Heteroskedasticity

We now consider the implications of relaxing the assumption that the conditional variance $\text{Var}(u_i|x_i) = \sigma^2$ is common to all observations $i = 1, ..., n$

In many applications, we may suspect that the conditional variance of the error term could vary over the observations, in ways that are related to (some of) the explanatory variables in x_i, and which may be difficult to model convincingly

For example, the variance of shocks to GDP per capita may be quite different for developing countries that are dependent on primary commodity exports compared to developing countries with more diversified export structures, or compared to OECD countries

Or the variance of shocks to firm-level total factor productivity (TFP) may be quite different for firms in high-tech sectors compared to firms in low-tech sectors, or for recent entrants compared to more established firms

Allowing the conditional variance $\text{Var}(u_i|x_i) = \sigma^2(x_i) = \sigma_i^2$ to be different for different observations $i = 1, ..., n$ with different values of x_i is referred to as allowing for **conditional heteroskedasticity**

We have seen that the consistency property of the OLS estimator does not require the assumption of conditional homoskedasticity

We have also seen that the asymptotic normality property of the OLS estimator does not require the assumption of conditional homoskedasticity

The variance matrix in the limit distribution of $\sqrt{n}(\widehat{\beta}_{\text{OLS}}-\beta)$ has a different form in the more general case of conditional heteroskedasticity, but can still be estimated consistently

This allows us to extend the (asymptotic) Wald tests of restrictions on the parameter vector β to this more general setting

Recall our earlier the asymptotic normality result, that under the assumptions:

i) $y_i = x_i'\beta + u_i$ for $i = 1, ..., n$ or $y = X\beta + u$

ii) The data on (y_i, x_i') are independently and identically distributed, with $\mathrm{E}(x_i u_i) = 0$ for all $i = 1, ..., n$

iii) The $K \times K$ matrix $M_{XX} = \mathrm{E}(x_i x_i')$ exists and is non-singular

iv) The $K \times K$ matrix $M_{X\Omega X} = \mathrm{E}(u_i^2 x_i x_i')$ exists and is non-singular

Then $\sqrt{n}(\widehat{\beta}_{\mathrm{OLS}} - \beta) \xrightarrow{\mathrm{D}} \mathrm{N}(0, M_{XX}^{-1} M_{X\Omega X} M_{XX}^{-1})$

Now consider the result

$$\sqrt{n}(\widehat{\beta}_{\text{OLS}} - \beta) \xrightarrow{\text{D}} \text{N}(0, V)$$

where we have $V = M_{XX}^{-1} M_{X\Omega X} M_{XX}^{-1}$

As before, we can use this limit distribution for $\sqrt{n}(\widehat{\beta}_{\text{OLS}} - \beta)$ to obtain an approximation to the distribution of $\widehat{\beta}_{\text{OLS}}$ that will be accurate in large finite samples

We obtain

$$\widehat{\beta}_{\text{OLS}} \overset{a}{\sim} N\left(\beta, V/n\right)$$

$$\text{with} \;\; V/n = \left(\frac{1}{n}\right) M_{XX}^{-1} M_{X\Omega X} M_{XX}^{-1}$$

To make use of this approximation, we require a consistent estimator of the $K \times K$ variance matrix V, where now $V = M_{XX}^{-1} M_{X\Omega X} M_{XX}^{-1}$

Our earlier result that

$$\left(\frac{X'X}{n}\right)^{-1} \xrightarrow{\mathrm{P}} M_{XX}^{-1}$$

gives us a consistent estimator of the $K \times K$ matrix M_{XX}^{-1}

The remaining task is to find a consistent estimator of the $K \times K$ matrix $M_{X\Omega X}$

If we knew the error terms u_i for $i = 1, 2, ..., n$, then with iid observations, the $K \times K$ matrix of sample means

$$\frac{1}{n}\sum_{i=1}^{n} u_i^2 x_i x_i' \xrightarrow{\mathrm{P}} \mathrm{E}(u_i^2 x_i x_i') = M_{X\Omega X}$$

by the Law of Large Numbers, and we would have a consistent estimator of $M_{X\Omega X}$

White (Econometrica, 1980) showed that, under reasonable assumptions, the unknown error terms u_i for $i = 1, 2, ..., n$ in this expression can be replaced by sample residuals based on a *consistent* estimator of β, such as the OLS residuals $\widehat{u}_i = y_i - x_i'\widehat{\beta}_{\mathrm{OLS}}$, for which we have $\widehat{u}_i \xrightarrow{\mathrm{P}} u_i$

The resulting estimator

$$\frac{1}{n}\sum_{i=1}^{n}\widehat{u}_i^2 x_i x_i' \xrightarrow{\mathrm{P}} \mathrm{E}(u_i^2 x_i x_i') = M_{X\Omega X}$$

remains consistent

This is not straightforward to prove, and additionally requires finite fourth moments for the explanatory variables in x_i ($\mathrm{E}[(x_{ij}x_{ik})^2]$ exists and is finite for all $j, k = 1, 2, ..., K$)

[Section 2.5 in Hayashi (2000) provides a sketch of the proof for the special case with $K = 1$]

Given this result, using Slutsky's theorem, the estimator

$$\begin{aligned}\widehat{V} &= \left(\frac{X'X}{n}\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\widehat{u}_i^2 x_i x_i'\right)\left(\frac{X'X}{n}\right)^{-1} \\ &= n(X'X)^{-1}\left(\sum_{i=1}^{n}\widehat{u}_i^2 x_i x_i'\right)(X'X)^{-1}\end{aligned}$$

provides a consistent estimator of $V = M_{XX}^{-1} M_{X\Omega X} M_{XX}^{-1}$

Since $\widehat{V} \xrightarrow{\mathrm{P}} V$, the difference between V and $\widehat{V}$ becomes negligible in the limit as $n \to \infty$, and we can replace the unknown V by the estimator $\widehat{V}$ without changing the form of our limit distribution result

This gives the approximation to the distribution of $\widehat{\beta}_{\text{OLS}}$ as

$$\widehat{\beta}_{\text{OLS}} \overset{\text{a}}{\sim} N\left(\beta, \widehat{V}/n\right)$$

where

$$\widehat{V}/n = (X'X)^{-1}\left(\sum_{i=1}^{n} \widehat{u}_i^2 x_i x_i'\right)(X'X)^{-1}$$

We can compute $\widehat{V}/n$ using the data on X and the OLS residuals $\widehat{u}$

We can then construct asymptotic t-test and Wald test statistics as before, using this **heteroskedasticity-consistent** estimator $\widehat{V}/n$ in place of the estimator of the variance of $\widehat{\beta}_{\text{OLS}}$ that we obtained in the special case of conditional homoskedasticity

White's (1980) paper which introduced this heteroskedasticity-consistent estimator for the variance of $\widehat{\beta}_{\text{OLS}}$ is one of the most cited papers in econometrics (or economics) in the last 35 years, and has had a huge impact on empirical research in economics

Similar ideas can be found earlier in the statistics literature, in papers by Huber and by Eicker (both in 1967)

The square roots of the scalar elements on the main diagonal of $\widehat{V}/n$ are variously referred to as heteroskedasticity-consistent standard errors, or heteroskedasticity-robust standard errors, or White standard errors (or some combination of Eicker-Huber-White standard errors)

Heteroskedasticity-consistent standard errors and test statistics are available in most econometric software

To obtain these in Stata, we can use the vce(robust) option within the regress command, e.g.

reg y x1 x2, vce(r)

Asymptotic inference based on the non-robust estimator $\widehat{V}/n = \widehat{\sigma}^2_{\text{OLS}}(X'X)^{-1}$ that we derived under conditional homoskedasticity is valid (in large finite samples) *only* under this restrictive assumption

But since conditional homoskedasticity ($\mathrm{E}(u_i^2|x_i) = \sigma^2$) is a *special case* of conditional heteroskedasticity, asymptotic inference based on the consistent estimator $\widehat{V}/n = (X'X)^{-1}\left(\sum_{i=1}^{n}\widehat{u}_i^2 x_i x_i'\right)(X'X)^{-1}$ that we obtain under conditional heteroskedasticity is *also* valid (in large finite samples) if the model happens to satisfy conditional homoskedasticity

Notice that at no point do we obtain a consistent estimator of the n conditional variance parameters $\sigma_i^2 = \mathrm{E}(u_i^2|x_i)$ for each $i = 1, 2, ...n$

All that we require is a consistent estimator of the $K \times K$ matrix $M_{X\Omega X} = \mathrm{E}(u_i^2 x_i x_i')$, which as we have seen can be estimated consistently as $n \to \infty$ with K fixed

In applications where large data samples are available, a common response to the suspicion that conditional heteroskedasticity may be relevant is to continue to use the OLS estimator, and to use heteroskedasticity-consistent standard errors (and test statistics) in place of the traditional standard errors (and test statistics)

The OLS estimator remains consistent and asymptotically normal (under the assumptions stated previously), and we have a consistent estimator of the variance matrix, so that asymptotic inference remains valid in large finite samples

That is, we will reject a correct null hypothesis approximately 5% of the time at the 5% significance level (the *level* or *size* of the test is approximately correct)

And the probability of rejecting a false null hypothesis (the *power* of the test) increases with the sample size, tending to one in the limit as $n \to \infty$ (the test is said to be *consistent*)

This is sometimes referred to as a 'passive' response to heteroskedasticity

The OLS estimator is **not** asymptotically efficient in the case of conditional heteroskedasticity, but can still be used to conduct valid hypothesis tests in large finite samples

Testing for heteroskedasticity then becomes less important if we have the luxury of using large data samples, and are content to follow this passive strategy

Various tests are available that have power to detect conditional heteroskedasticity based on the OLS residuals, or to reject the null hypothesis of conditional homoskedasticity

White (Econometrica, 1980) suggested regressing the squared OLS residuals $\widehat{u}_i^2$ on a constant and on all the explanatory variables, and their squares and cross-products

For example, in the model with $K = 3$ and an intercept term

$$y_i = \beta_1 + \beta_2 x_{2i} + \beta_3 x_{3i} + u_i$$

we run the regression

$$\widehat{u}_i^2 = \gamma_1 + \gamma_2 x_{2i} + \gamma_3 x_{3i} + \gamma_4 x_{2i}^2 + \gamma_5 x_{3i}^2 + \gamma_6 (x_{2i} x_{3i}) + v_i$$

and test the restriction $H_0 : \gamma_2 = \gamma_3 = ... = \gamma_6 = 0$ (which is implied by the conditional homoskedasticity assumption $\mathrm{E}(u_i^2 | x_i) = \sigma^2$ for all $i = 1, ..., n$)

The basic idea is that we let $\sigma_i^2 = \mathrm{E}(u_i^2|x_i)$ be some unknown function $f(z_i)$ of a vector of observed variables z_i (which may include some or all of the explanatory variables in x_i)

We use the squared residuals $\widehat{u}_i^2$ as a proxy for u_i^2, and we use this polynomial approximation to the unknown function $f(z_i)$

Earlier tests for heteroskedasticity based on similar ideas include those proposed by Glejser (JASA, 1969), Ramsey (JRSS(B), 1969), Goldfeld and Quandt (1972) and Breusch and Pagan (Econometrica, 1979)

In some models with conditional heteroskedasticity, we can obtain more efficient estimators than OLS if we are willing to *model* the form that this conditional heteroskedasticity takes

This 'active' response to heteroskedasticity may be more appropriate in applications where efficiency is considered to be a more important concern

To see the basic idea, we first consider a version of the classical linear regression model with a **known form** of conditional heteroskedasticity

Generalized Least Squares

We assume

$$\mathrm{E}(y|X) = X\beta$$

$\mathrm{Var}(y|X) = \Omega$, with $\Omega \neq \sigma^2 I$ a known, positive definite $n \times n$

conditional variance matrix

X has full rank (with probability one)

Because Ω is positive definite and known, we can find a non-stochastic $n \times n$ matrix B with the properties that $B'B = \Omega^{-1}$ and $B\Omega B' = I$

Let $y^* = By$ and $X^* = BX$

Now

$$\mathrm{E}(y^*|X) = B\mathrm{E}(y|X) = BX\beta = X^*\beta$$

$$\mathrm{Var}(y^*|X) = B\mathrm{Var}(y|X)B' = B\Omega B' = I \ (= \sigma^2 I \text{ for } \sigma^2 = 1)$$

$X^* = BX$ has full rank (with probability one)

Since B is non-stochastic, conditioning on X and conditioning on $X^* = BX$ are equivalent

The transformed model

$$y^* = X^*\beta + u^* \quad \text{with} \ \ \mathrm{Var}(u^*|X^*) = I$$

is a classical linear regression model with conditional homoskedasticity

Aitken's theorem

As the transformed model satisfies the assumptions of the Gauss-Markov theorem, the OLS estimator of β in this transformed model, known as the Generalized Least Squares (GLS) estimator, is efficient (in the class of linear, unbiased estimators)

$$\begin{aligned}\widehat{\beta}_{\text{GLS}} &= (X^{*\prime}X^{*})^{-1}X^{*\prime}y^{*} \\ &= (X'B'BX)^{-1}X'B'By \\ &= (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}y\end{aligned}$$

We could replace B here by any matrix which is proportional to B, and still obtain the same GLS estimator

If we use $\widetilde{B} = aB$ for some scalar a, we have the transformed variables $\widetilde{y} = \widetilde{B}y = aBy$ and $\widetilde{X} = \widetilde{B}X = aBX$, giving

$$\begin{aligned}(\widetilde{X}'\widetilde{X})^{-1}\widetilde{X}'\widetilde{y} &= (a^2X'B'BX)^{-1}a^2X'B'By \\ &= (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}y = \widehat{\beta}_{\text{GLS}}\end{aligned}$$

Indeed we could replace B by QB, where Q is an $n \times n$ matrix with the property that $Q'Q = I$

Under the further normality assumption that

$$y|X \sim \mathrm{N}(X\beta, \Omega)$$

the GLS estimator is also the (conditional) Maximum Likelihood estimator in this particular model where Ω is known

In this case the exact finite sample distribution of $\widehat{\beta}_{\mathrm{GLS}}$ is also normal, with

$$\widehat{\beta}_{\mathrm{GLS}}|X \sim \mathrm{N}\left(\beta, (X'\Omega^{-1}X)^{-1}\right)$$

In practice this GLS estimator cannot be computed, since we don't know the conditional variance matrix Ω

The **Feasible Generalized Least Squares** (FGLS) estimator replaces the unknown Ω by an estimator $\widehat{\Omega}$, giving

$$\widehat{\beta}_{\text{FGLS}} = (X'\widehat{\Omega}^{-1}X)^{-1}X'\widehat{\Omega}^{-1}y$$

This can again be computed as an OLS estimator, using the transformed variables $y^* = \widehat{B}y$ and $X^* = \widehat{B}X$, where we now require $\widehat{B}'\widehat{B} = \widehat{\Omega}^{-1}$ and $\widehat{B}\widehat{\Omega}\widehat{B}' = I$

[or we could use any matrix that is proportional to $\widehat{B}$]

The properties of the FGLS estimator depend on the properties of $\widehat{\Omega}$ as an estimator of Ω

If $\widehat{\Omega}$ is a **consistent** estimator of Ω, then under quite general conditions we find that $\widehat{\beta}_{\text{FGLS}}$ has the same **asymptotic distribution** as the infeasible $\widehat{\beta}_{\text{GLS}}$, giving

$$\widehat{\beta}_{\text{FGLS}} \overset{\text{a}}{\sim} \text{N}\left(\beta, (X'\widehat{\Omega}^{-1}X)^{-1}\right)$$

In this case, $\widehat{\beta}_{\text{FGLS}}$ is also asymptotically efficient

However obtaining a consistent estimator of Ω is *not* straightforward

In general, the $n \times n$ symmetric matrix Ω has $n(n+1)/2$ distinct elements

Even if we restrict all the off-diagonal elements to be zero, as is natural in a cross-section regression context with independent observations, so that

$$\Omega = \begin{pmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_n^2 \end{pmatrix}$$

we still have n distinct elements

These *cannot* be estimated consistently from a sample of size n

Consistent estimation requires us to specify a (parametric) **model** for the conditional variance matrix $\mathrm{Var}(y|X) = \Omega$, of the form $\Omega = \Omega(\phi)$, in which the $n \times n$ matrix $\Omega(\phi)$ is a function of the vector ϕ, which contains a finite number of additional parameters, not increasing with the sample size n, and which *can* be estimated consistently from the data

If this specification of the conditional variance matrix $\mathrm{Var}(y|X) = \Omega(\phi)$ is *correct*, *and* we can find a consistent estimator $\widehat{\phi}$ of the vector ϕ, we can then use the consistent estimator $\widehat{\Omega} = \Omega(\widehat{\phi})$ to obtain the asymptotically efficient FGLS estimator

As a very simple example (the implementation of which does not require the estimation of any additional parameters), we could specify the conditional variance $\text{Var}(y_i|X) = \text{Var}(u_i|X) = \sigma_i^2$ to be proportional to the *squared* values of one of the regressors, say x_{Ki}, giving $\sigma_i^2 = \sigma^2 x_{Ki}^2$

We then let $y_i^* = \frac{y_i}{x_{Ki}}$ and $x_{ki}^* = \frac{x_{ki}}{x_{Ki}}$ for each $k = 1, ..., K$ (notice how this transformation affects the intercept in the model)

In the transformed model $y_i^* = x_i^{*\prime}\beta + u_i^*$ we then have

$$\text{Var}(y_i^*|X) = \text{Var}(u_i^*|X) = \frac{\sigma_i^2}{x_{Ki}^2} = \frac{\sigma^2 x_{Ki}^2}{x_{Ki}^2} = \sigma^2 \quad \text{for all } i = 1, ..., n$$

This transformed model then satisfies conditional homoskedasticity, and we can compute the FGLS estimator here simply as the OLS estimator in the transformed model

Feasible GLS estimators of this kind are also known as Weighted Least Squares estimators, since we weight each observation by a factor which is proportional to $\frac{1}{\sigma_i}$ (or, more generally, to an estimator $\frac{1}{\widehat{\sigma}_i}$ of $\frac{1}{\sigma_i}$)

Note that the transformation gives *less* weight to observations where the variance of u_i is (estimated to be) relatively high, and *more* weight to observations where the variance of u_i is (estimated to be) relatively low

If our specification for the conditional variance $\mathrm{Var}(y|X) = \Omega(\phi)$ is correct, and we estimate $\Omega(\phi)$ consistently, this weighting is the source of the efficiency gain compared to OLS

If we specify the conditional variance $\mathrm{Var}(y|X) = \Omega = \Omega(\phi)$ and further assume that

$$y|X \sim \mathrm{N}(X\beta, \Omega(\phi))$$

the *feasible* GLS estimator is **not** the (conditional) Maximum Likelihood estimator, in the case where ϕ is unknown and has to be estimated

FGLS uses a consistent estimator of ϕ to construct a consistent estimator of Ω, and then maximizes $\mathcal{L}(\beta, \widehat{\Omega}) = \mathcal{L}(\beta, \Omega(\widehat{\phi}))$ with respect to β

The (conditional) Maximum Likelihood estimator maximizes the likelihood function $\mathcal{L}(\beta, \Omega) = \mathcal{L}(\beta, \Omega(\phi))$ with respect to β and ϕ jointly

These estimators are different, *unless* we happen to have $\widehat{\phi} = \widehat{\phi}_{\mathrm{ML}}$, giving $\widehat{\Omega} = \Omega(\widehat{\phi}_{\mathrm{ML}}) = \widehat{\Omega}_{\mathrm{ML}}$, in which case $\mathcal{L}(\beta, \widehat{\Omega}_{\mathrm{ML}})$ is a **concentrated likelihood function**, and maximizing $\mathcal{L}(\beta, \widehat{\Omega}_{\mathrm{ML}})$ with respect to β does yield the (conditional) Maximum Likelihood estimator $\widehat{\beta}_{\mathrm{ML}}$

In most applications of Feasible GLS, we do *not* have $\widehat{\phi} = \widehat{\phi}_{\mathrm{ML}}$

Then $\widehat{\beta}_{\mathrm{FGLS}} \neq \widehat{\beta}_{\mathrm{ML}}$, although the two estimators are asymptotically equivalent (i.e. they have the same asymptotic distribution) under quite general conditions

The *consistency* properties of $\widehat{\beta}_{\text{FGLS}}$ and $\widehat{\beta}_{\text{ML}}$ in this linear regression model do not depend on the parametric specification of $\Omega = \Omega(\phi)$ being correct

But the *efficiency* advantages of these estimators relative to $\widehat{\beta}_{\text{OLS}}$ may not hold if this specification for the form of the conditional variance matrix is not correct

These estimators also do not extend straightforwardly to linear models that do not satisfy the linear conditional expectation assumption $\text{E}(y_i|x_i) = x_i'\beta$