ECON-122
Introduction to Econometrics

Agnieszka Postepska

August 4th, 2011

Weighted least squares estimation - known form of heteroskedasticity

- recall that we define heteroskedastcity as: $Var(u|x) = \sigma^2 h(x)$
- in this expression $h(x)$ is the form of heteroskedasticity
- this is a rather unlikely situation but derivations are quite simple so we will start with this scenario
- we will show that if we specify the function of the variance correctly, than WLS is more efficient than OLS
- we will also talk about consequences of using the wrong form of the variance in the WLS procedure
- consider the following example:

$$\begin{aligned} sav_i &= \beta_0 + \beta_1 inc_i + u_i \\ Var(u_i|inc_i) &= \sigma^2 inc_i \end{aligned}$$

- so in the example above we have the following form of heteroskedasticty: $h(x) = h(inc) = inc$

WLS - known form of heteroskedasticity cont.

- the general form for a model with heteroskedasticity:

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_{i1} + ... + \beta_k x_k + u_i \\ Var(u_i|X_i) &= \sigma^2 h(X_i) \end{aligned}$$

- so in the example above we have the following form of heteroskedasticty: $h(x) = h(inc) = inc$
- it is important to remember that for now we are talking about a known form of heteroskedastcity
- in order to correct for heteroskedasticty we will transform the original model so that the transformed model satisfies the assumption on homoskedasticity

WLS - known form of heteroskedasticity cont.

- divide the model by $\sqrt{h(X_i)}$

$$\begin{aligned} \frac{y_i}{\sqrt{h(X_i)}} &= \frac{\beta_0}{\sqrt{h(X_i)}} + \frac{\beta_1 x_{i1}}{\sqrt{h(X_i)}} + ... + \frac{\beta_k x_k}{\sqrt{h(X_i)}} + \frac{u_i}{\sqrt{h(X_i)}} \\ y_i^* &= \beta_0 x_{i0}^* + \beta_1 x_{i1}^* + ... + \beta_k x_k^* + u_i^* \end{aligned}$$

  where $x_{i0}^* = \frac{1}{\sqrt{h(X_i)}}$

- notice that the transformed model does not have a constant term - the regressors in this model are meaningless, however, this model allows us to identify the parameters of the *original model*
- intuition behind this transformation: every observation i is weighted by $\frac{1}{\sqrt{h(X_i)}}$, so less weight is given to observations with greater error variance (bigger denominator, so smaller fraction) and more weight is put on observation with smalled error variance

WLS - known form of heteroskedasticity cont.

- most importantly, the transformed model satisfies assumption A5 - error term u^* is homoskedastic:

$$\begin{aligned}
Var(u_i|X_i) &= \sigma^2 h(X_i) \text{ by assumption} \\
E(u_i^2|X_i) - E(u_i|X_i)^2 &= \sigma^2 h(X_i) \text{ by definition of a variance} \\
\frac{E(u_i^2|X_i) - E(u_i|X_i)^2}{h(X_i)} &= \sigma^2 \text{ divide both sides by } h(X_i) \\
\frac{E(u_i^2|X_i)}{h(X_i)} - \frac{E(u_i|X_i)^2}{h(X_i)} &= \sigma^2 \\
E(\frac{u_i^2}{h(X_i)}|X_i) - E(\frac{u_i}{\sqrt{h(X_i)}}|X_i)^2 &= \sigma^2 \\
E(u_i^{*2}|X_i) - E(u_i^*|X_i)^2 &= \sigma^2 \\
Var(u_i^*|X_i) &= \sigma^2 \text{ WWTBS}
\end{aligned}$$

- so the *transformed model* satisfies all five assumptions and we can proceed with OLS estimation

WLS - known form of heteroskedasticity cont.

Interpretation: Recall the original and starred model:

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_k + u_i \\ y_i^* &= \beta_0 x_{i0}^* + \beta_1 x_{i1}^* + \ldots + \beta_k x_k^* + u_i^* \end{aligned}$$

- the key here is that we are estimating the starred model and it gives us estimates of the parameters in our original model
- in the original model the parameters have their usual interpretation
- in the starred model the parameters have no meaningful interpretation (similarly R^2)

WLS - known form of heteroskedasticity cont.

- to summarize, with KNOWN form of heteroskedasticity we transform our variables to get the variables in the starred model
- thus we estimate a model that satisfies all 5 assumptions - and we know that if A1-A5 are satisfied OLS is BLUE
- so, if form of heteroskedasticity is correctly specified, WLS is BLUE
- note that OLS estimates of the original model are still unbiased but WLS uses more information and therefore the estimates have smaller variance (WLS is taking more information from observations less "polluted" by variance in error)
- t-statistics and F-statistics can be computed from the starred model as before
- we can think of OLS as a special case of WLS as it assigns equal weights to all observations

Unknown form of heteroskedasticity

- WLS was a solution to KNOWN form of heteroskedasticity
- when we don't know the form of heteroskedasticity, then WLS is not feasible - we cannot compute the weights
- actually, it is very unlikely that a researcher know the form of heteroskedasticity and if the assumed is incorrect the estimator is no longer BLUE
- we said before that we can think of OLS as a special case of WLS
- WLS, on the other hand, is a special case of GENERALIZED LEAST SQUARES (GLS) ESTIMATORS
- now we will meet FEASIBLE GENERALIZED LEAST SQUARES estimation and we will see how we can estimate the form heteroskedasticity if we don't know in advance
- FGLS is similar to WLS except that it uses the estimated variance -covariance matrix since the true one is unknown (unknown form of heteroskedastcity)

FGLS - estimator for $h(X_i)$

- there are many ways in which one can model heteroskedasticity - we will focus on a fairly flexible one:

  $$Var(u|X) = \sigma^2 exp(\delta_0 + \delta_1 x_1 + \delta_2 x_2 + ... + \delta_k x_k)$$

  where $x_1, ..., x_k$ are the independent variables in the main regression model and the $\delta's$ are the unknown parameters
- therefore, $h(x) = exp(\delta_0 + \delta_1 x_1 + \delta_2 x_2 + ... + \delta_k x_k)$
- note that this functional form ensures that our estimated variances are positive (if we have assumed linear function we could obtain negative predicted values)

FGLS - estimator for $h(X_i)$

- using the assumed form of heteroskedasticity, the variance of the error term can be modeled in the following way:

  $$u^2 = \sigma^2 exp(\delta_0 + \delta_1 x_1 + \delta_2 x_2 + ... + \delta_k x_k)v$$

  where $E(v|x) = 1$
- to be able to estimate this equation we first need to transform this model into a linear form - take logs:

  $$log(u^2) = \alpha_0 + \delta_1 x_1 + \delta_2 x_2 + ... + \delta_k x_k + e$$

  where e has mean zero and is independent of $x's$
- and as we do not observe errors, the $u's$, we have to replace them with residuals, and we run the regression of $ln(\hat{u}^2)$ on $x_1, ..., x_k$
- lets define the fitted values from this regression as $\hat{g}_i$
- then, $\hat{h}(X_i) = exp(\hat{g}_i)$
- and we are done as we have estimated the weights - each observation is now weighted by $\frac{1}{\sqrt{\hat{h}(X_i)}}$

FGLS - cook book procedure

1 Run OLS regression of y on $x_1, ..., x_k$ and obtain residuals, $\hat{u}$

2 Create $ln(\hat{u}^2)$ - first squared the residuals than obtain the natural log

3 Run the regression of $ln(\hat{u}^2)$ on $x_1, ..., x_k$ and obtain residuals, $\hat{g}$

4 Exponentiate the fitted values $\hat{g}$ and obtain $\hat{h}(X_i) = exp(\hat{g}_i)$

5 Transform the original model using the estimated weights, $\hat{h}(X_i)$

6 Estimate the transformed model using WLS

7 Interpret coefficients of the original model

FGLS: final remarks

- FGLS is biased (and therefore not BLUE) but *consistent* - consistency is a concept similar to unbiasedness but refers to large samples - so we can say that FGLS is asymptotically unbiased
- this means that it will only work in large samples
- interpretation is the same is in WLS - so even though we estimate the starred model (the weighted model) we obtain parameters of the true model
- the transformed model has no useful interpretation - we need it to obtain efficient and consistent estimates of the parameters of the main model
- similarly as in WLS all test that we have seen are valid for the transformed model but in F test both restricted and unrestricted models must be weighted

OLS vs. FGLS

- OLS and FGLS will always produce different coefficients due to different sampling error (we are transforming the model in a way that doesn't preserve the variance of $x's$)
- also, as estimates of the standard errors are going to be different, our conclusions concerning significance of coefficients will change
- as long as the signs do not flip between OLS and FGLS or the magnitudes don't differ significantly, there is nothing to worry about
- if one of the above occurs, it usually indicates violation of some other assumption
- there are no particular rules about when you should start being worried but if this problem occurs you should inspect your model

Consequences of assuming wrong form of heteroskedasticity

- as we do need homoskedasticity to prove unbiasedness, assuming incorrect form of heteroskedasticity does not bias our estimates
- however, we are not fixing the initial problem - we are just changing the form of heteroskedasticity that the model is suffering from:

$$Var(u_i|X_i) = Var(\frac{u_i}{\sqrt{f(X_i)}}|X_i) = \frac{\sigma^2 h(X_i)}{f(X_i)}$$

  where $h(X_i)$ is the true form and $f(X_i)$ is the wrongly assumed form
- if the functional form for the variance is misspecified than we cannot say which one is better OLS or WLS
- what we can do is apply a heteroskedasticity-robust standard erros (that are good for any form of heteroskedasticity to fix this problem)
- in practice it is usually better to do WLS even if the assumed form is wrong than do simple OLS (asymptotics involved in explanation so we'll skip it)