens, researchers often find themselves in between these two situations. The paradigmatic example in political science is with pooled time-series cross-sectional data (say, data collected on countries over time). In such a situation a researcher might have a strong suspicion that the error variance for (say) Italy is different from that for Norway, but will not really have much more information than that.

More generally, suppose we have "nested" (or "grouped") data, where we have a total of N units ("groups," or "clusters") $i = \{1, 2, ...N\}$ and each unit in turn has n_i observations $j = \{1, 2, ...n_i\}$; the model is then:

$$Y_{ij} = \mathbf{X}_{ij}\boldsymbol{\beta} + u_{ij} \tag{8}$$

For whatever reason, we might believe that the error variance within a particular "cluster" is the same (perhaps because that cluster is a repeated observation on the same unit, or whatever), but that the error variances across clusters are different. In that case, we can modify our estimate of the "weight" $\widehat{\mathbf{Q}}$ to account for this fact. In particular, we can use

$$\widehat{\text{Var}(\boldsymbol{\beta})}_{\text{Clustered}} = (\mathbf{X}'\mathbf{X})^{-1}\left\{\mathbf{X}'\left[\sum_{i=1}^{N}\left(\sum_{j=1}^{n_j}\hat{u}_{ij}^2\mathbf{X}_{ij}\mathbf{X}_{ij}'\right)\right]^{-1}\mathbf{X}\right\}(\mathbf{X}'\mathbf{X})^{-1}$$

That is,

- We first build an estimate of σ_j^2 from the squared residuals u_{ij}^2 *within* each cluster.
- We then aggregate *across* clusters, to come up with a final "weight" for inclusion in the sandwich estimator of the variances.

These "clustered" robust standard errors are a nice mid-way point between the standard robust VCV estimates and full-blown (F)GLS. In particular, they allow the analyst to incorporate information about the possible structure or location of the heteroscedasticity if/when it is available, but don't require a great deal of information about its form. This approach is thus increasingly popular with applied researchers, particularly those working with panel, TSCS, or multilevel/grouped data.

Transformations, Generalized Ws, etc.

Weisberg also talks about variance-smoothing transformations, and about models with a general structure for $\mathbf{W}$, including addressing the situation where $\mathbf{W}$ has non-zero off-diagonal elements. We'll discuss these a bit later, in somewhat different contexts (including a bit in the next class...).