

More on FGLS and How to Test for Heteroskedasticity

Generalized Least Squares and Feasible GLS

Here's a little more detail on how to perform GLS/FGLS:
Suppose you have a standard multivariate model such as

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$$

which satisfies all of the standard assumptions except $Var(\varepsilon_i|x_{1i}, x_{2i}) \neq \sigma_\varepsilon^2$. Instead, you are told (or you guess) that

$$Var(\varepsilon_i|x_{1i}, x_{2i}) = \sigma_\varepsilon^2 g(x_{1i}, x_{2i}) = \sigma_\varepsilon^2 (\alpha_0 + \alpha_1 |x_{1i}| + \alpha_2 |x_{2i}|)$$

This is definitely not the only way to model the standard errors. For example, you can use interactions, higher powers of the x's, etc. There are two issues to watch out for:

1. All Variances must be positive (the reason for using the exponential function above)
2. You need to believe in your model! In most cases, there is no strong theory as to what the heteroskedasticity should look like. This is why many empirical economists just "punt" and just use robust standard errors.

If you are confident in your model and you know α_0, α_1 , and α_2 , then you can just use generalized least squares. Instead of running the model

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$$

you need to divide all of the variables by $\sqrt{g(x_{1i}, x_{2i})}$ and regress

$$\frac{y_i}{\sqrt{g(x_{1i}, x_{2i})}} = \frac{\beta_0}{\sqrt{g(x_{1i}, x_{2i})}} + \frac{\beta_1 x_{1i}}{\sqrt{g(x_{1i}, x_{2i})}} + \frac{\beta_2 x_{2i}}{\sqrt{g(x_{1i}, x_{2i})}} + \frac{\varepsilon_i}{\sqrt{g(x_{1i}, x_{2i})}}$$

Now, we have eliminated the heteroskedasticity, since

$$Var\left(\frac{\varepsilon_i}{\sqrt{g(x_{1i}, x_{2i})}}\right) = \frac{1}{g(x_{1i}, x_{2i})} Var(\varepsilon_i) = \sigma_\varepsilon^2$$

This procedure is relatively easy, but not feasible since we almost never know the parameters of the variance equation, α_0, α_1 , and α_2 . Hence the need for Feasible Generalized Least Squares.

The general idea of FGLS is to use the estimated $\hat{\varepsilon}_i$ from the standard OLS equation, then estimate $\hat{\alpha}_0, \hat{\alpha}_1$, and $\hat{\alpha}_2$ using another OLS regression. The key here is that even though OLS standard errors are wrong, the estimates are still consistent. This means that we can use the estimated errors as consistent estimators of the actual standard errors.

FGLS is a straightforward process, but there are more steps:

1. Run the OLS Regression $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$ to get the estimated residuals $\hat{\varepsilon}_i$.

2. Run the regression $\hat{\varepsilon}_i^2 = g(x_{1i}, x_{2i})$, in our case we have

$$\hat{\varepsilon}_i^2 = \alpha_0 + \alpha_1 |x_{1i}| + \alpha_2 |x_{2i}| + v$$

The fitted values from this regression, call them $\hat{\varepsilon}$ are our estimated $\widehat{g(x_{1i}, x_{2i})}$ (up to a multiplicative constant, which doesn't matter).

3. We can now use a similar formula as before

$$\frac{y_i}{\sqrt{\widehat{g(x_{1i}, x_{2i})}}} = \frac{\beta_0}{\sqrt{\widehat{g(x_{1i}, x_{2i})}}} + \frac{\beta_1 x_{1i}}{\sqrt{\widehat{g(x_{1i}, x_{2i})}}} + \frac{\beta_2 x_{2i}}{\sqrt{\widehat{g(x_{1i}, x_{2i})}}} + \frac{\varepsilon_i}{\sqrt{\widehat{g(x_{1i}, x_{2i})}}}$$

Computing the actual variance of the new residual $\frac{\varepsilon_i}{\sqrt{\widehat{g(x_{1i}, x_{2i})}}}$ is tricky, but suffice to

say that in large samples

$$\text{Var}\left(\frac{\varepsilon_i}{\sqrt{\widehat{g(x_{1i}, x_{2i})}}}\right) = \text{Var}\left(\frac{\varepsilon_i}{\sqrt{g(x_{1i}, x_{2i})}}\right) = \frac{1}{g(x_{1i}, x_{2i})} \text{Var}(\varepsilon_i) = \sigma_\varepsilon^2$$

which is exactly the same as in GLS, as if we really knew the parameters of $g(x_{1i}, x_{2i})$!

How to check for Heteroskedasticity

Here are two ways to test for heteroskedasticity, and they use the same concepts as we use above: they test whether the squared estimated residuals are related to the x's.

The Breusch-Pagan Test

1. Run the OLS Regression $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$ to get the estimated residuals.

2. Run the OLS regression of the estimated residuals on the independent variables, that is,

$$\hat{\varepsilon}_i^2 = \alpha_0 + \alpha_1 x_{1i} + \alpha_2 x_{2i} + v$$

3. Construct and F-test of the joint hypothesis that $\alpha_1 = 0$ and $\alpha_2 = 0$, as you would in any other OLS situation.

The White Test

1. Run the OLS Regression $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$ to get the estimated residuals

2. Run the OLS regression of the estimated residuals on the independent variables *and* interactions of the independent variables, that is,

$$\hat{\varepsilon}_i^2 = \alpha_0 + \alpha_1 x_{1i} + \alpha_2 x_{2i} + \alpha_3 x_{1i} x_{2i} + v$$

3. Construct and F-test of the joint hypothesis that $\alpha_1 = 0$, $\alpha_2 = 0$ and $\alpha_3 = 0$, as you would in any other OLS situation.

Q: Which of these two tests is better?

SAS can automatically compute a variant of the White test using the SPEC option in the model statement. (It uses a Chi-squared LM-type test instead of an F-test to check for joint significance, but it doesn't really matter which one you use.)

Punting with White Standard Errors

Finally, recall that if we decide to "punt" and assume general heteroskedasticity just in case it's there, we simply use the formula

$$\widehat{Var}(\hat{\beta}) = \frac{\sum \tilde{x}_{1i} \hat{u}_i^2}{(s_{\tilde{x}_1}^2)^2}$$

I put \tilde{x} in there because you need to partial out other variables when doing a multivariate regression.

To do this numerically in SAS, you need to use the option ACOV in the MODEL

statement. To get the standard errors of each variable, take the square root of each diagonal element of the matrix that turns up in the output. Check out the output on the next page using a standard earnings-schooling regression with both the ACOV and SPEC options:

```
proc reg data=one;  
model rearnings=ed_comp exp / acov spec;
```

This was done with Kreuger's 1984 data from problem set 5.