Heteroskedasticity-Robust Inference

Consider the linear regression model with exogenous regressors,

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}, \quad \mathrm{E}(\boldsymbol{u}) = \boldsymbol{0}, \ \ \mathrm{E}(\boldsymbol{u}\boldsymbol{u}^\top) = \boldsymbol{\Omega}, \tag{1}$$

where $\boldsymbol{\Omega}$ is an $N \times N$ matrix with i^{th} diagonal element equal to $\omega_i^2 > 0$ and all the off-diagonal elements equal to 0.

Since $\boldsymbol{X}$ is assumed to be exogenous, the expectations in (1) can be treated as conditional on $\boldsymbol{X}$.

We would get the same asymptotic results if, instead of treating $\boldsymbol{X}$ as exogenous, we assumed that $\mathrm{E}(u_i \,|\, \boldsymbol{X}_i) = 0$.

The disturbances in (1) are uncorrelated and have mean 0, but their variances differ. They are said to be **heteroskedastic**.

We assume that the investigator knows nothing about the ω_i^2. In other words, the form of the heteroskedasticity is completely unknown.

Whatever the form of $\boldsymbol{\Omega}$, the covariance matrix of $\hat{\boldsymbol{\beta}}$ is

$$\begin{aligned} \mathrm{E}\big((\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^\top\big) &= (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \mathrm{E}(\boldsymbol{u}\boldsymbol{u}^\top) \boldsymbol{X} (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \\ &= (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{\Omega} \boldsymbol{X} (\boldsymbol{X}^\top \boldsymbol{X})^{-1}. \end{aligned} \tag{2}$$

This is often called a **sandwich covariance matrix**, for obvious reasons.

If we knew the ω_i^2, we could evaluate (2). In fact, we could do better and obtain efficient estimates of $\boldsymbol{\beta}$.

Observations with low variance convey more information than ones with high variance, and so the former should be given greater weight.

But it is assumed that we do not know the ω_i^2. We cannot hope to estimate them consistently without making additional assumptions, because there are N of them.

For the purposes of asymptotic theory, we wish to consider the covariance matrix of $N^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$, the limit of N times the matrix (2).

The asymptotic covariance matrix of $N^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$ is

$$\left(\lim_{N\to\infty} \frac{1}{N}\boldsymbol{X}^\top\boldsymbol{X}\right)^{-1}\left(\lim_{N\to\infty} \frac{1}{N}\boldsymbol{X}^\top\boldsymbol{\Omega}\boldsymbol{X}\right)\left(\lim_{N\to\infty} \frac{1}{N}\boldsymbol{X}^\top\boldsymbol{X}\right)^{-1}. \tag{3}$$

Under standard assumptions, the factor $(\lim N^{-1}\boldsymbol{X}^\top\boldsymbol{X})^{-1}$ tends to the positive definite matrix $\boldsymbol{S}^{-1}_{\boldsymbol{X}^\top\boldsymbol{X}}$.

To estimate $\boldsymbol{S}^{-1}_{\boldsymbol{X}^\top\boldsymbol{X}}$, we can simply use the matrix $(N^{-1}\boldsymbol{X}^\top\boldsymbol{X})^{-1}$ itself.

In a very famous paper, White (1980) showed that, under certain conditions, the middle matrix can be estimated consistently by

$$\frac{1}{N}\boldsymbol{X}^\top\hat{\boldsymbol{\Omega}}\boldsymbol{X}, \tag{4}$$

where $\hat{\boldsymbol{\Omega}}$ is an *inconsistent* estimator of $\boldsymbol{\Omega}$.

The simplest version of $\hat{\boldsymbol{\Omega}}$ is a diagonal matrix with i^{th} diagonal element equal to $\hat{u}_i^2$, the i^{th} squared OLS residual.

The matrix $\lim(N^{-1}\boldsymbol{X}^\top\boldsymbol{\Omega}\boldsymbol{X})$ is a $k \times k$ symmetric matrix. Therefore, it has exactly $(k^2 + k)/2$ distinct elements.

Since this number is independent of the sample size, the matrix can be estimated consistently. Its jl^{th} element is

$$\lim_{N\to\infty} \left(\frac{1}{N} \sum_{i=1}^{N} \omega_i^2 x_{ij} x_{il} \right). \tag{5}$$

This is estimated by the jl^{th} element of (4). For the simplest version of $\hat{\boldsymbol{\Omega}}$, the estimator is

$$\frac{1}{N} \sum_{i=1}^{N} \hat{u}_i^2 x_{ij} x_{il}. \tag{6}$$

Because $\hat{\boldsymbol{\beta}}$ is consistent for $\boldsymbol{\beta}_0$, $\hat{u}_i$ must be consistent for u_i, and $\hat{u}_i^2$ is therefore consistent for u_i^2.

However, $\hat{u}_i^2$ does not estimate ω_i^2 consistently.

Asymptotically, expression (6) is equal to

$$\frac{1}{N}\sum_{i=1}^{N} u_i^2 x_{ij} x_{il} = \frac{1}{N}\sum_{i=1}^{N} (\omega_i^2 + v_i) x_{ij} x_{il} \tag{7}$$

$$= \frac{1}{N}\sum_{i=1}^{N} \omega_i^2 x_{ij} x_{il} + \frac{1}{N}\sum_{i=1}^{N} v_i x_{ij} x_{il}, \tag{8}$$

where v_i is defined to equal u_i^2 minus its mean of ω_i^2.

Under suitable assumptions about the x_{ij} and the ω_i^2, we can apply a law of large numbers to the second term in (8).

Since $\mathrm{E}(v_i) = 0$, this term converges to 0, while the first term converges to expression (5).

Because $N^{-1}\sum_{i=1}^{N} \hat{u}_i^2 x_{ij} x_{il} \stackrel{a}{=} N^{-1}\sum_{i=1}^{N} u_i^2 x_{ij} x_{il}$, these arguments imply that (6) consistently estimates (5).

In practice, of course, we omit the factors of $1/N$. We simply use the matrix

$$\widehat{\mathrm{Var}}_{\mathrm{h}}(\hat{\boldsymbol{\beta}}) \equiv (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \hat{\boldsymbol{\Omega}} \boldsymbol{X} (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \tag{9}$$

directly to estimate the covariance matrix of $\hat{\boldsymbol{\beta}}$.

A more revealing way to write (9) is

$$\widehat{\mathrm{Var}}_{\mathrm{h}}(\hat{\boldsymbol{\beta}}) \equiv (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \left(\sum_{i=1}^{N} \hat{u}_i^2 \, \boldsymbol{X}_i^\top \boldsymbol{X}_i \right) (\boldsymbol{X}^\top \boldsymbol{X})^{-1}, \tag{10}$$

where $\boldsymbol{X}_i$ is the i^{th} row of $\boldsymbol{X}$. This makes it clear that the $N \times N$ matrix $\hat{\boldsymbol{\Omega}}$ is never used.

The sandwich estimator (10) is a **heteroskedasticity-consistent covariance matrix estimator**, or **HCCME**. It is valid for heteroskedasticity of unknown form.

By taking square roots of the diagonal elements of (10), we can obtain **heteroskedasticity-robust standard errors**.

Asymptotic Theory for OLS

With heteroskedasticity of unknown form, Theorem 4.3 needs to be replaced by

$$\mathrm{N}^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \xrightarrow{d} \mathrm{N}\left(\mathbf{0},\ \boldsymbol{S}_{\boldsymbol{X}^\top \boldsymbol{X}}^{-1} \left(\lim_{N\to\infty} \frac{1}{N} \boldsymbol{X}^\top \boldsymbol{\Omega} \boldsymbol{X} \right) \boldsymbol{S}_{\boldsymbol{X}^\top \boldsymbol{X}}^{-1} \right) \tag{11}$$

and

$$\lim_{N\to\infty} \frac{1}{N} \widehat{\mathrm{Var}}_{\mathrm{h}}(\hat{\boldsymbol{\beta}}) = \boldsymbol{S}_{\boldsymbol{X}^\top \boldsymbol{X}}^{-1} \left(\lim_{N\to\infty} \frac{1}{N} \boldsymbol{X}^\top \boldsymbol{\Omega} \boldsymbol{X} \right) \boldsymbol{S}_{\boldsymbol{X}^\top \boldsymbol{X}}^{-1}. \tag{12}$$

We conclude that $\hat{\boldsymbol{\beta}}$ is root-N consistent and asymptotically normal, with (12) providing a consistent estimator of its covariance matrix.

Of course, all the factors of N are omitted when we actually make inferences about $\hat{\boldsymbol{\beta}}$.

Alternative Forms of HCCME

The original HCCME (10) of White (1980), often called HC_0, uses squared residuals to estimate the diagonal elements $\boldsymbol{\Omega}$.

But least-squares residuals tend to be too small. Better estimators inflate the squared residuals (MacKinnon and White, 1985).

HC_1: Use $\hat{u}_i^2$ in $\hat{\boldsymbol{\Omega}}$ and then multiply the entire matrix by the scalar $N/(N-k)$, for a standard degrees-of-freedom correction.

HC_2: Use $\hat{u}_i^2/(1-h_i)$ in $\hat{\boldsymbol{\Omega}}$, where

$$h_i \equiv \boldsymbol{X}_i(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}_i^\top \tag{13}$$

is the i^{th} diagonal element of the "hat" matrix $\boldsymbol{P}_{\boldsymbol{X}}$.

Recall the result that, when $\text{Var}(u_i) = \sigma^2$ for all i, $\text{E}(\hat{u}_i^2) = \sigma^2(1-h_i)$. Therefore, the ratio of $\hat{u}_i^2$ to $1-h_i$ would have expectation σ^2 if the disturbances were homoskedastic.

HC_3: Use $\hat{u}_i^2/(1-h_i)^2$ in $\hat{\mathbf{\Omega}}$. This is a computationally efficient approximation to a **jackknife estimator**.

A jackknife estimator omits one observation at a time when obtaining estimates and fitted values.

The usual way to define a jackknife covariance matrix is

$$\widehat{\text{Var}}_{\text{JK}}(\hat{\boldsymbol{\beta}}) = \frac{N-1}{N}\sum_{i=1}^{N}\left(\hat{\boldsymbol{\beta}}^{(i)}-\hat{\boldsymbol{\beta}}\right)\left(\hat{\boldsymbol{\beta}}^{(i)}-\hat{\boldsymbol{\beta}}\right)^{\top}, \tag{14}$$

where $\hat{\boldsymbol{\beta}}^{(i)}$ is the vector of estimates when the i^{th} observation is omitted. We can also omit groups of observations.

It is not at all obvious that (14) is numerically equal to

$$\frac{N-1}{N}(\boldsymbol{X}^{\top}\boldsymbol{X})^{-1}\left(\sum_{i=1}^{N}\frac{\hat{u}_i^2}{(1-h_i)^2}\boldsymbol{X}_i^{\top}\boldsymbol{X}_i\right)(\boldsymbol{X}^{\top}\boldsymbol{X})^{-1}, \tag{15}$$

but it can be verified numerically.

MacKinnon and White (1985) actually proposed a slightly more complicated version of (15), based on a different version of (14),

$$\widehat{\mathrm{Var}}_{\mathrm{JK}}(\hat{\boldsymbol{\beta}}) = \frac{N-1}{N} \sum_{i=1}^{N} (\hat{\boldsymbol{\beta}}^{(i)} - \bar{\boldsymbol{\beta}})(\hat{\boldsymbol{\beta}}^{(i)} - \bar{\boldsymbol{\beta}})^{\top}, \tag{16}$$

where $\bar{\boldsymbol{\beta}} = N^{-1} \sum_{i=1}^{N} \hat{\boldsymbol{\beta}}^{(i)}$, the mean of the omit-one estimates.

Dividing by $(1 - h_i)^2$ actually seems to overcorrect the residuals.

But observations with large variances often tend to have residuals that are very much too small. Thus, HC_3 may be attractive if large variances are associated with large values of h_i.

Inferences based on any HCCME, especially HC_0 and HC_1, may be seriously inaccurate even when the sample size is moderately large if some observations have much higher leverage than others.

By default, Stata uses HC_1 with the "robust" and "vec(robust)" options. But "vce(hc2)" and "vce(hc3)" provide HC_2 and HC_3.

When Does Heteroskedasticity Matter?

Even when the disturbances are heteroskedastic, we do not necessarily have to use an HCCME.

Consider the jl^{th} element of $N^{-1}\boldsymbol{X}^{\top}\boldsymbol{\Omega}\boldsymbol{X}$, which is

$$\frac{1}{N}\sum_{i=1}^{N}\omega_i^2 x_{ij}x_{il}. \tag{17}$$

If the limit as $N \to \infty$ of the average of the ω_i^2 exists and is denoted σ^2, then expression (17) can be rewritten as

$$\sigma^2\frac{1}{N}\sum_{i=1}^{N}x_{ij}x_{il} + \frac{1}{N}\sum_{i=1}^{N}(\omega_i^2 - \sigma^2)x_{ij}x_{il}. \tag{18}$$

The first term here is just the jl^{th} element of $\sigma^2 N^{-1}\boldsymbol{X}^{\top}\boldsymbol{X}$.

If it happens that

$$\lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} (\omega_i^2 - \sigma^2) x_{ij} x_{il} = 0 \tag{19}$$

for $j, l = 1, \ldots, k$, then we find that

$$\lim_{N\to\infty} \frac{1}{N} \boldsymbol{X}^\top \boldsymbol{\Omega} \boldsymbol{X} = \sigma^2 \lim_{N\to\infty} \frac{1}{N} \boldsymbol{X}^\top \boldsymbol{X}. \tag{20}$$

If so, the asymptotic covariance matrix of $N^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$ is just

$$\Big(\lim_{N\to\infty} \frac{1}{N} \boldsymbol{X}^\top \boldsymbol{X} \Big)^{-1} \sigma^2 \Big(\lim_{N\to\infty} \frac{1}{N} \boldsymbol{X}^\top \boldsymbol{X} \Big) \Big(\lim_{N\to\infty} \frac{1}{N} \boldsymbol{X}^\top \boldsymbol{X} \Big)^{-1} = \sigma^2 \boldsymbol{S}_{\boldsymbol{X}^\top \boldsymbol{X}}^{-1}. \tag{21}$$

The usual OLS estimate of σ^2 is $s^2 = \big(1/(N-k)\big) \sum_{i=1}^{N} \hat{u}_i^2$.

If we assume that we can apply a law of large numbers, the probability limit of $N^{-1}\sum_{i=1}^{N}\hat{u}_i^2$ is

$$\lim_{N\to\infty}\frac{1}{N}\sum_{i=1}^{N}\omega_i^2 = \sigma^2. \tag{22}$$

In this special case, the usual OLS covariance matrix estimator $s^2(\boldsymbol{X}^\top\boldsymbol{X})^{-1}$ is valid asymptotically.

If we are estimating a sample mean, then $\boldsymbol{X} = \boldsymbol{\iota}$, and

$$\frac{1}{N}\sum_{i=1}^{N}\omega_i^2 x_{ij}x_{il} = \frac{1}{N}\sum_{i=1}^{N}\omega_i^2\iota_i^2 = \frac{1}{N}\sum_{i=1}^{N}\omega_i^2 \to \sigma^2 \text{ as } N\to\infty. \tag{23}$$

Thus (19) holds, and we do not have to worry about heteroskedasticity.

Only heteroskedasticity related to the squares and cross-products of the x_{ij} affects the validity of the usual OLS covariance matrix estimator.

HAC Covariance Matrix Estimation

The assumption that the matrix $\boldsymbol{\Omega}$ is diagonal is what makes it possible to estimate $N^{-1}\boldsymbol{X}^\top\boldsymbol{\Omega}\boldsymbol{X}$ consistently and obtain an HCCME, even though $\boldsymbol{\Omega}$ itself cannot be estimated consistently.

The matrix $N^{-1}\boldsymbol{X}^\top\boldsymbol{\Omega}\boldsymbol{X}$ can sometimes be estimated consistently for a model that uses time-series data when the disturbances are correlated across time periods.

Observations that are close to each other may be strongly correlated, but observations that are far apart may be uncorrelated or nearly so.

If so, only the elements of $\boldsymbol{\Omega}$ that are on or close to the principal diagonal are large.

We may be able to obtain an estimate of the covariance matrix of the parameter estimates that is **heteroskedasticity and autocorrelation consistent**, or **HAC**.

Cluster-Robust Inference

Data are often collected at the individual level, but each observation is associated with a higher-level entity, such as a city, state, province, or country, a classroom or school, a hospital, or perhaps a time period.

Thus each observation belongs to a **cluster**, and the regression disturbances may be correlated within the clusters.

It seems natural to allow for any form of correlation within each of G clusters, while assuming no correlation across clusters.

The resulting covariance matrix is called a **cluster-robust variance estimator** or **CRVE**.

Instead of presenting an elementary exposition based on ETM2, I will present a more detailed one based on:

James G. MacKinnon, Morten Ø. Nielsen, and Matthew D. Webb, "Cluster-robust inference: A guide to empirical practice." *Journal of Econometrics*, 2023, **232**, 272–299.