Heteroskedasticity

Recall the assumption of the CLNM

S1: $y = X\beta + u$

S2: X has full column rank

S3: $E(u|X) = 0$

S4: $E(uu'|X) = \sigma^2 I_n$

S5: $u|X \sim MVN$

We have also discussed some reasons why these assumptions may be violated, and some potential solutions for deviations from this list. Below, I give a partial list and mention a few things that we won’t have time to cover.

Deviations from S1:

- OLS estimates the coefficients of the BLP. But, in general, we would expect the regression function to be nonlinear.
- We can introduce some nonlinearity into our model by adding nonlinear transformations (such as x_j^2) or splines (linear or cubic, with or without continuity restrictions, implemented by introducing dummy variables) but stay within the linear in parameters framework.
- The logical extension of these ideas is nonparametric regression, i.e. estimate $E(y|X) \equiv f(x)$ directly. This is logically satisfying but suffers from the *curse of dimensionality*: we need enormous amounts of data once the number of regressors exceeds around 3. We won’t be studying nonparametrics in this course.

- Related to nonlinearity is the idea that the parameters may vary with observations as functions of their regressors or other observable characteristics. In practice, we can deal with variation that depends on observables by building models of the form

  $$\beta_{ik} = \beta_k + x_i\delta_k + \sum D_{ikj}\gamma_{kj}$$

  where D_{ikj} are dummy variables.

Deviations from S2:

- If X doesn't have full column rank, then there is no unique solution to the normal equations. Although $\widehat{\beta}$ is not unique, the OLS fitted values and residuals, $\hat{y}$ and $\hat{u}$, are unique.
- If the columns of X are singular "in population", then no sample could identify β and we should think seriously about redefining the model (eg. the dummy variable trap).
- If the singularity of X is a problem for our sample (not in population), then we could either focus on the l.c. of β that are estimable, or bring in extraneous information (restricted OLS).
- Older texts used to talk about "multicollinearity" as somehow a kind of violation of S2. It's not.

Deviations from S3:

- This is probably the most critical assumption.
- We can weaken the mean independence assumption $E(u|X)$ to a "zero covariance assumption" $E(x_i u_i) = 0$. This makes finite sample results difficult to obtain, but we can still obtain consistency (i.e. show that the OLS estimator $\widehat{\beta}$ converges to the coefficient vector of the BLP). However, in many settings, we are not interested in the BLP of y_i given x_i
- If we have "left out variables", then OLS will not estimate the partial response of y_i to x_i, but the sum of that response and an indirect proxy for the effects of left out variables.
- We’ll see other reasons for correlation between x_i and u_i and some solutions later in the lectures.

Deviations from S5:

- We can still do OLS and our standard tests, but base our results on asymptotic theory. But we lose optimality properties for the OLS estimator and for our testing procedures.
- We can try to improve on the "first order"asymptotics by using computer intensive methods such as the bootstrap for better finite sample approximations. (We won’t cover this or anything below).
- If the errors are not conditionally normal, but we know their distribution, then we can do ML. This often gives us good large sample properties, but without a finite sample theory.
- With a known distribution, we can use analytical approximations (Edgeworth or saddlepoint), or Bayesian methods.

Deviations from S4:

- There are many important applications which involve violations of the conditional scalar covariance matrix assumption, i.e. where we replace

$$\text{S4: } E(uu'|X) = \sigma^2 I_n \text{ with } \text{S4}' : E(uu'|X) = \Omega$$

  where Ω is a positive definite matrix.

- If Ω is only positive semidefinite, then formally it is equivalent to having fewer observations but imposing linear restrictions on the coefficient vector.

- Leading applications include time series, systems of equations, panel data models and
- *heteroskedasticity*

$$\Omega = diag(\sigma_i^2) \quad \sigma_i^2 > 0$$

:Consequences of nonscalar covariance matrix for OLS

a) Sampling Distribution

- From S1 and S2

$$\widehat{\beta} = \beta + Lu$$

- From S1-S3

$$E(\widehat{\beta}|X) = \beta = E(\widehat{\beta})$$

- Adding S4$'$

$$V(\widehat{\beta}|X) = LV(u)L' = (X'X)^{-1}X'\Omega X(X'X)^{-1}$$

- Adding S5

$$\widehat{\beta}|X \sim MVN(\beta, (X'X)^{-1}X'\Omega X(X'X)^{-1})$$

b) Optimality Properties

- OLS is no longer Gauss-Markov
- OLS is no longer MVUE
- If we use the OLS covariance matrix, $\widehat{\sigma}^2(X'X)^{-1}$, then t and F statistics for linear hypotheses will not have the correct size. Test procedures that use a "consistent" estimator of $V(\widehat{\beta}|X)$ will have the correct asymptotic size, but they are no longer optimal. For example, the (asymptotic) t-test of $H_0 : \beta_i = 0$ vs $H_0 : \beta_i > 0$ based on

$$\frac{\widehat{\beta}_i}{s.e.(\widehat{\beta}_i)}$$

is no longer UMP (uniformly most powerful).

:Heteroskedasticity consistence covariance matrix estimator (HCCME) aka "Heteroskedasticity Robust" CME

- With $\Omega = diag(\sigma_i^2)$, we can write

$$\begin{aligned} X'\Omega X &= \sum_{i=1}^{n} \sigma_i^2 x_i' x_i \\ &= \sum_{i=1}^{n} E(u_i^2 x_i' x_i | X) \end{aligned}$$

  (Note that both right and left hand sides live in $\mathbb{R}^{KxK}$)

- If we knew σ_i^2 we could use the first row on the rhs.
- Otherwise, if we saw the disturbances, we can use fact that rhs (divided by sample size) is a matrix of sample means and invoke a LLN to estimate it.

- White (1980) [and earlier Eicker (1967)] showed that we can replace the unobserved disturbances with the OLS residuals in the argument above, i.e.

$$p\lim\left(\frac{1}{n}X'\widehat{\Omega}X - \frac{1}{n}X'\Omega X\right) = 0$$

$$\text{where } X'\widehat{\Omega}X = \sum_{i=1}^{n}\hat{u}_i^2 x_i' x_i$$

Rk: Notice that standard OLS uses $X'\widehat{\Omega}X = \hat{\sigma}^2 \sum_{i=1}^{n} x_i' x_i$

- Moreover (using my abuse of notation)

$$\widehat{\beta}|X \sim^a N(\beta, (X'X)^{-1}X'\widehat{\Omega}X(X'X)^{-1})$$

so the usual tests are asymptotically correct if we use this new estimator for

$$\widehat{V}(\widehat{\beta}|X) = (X'X)^{-1}X'\widehat{\Omega}X(X'X)^{-1}$$

- Wald test for the general linear hypothesis $R\beta = r$ is derived just as before. Under the null,

$$R\hat{\beta}|X \sim^a N(r, R\widehat{V}(\hat{\beta}|X)\,R')$$

Therefore, by the continuous mapping theorem

$$\left(R\hat{\beta} - r\right)'\left[R\widehat{V}(\hat{\beta}|X)\,R'\right]^{-1}\left(R\hat{\beta} - r\right) \sim^a \chi^2(q)$$

- Notice that this tests statistic will NOT reduce to an expression involving restricted and unrestricted sum of squares.
- There are other HCCMEs that appear to have better finite sample properties that the White’s.

- Should we always use a HCCME?
  - If no heteroskedasticity, then we can get exact distribution use standard OLS estimator for covariance matrix.
  - If the heteroskedasticity is "small", then we can do worse by trying to estimate it than by acting as if it is zero (usual bias vs. variance tradeoff).
  - For large sample sizes, it makes sense to report only standard erors and test statistics that use a HCCME (assuming that we don’t have to worry about correlations in disturbances across observations).

Testing for heteroskedasticity

- If we saw the disturbances, then we could build skedastic models

  $$u_i^2 = \delta_0 + \delta_1 z_1 + \delta_2 z_2 + \cdots + \delta_p z_p + \nu_i$$

  Note that $\{z_i\}_1^p$ and $\{x_i\}_1^k$ may have some elements in common, but there may be some variables that appear only in the population regression function (PRF) $E(y|X)$ or the skedastic function.

- The null hypothesis of homoskedasticity reduces to

  $$H_0 : \delta_1 = \delta_2 = \cdots = \delta_p = 0$$

  We could test this using a variety of asymptotic tests.

- In practice, we can use the OLS residuals in place of the unobserved disturbances, and model

  $$\hat{u}_i^2 = \delta_0 + \delta_1 z_1 + \delta_2 z_2 + \cdots + \delta_p z_p + \nu_i$$

  then construct the test statistics

  $$\frac{R_{\hat{u}^2}^2/p}{\left(1 - R_{\hat{u}^2}^2\right)/(n-p-1)} \sim^a F(p, n-p-1)$$

  or $\quad LM = nR_{\hat{u}^2}^2 \sim^a \chi^2(p)$

  Rk: *LM* is the LM test suggested by Breusch and Pagan (1979), but with a covariance matrix that is robust to deviations from normality.

- Breusch and Pagan showed that the form of the LM test was the same if we used skedastic functions of the form

  $$E(u_i^2|Z) = h(\delta_0 + \delta_1 z_1 + \delta_2 z_2 + \cdots + \delta_p z_p)$$

where h is any C^1 function (because a first order Taylor series is good enough for local alternatives, the form of the test statistic doesn't vary with h).

- Examples (choices for z):
  - Choose some or all of the components of x_i (the latter is the original Breusch-Pagan test)
  - Include all linearly independent levels and cross-products of the x_i. This gives us the White (1980) test. This is a test of whether $V(\widehat{\beta}|X) = \sigma^2(X'X)$
  - If $z_i = (\widehat{y}_i, \widehat{y}_i^2)$ we get another common test.

- Tests for heteroskedasticity detect deviations from S1-S4(S5). If we reject the null, it could be due to misspecification of the PRF. For example, there could be neglected nonlinearity.

- It is straightforward to take residuals from a regression and running our skedastic regressions as tests. In STATA, after a regression, we can use the postestimation commands to perform our tests automatically (you should do it both ways and compare to make sure you understand what the package is doing):
    - we can use the option estat hettest for the BP test. NOTE: the default option assumes normal residuals. Using the option "iid" gives the nR^2 form of the LM test. Using the option "fstat" gives the robust form of the *F*-test. The default uses all the regressors. You can use a subset or add some.
    - STATA also implements a test that uses *ranks* of the regressors as the z variables (szroeter)

Efficient estimation with heteroskedasticity

- Suppose $E(uu'|X) = \sigma^2 diag(h_i)$ where h_i is known, i.e., we know the form of the heteroskedasticity, perhaps up to some constant. This is a bit more general than the case where we know $diag(\sigma_i^2)$. It's easy to see that we can transform the model into a form where S1-S4 hold, and then use our previous results.

- Consider a typical observation. We have the model

$$y_i = x_i\beta + u_i$$

where $E(u_i|X) = 0$ and $E(u_iu_j|X) = \sigma_i^2\delta_{ij}$, where δ_{ij} denotes the so-called *Kronecker delta* function: $\delta_{ij} = 1$ if $i = j$, and 0 otherwise.

- If we multiply each observation by a different constant, say c_i, we obtain

$$\begin{aligned} c_i y_i &= c_i x_i \beta + c_i u_i \\ \Leftrightarrow y_i^* &= x_i^* \beta + u_i^* \end{aligned}$$

  where $E(u_i^*|X) = c_i E(u_i|X) = 0$ and $E(u_i^* u_j^*|X) = c_i c_j E(u_i u_j|X) = c_i^2 \sigma_i^2 \delta_{ij}$.

- So scaling each observation by its own constant doesn't destroy the zero conditional mean property or the lack of correlation across distinct observations. But it does change the conditional variance.

- If we pick c_i^2 to be proportional to σ_i^{-2}, we can induce homoskedasticity. Given our assumption above, this says that the model

$$\frac{y_i}{\sqrt{h_i}} = \frac{x_i}{\sqrt{h_i}}\beta + \frac{u_i}{\sqrt{h_i}}$$

$$\Leftrightarrow y_i^* = x_i^*\beta + u_i^*$$

  satisfies S1-S4(S5).

- The GM (MLE) estimator is OLS on the transformed data.
- Notice that even if x_i contains an intercept, x_i^* won’t. So R^2 may not make much sense.

In matrix notation

- Suppose we have the model
$$y = X\beta + u$$
where $E(u|X) = 0$ and $E(uu'|X) = \sigma^2 H$ where H is known.
- Take any matrix $H^{-1/2}$ that satisfies $H^{-1/2}H(H^{-1/2})' = I_n$. Premultiplying yields
$$H^{-1/2}y = H^{-1/2}X\beta + H^{-1/2}u \quad \Leftrightarrow$$
$$y^* = X^*\beta + u^*$$
where $E(u^*|X) = 0$ and $E(u^*u^{*\prime}|X) = \sigma^2 I_n$. So the transformed model satisfies S1-S4.
- OLS on the transformed model is GM:
$$\begin{aligned}\widehat{\beta}_{GM} &= (X^{*\prime}X^*)^{-1}X^{*\prime}y^* \\ &= (X'H^{-1}X)^{-1}X'H^{-1}y\end{aligned}$$

- $\widehat{\beta}_{GM}$ is also called the *Generalized Least Squares (GLS)* estimator, denoted $\widehat{\beta}_{GLS}$
- If $u|X \sim MVN$, then $u^*|X^* \sim MVN$ so the transformed data satisfy S1-S5.
- The OLS estimator of σ^2 and test procedures are exactly what we've already seen, except that we replace (y, X) with (y^*, X^*)

Weighted Least Squares

- If we express the GLS estimator with heteroskedasticity on an observation by observation basis, we get

$$\begin{aligned}\widehat{\beta}_{GLS} &= \arg\min_{\widetilde{\beta}\in\mathbb{R}^K} \sum (y_i^* - x_i^*\beta)^2 \\ &= \arg\min_{\widetilde{\beta}\in\mathbb{R}^K} \sum w_i(y_i - x_i\beta)^2\end{aligned}$$

  where $w_i = h_i^{-1}$.
- The GLS estimator with heteroskedasticity belongs to the class of *weighted least squares estimators.*
- The optimal weights are proportional to the inverse of the variance of the disturbance u_i. Observations with large variances are given small weights; observations with small variances are given large weights.

Feasible GLS

- Suppose we believe the skedastic function is of the form

$$E(u_i^2|Z) = \sigma^2 \exp(\delta_1 z_1 + \delta_2 z_2 + \cdots + \delta_p z_p)$$

- If we knew the parameters $\delta_1, \delta_2, \cdots, \delta_p$ we'd do GLS.
- *Feasible GLS (FGLS)* replaces the unknown parameter vector δ with a consistent estimator.

Rks:

- We can include variables in Z that appear in the PRF *plus* variables that are irrelevant to the mean but matter for the variance.
- For testing purposes, it's OK to use a linear skedastic function (local alternatives are linear), but a model for the skedastic function used in FGLS must ensure that the estimated variances are positive.

Strategy:

1. Regress $\ln \hat{u}^2$ on z_1 to z_p.   Call the fitted values $\hat{g}_i$.

2. Act as if $h_i = \exp(\hat{g}_i)$ and follow the procedure for GLS.

Rks:

- Because we only have to get something proportional to h_i we don’t have to worry about scaling our prediction as in W6.42.
- If we have any observations with $\hat{u} = 0,$ we’ll have to estimate the skedastic function by nonlinear OLS, or replace all the zeros with a "small number".

Properties of Feasible GLS estimator

$$p\lim \sqrt{n}\,(\hat{\beta}_{GLS} - \hat{\beta}_{FGLS}) = 0$$

- Therefore, $\hat{\beta}_{FGLS}$ has the same asymptotic distribution as the GM estimator; we don’t lose anything (according to the standard first order asymptotic theory) from estimating the skedastic function!

A summing up

- Using OLS and a HCCME allows us to do correct inference, asymptotically. But we give up efficiency. Also, in practice, the HCCME estimator may not work well if there is "heavy" heteroskedasticity.
- Building a model for the scedastic function allows us to gain efficiency. But it’s a lot of work given our interest is in the

PRF. And there’s not much economic theory available to us for second moment specifications. It’s easy to get the scedastic function wrong which could make matters worse than using OLS.

- I recommend if specification tests point to lots of evidence of heteroskedasticity, then try to model it. Use specification tests on the *transformed* model to see how well you do in removing it.
- After an attempt to clean up the heteroskedasticity, use HCCME to get consistent standard errors and correctly sized tests.

Some Tips

- If you build a model for $E(u_i^2)$, use $h_i = \max(\widehat{E}(u_i^2), 10^{-3}\widehat{\sigma}^2)$. My constant 10^{-3} is pretty arbitrary, but the message is don’t let any single observation have too much weight.
- Tests for heteroskedasticity can pick up other deviations from the CLM. We’ll develop other specification tests in the next few lectures. Try to see which direction the tests are pointing toward before you decide whether the problem is heteroskedasticity, functional form, or serial correlation (or some combination).
- If OLS and GLS estimates are too different then something went wrong. Typically, it means that you have a violation of the assumption that $E(u|Z) = 0$, i.e. either you have missed important nonlinearities or you have included variables in your scedastic function that should have been in the PRF.