Econ 452 Section 14 - STATA

Connor Cole

January 6, 2016

Responses to Heteroskedasticity, Continued

Response Two: Weighted Least Squares

Under heteroskedasticity, the BLUE estimator of β is a method that adjusts observations for the variances of the error term. Therefore, our regression estimates will have less variance if we weight observations by the variance of the error term called generalized least squares (GLS). Suppose our model is of the following form, where ϵ_i is a random independent variable with variance 1, and where $\sqrt{h_i}$ is a term that varies across observations i.

$$VariableName1 = \beta_0 + \beta_1 VariableName2 + \sqrt{h_i}\sigma \cdot \epsilon$$

Note that:

$$\begin{aligned} V(VariableName1|VariableName2) =& V(\beta_0 + \beta_1 VariableName2 + \sqrt{h_i}\sigma \cdot \epsilon|VariableName2) \\ =& V(\sqrt{h_i}\sigma \cdot \epsilon|VariableName2) \\ =& h_i\sigma^2 V(\epsilon|VariableName2) \\ =& h_i\sigma^2 \end{aligned}$$

Hence the $h_i\sigma^2$ values determine the way in which variances of the error term differ across observations. Now, consider reweighting the data by the square root of this variance. Then:

$$\frac{VariableName1}{\sqrt{h_i}\sigma} = \beta_0 \frac{1}{\sqrt{h_i}\sigma} + \beta_1 \frac{VariableName2}{\sqrt{h_i}\sigma} + \epsilon$$

Note that weighting by $\sqrt{h_i}\sigma$ returns the error term to a homoskedastic error, as the error term now has constant variance across all observations. Then, if we estimated a regression for this model using OLS with these reweighted data, we would then have estimates that are BLUE. GLS works then by reweighting the data by the standard deviation of the error term for each observation and then applying OLS to the new data.

The question then naturally becomes how we determine $\sigma^2 h_i$. When our data are averages and we have the number of observations in each average, our $\sigma^2 h_i$ is in fact just the number of observations in each group., q. Consider a data generating process where $VariableName1$ is a linear function of $VariableName2$ and an error term. Then:

$$\begin{aligned}
VariableName1 =& \beta_0 + \beta_1 VariableName2 + \epsilon \\
\sum_{i=1}^{q} VariableName1 =& \sum_{i=1}^{q} \beta_0 + \sum_{i=1}^{q} \beta_1 VariableName2 + \sum_{i=1}^{q} \epsilon \\
\frac{\sum_{i=1}^{q} VariableName1}{q} =& \frac{\sum_{i=1}^{q} \beta_0}{q} + \frac{\sum_{i=1}^{q} \beta_1 VariableName2}{q} + \frac{\sum_{i=1}^{q} \epsilon}{q}
\end{aligned}$$

If we only observe these averages across groups of size q (which is more common than you might think - e.g. average performance across schools). Suppose we are in a homoskedastic world in the original model. Then consider the variance of the error term here:

$$\begin{aligned}
V(\frac{\sum_{i=1}^{q} VariableName1}{q} | \frac{\sum_{i=1}^{q} \beta_1 VariableName2}{q}) =& V(\frac{\sum_{i=1}^{q} \epsilon}{q} | \frac{\sum_{i=1}^{q} \beta_1 VariableName2}{q}) \\
=& \frac{1}{q^2} V(\sum_{i=1}^{q} \epsilon | \frac{\sum_{i=1}^{q} \beta_1 VariableName2}{q}) \\
=& \frac{1}{q} V(\epsilon | \frac{\sum_{i=1}^{q} \beta_1 VariableName2}{q}) \\
=& \frac{\sigma_\epsilon^2}{q}
\end{aligned}$$

Note that the error term here is heteroskedastic across observations, as it depends on the number of observations in each group. Hence, estimating the relationship between means using OLS is inefficient. Note that if we reweight the data by $\sqrt{q}$, then we return to a homoskedastic error:

$$\begin{aligned}
\sqrt{q}\frac{\sum_{i=1}^{q} VariableName1}{q} =& \sqrt{q}\frac{\sum_{i=1}^{q} \beta_0}{q} + \sqrt{q}\frac{\sum_{i=1}^{q} \beta_1 VariableName2}{q} + \sqrt{q}\frac{\sum_{i=1}^{q} \epsilon}{q} \\
\frac{\sum_{i=1}^{q} VariableName1}{\sqrt{q}} =& \frac{\sum_{i=1}^{q} \beta_0}{\sqrt{q}} + \frac{\sum_{i=1}^{q} \beta_1 VariableName2}{\sqrt{q}} + \frac{\sum_{i=1}^{q} \epsilon}{\sqrt{q}}
\end{aligned}$$

Where:

$$\begin{aligned} V(\frac{\sum_{i=1}^{q} VariableName1}{\sqrt{q}}|\frac{\sum_{i=1}^{q} \beta_1 VariableName2}{\sqrt{q}}) =& V(\frac{\sum_{i=1}^{q} \epsilon}{\sqrt{q}}|\frac{\sum_{i=1}^{q} \beta_1 VariableName2}{\sqrt{q}}) \\ =& V(\epsilon|\frac{\sum_{i=1}^{q} \beta_1 VariableName2}{q}) \\ =& \sigma_\epsilon^2 \end{aligned}$$

Hence, when we have averages and homoskedastic errors in the underlying data process, reweighting the data by the square root of the number of observations in each group makes the estimation of the coefficients from the original model efficient.

The question becomes then what we do when we don't know this variance, which is usually the case. If ϵ is continuous, then one way of determining the variance of the error term would be to make a functional form assumption about the structure of $\sigma^2 h_i$. We will assume that the variance of the error term, $\sigma^2 h_i$ is exactly:

$$\begin{aligned} V(\psi_i|VariableName2) = \sigma^2 h_i =& \sigma^2 e^{\gamma_0 + \gamma_1 VariableName2} \delta_i \\ \ln(V(\psi_i|VariableName2)) =& \ln(\sigma^2) + \gamma_0 + \gamma_1 VariableName2 + \ln(\delta_i) \\ \ln(V(\psi_i|VariableName2)) =& (\ln(\sigma^2) + \gamma_0 + \mathbb{E}[\ln(\delta_i)]) + \gamma_1 VariableName2 + \ln(\delta_i) - \mathbb{E}[\ln(\delta_i)] \\ \ln(V(\psi_i|VariableName2)) =& \gamma_0^{New} + \gamma_1 VariableName2 + \epsilon_i \end{aligned}$$

Where $\epsilon_i = \ln(\delta_i) - \mathbb{E}[\ln(\delta_i)]$ and $\gamma_0^{New} = \ln(\sigma^2) + \gamma_0 + \mathbb{E}[\ln(\delta_i)]$.

Note that we can consistently estimate this relationship, assuming that $\mathbb{E}[\epsilon_i|VariableName2] = 0$, by taking the squared residuals from a regression and regressing the log of those squared residuals on the covariates X. Then, if we take the predicted values of these regressions and then put them in an exponent we have a consistent estimator of the expected value of $\sigma^2 h_i$. Finally, then we use these values and reweight our data for feasible generalized least squares (FGLS) estimate. Note that this estimate, since we need to estimate the relationship and do not have the true variances as in GLS, will likely be biased but, as noted before, it will be consistent and have asymptotically lower variance than OLS. Thus, FGLS is 'more efficient' than OLS.

Thus, the process for implementing FGLS in a regression of $VariableName1$ on $VariableName2$ would be:

1. Regress $VariableName1$ on $VariableName2$
2. Take residuals from the regression, square them, and lastly take a log of them. Save the new values as *lnsqresid*.
3. Regress the log squared residuals in *lnsqresid* on $VariableName2$

4. Pull the predicted values from the last regression in step 3, and take an exponent of them: $e^{\hat{y}}$
5. Take a reciprocal of the values created in step 4 and save as a new variable as *inversevarhat*.
6. Regress $VariableName1$ on $VariableName2$ using *inversevarhat* created in step 5 as an inverse variance weight.

Let's say we wanted to estimate a regression of $VariableName1$ on $VariableName2$ using FGLS in STATA:

```
regress VariableName1 VariableName2

predict residuals, resid

generate residualssq =residuals*residuals

generate lnresidualssq=log(residualssq)

regress lnresidualssq VariableName2

predict lnresidualssq, xb

generate variancehat=exp(lnresidualssq)

generate inversevarhat=1/variancehat

regress VariableName1 VariableName2 [aweight= inversevarhat]
```

The `aweight` option performs the weighting operation we did above, where we divide all components of the regression by the square root of the estimated variance. Here's how we would do the same regression long-hand without using the `aweight` option:

```
regress VariableName1 VariableName2

predict residuals, resid

generate residualssq =residuals*residuals

generate lnresidualssq=log(residualssq)

regress lnresidualssq VariableName2

predict lnresidualssq, xb

generate variancehat=exp(lnresidualssq)

generate inversevarhat=1/variancehat
```

```
generate inversevarhatsqrt=(1/variancehat)^1/2

generate VariableName1new=VariableName1*inversevarhatsqrt

generate VariableName2new=VariableName2*inversevarhatsqrt

generate constantnew = 1*inversevarhatsqrt

regress VariableName1new VariableName2new constantnew, noconstant
```

Note that we need to create a new constant term for our regression and specify the `noconstant` option.

Now, suppose that our data come from a linear probability model where $VariableName1$ is equal to either 1 or 0 and $VariableName1 = \beta_0 + \beta_1 VariableName2 + \epsilon$. Suppose furthermore that $VariableName2$ is at some value. Note that if $VariableName1 = 1$ then ϵ takes on value $1 - (\beta_0 + \beta_1 VariableName2)$ and if $VariableName1 = 0$ then $\epsilon = -\beta_0 - \beta_1 VariableName2$. Note then for any given value of $VariableName2$ then ϵ can only take on two values. Hence, our previous estimation procedure, where we essentially assumed that the error term was from a continuous distribution, doesn't work.

However, we can derive the variance of $VariableName1$ by using facts about the linear probability model. First, note that $VariableName1$ is a discrete random variable that takes on values 0 or 1. Hence:

$$\begin{aligned}\mathbb{E}[VariableName1|VariableName2] \\ &= 0 \cdot P(VariableName1 = 0|VariableName2) + 1 \cdot P(VariableName1 = 1|VariableName2) \\ &= P(VariableName1 = 1|VariableName2)\end{aligned}$$

And note furthermore that $\mathbb{E}[VariableName1|VariableName2] = \beta_0 + \beta_1 VariableName2$. Therefore:

$$P(VariableName1 = 1|VariableName2) = \beta_0 + \beta_1 VariableName2$$

ϵ takes on value $1-(\beta_0+\beta_1 VariableName2)$ when $VariableName1 = 1$, and hence takes on this value with the probability that $VariableName1 = 1$, and ϵ takes on value $-\beta_0-\beta_1 VariableName2$ when $VariableName1 = 0$ and hence takes on this value with the probability that $VariableName1 = 0$. Therefore:

$$
\begin{aligned}
&V(\epsilon|VariableName2) = \mathbb{E}[\epsilon^2|VariableName2] - \mathbb{E}[\epsilon|VariableName2]^2 \\
&=(\epsilon_1)^2 \cdot P(VariableName1 = 1|VariableName2) + (\epsilon_0)^2 \cdot P(VariableName1 = 0|VariableName2) - \\
&\quad \big(\epsilon_1 \cdot P(VariableName1 = 1|VariableName2) + \epsilon_0 \cdot P(VariableName1 = 0|VariableName2)\big) \\
&=(1 - (\beta_0 + \beta_1 VariableName2))^2 \cdot (\beta_0 + \beta_1 VariableName2) + \\
&\quad (-\beta_0 - \beta_1 VariableName2)^2 \cdot (1 - (\beta_0 + \beta_1 VariableName2)) - \\
&\quad ((1 - (\beta_0 + \beta_1 VariableName2)) \cdot (\beta_0 + \beta_1 VariableName2) + \\
&\quad (-\beta_0 - \beta_1 VariableName2)(((1 - (\beta_0 + \beta_1 VariableName2)))) \\
&=(1 - \beta_0 - \beta_1 VariableName2)(\beta_0 + \beta_1 VariableName2) \\
&=P(VariableName1 = 1|VariableName2) \cdot (1 - P(VariableName1 = 1|VariableName2))
\end{aligned}
$$

For an example, assume that $\beta_0 + \beta_1 VariableName2 = .3$. Then:

$$
\begin{aligned}
V[\epsilon|VariableName2] &= \mathbb{E}[\epsilon^2|VariableName2] - \mathbb{E}[\epsilon|VariableName2]^2 \\
&= 0.3 \cdot (0.7)^2 + 0.7 \cdot (-0.3)^2 - (0.3 \cdot 0.7 - 0.7 \cdot 0.3) \\
&= 0.3 \cdot (0.7)^2 + 0.7 \cdot (-0.3)^2 = .21
\end{aligned}
$$

Note then that a natural estimator for the variance of the error term in the linear probability model case would be:

$$
\begin{aligned}
&V[\epsilon|VariableName2] \\
&\quad =P(VariableName1 = 1|VariableName2) \cdot (1 - P(VariableName1 = 1|VariableName2)) \\
&\quad \approx \hat{P}(VariableName1 = 1|VariableName2) \cdot (1 - \hat{P}(VariableName1 = 1|VariableName2)) \\
&\quad \approx \hat{P}(VariableName1 = 1|VariableName2) \cdot (1 - \hat{P}(VariableName1 = 1|VariableName2)) \\
&\quad \approx \widehat{VariableName1} \cdot (1 - \widehat{VariableName1})
\end{aligned}
$$

This fact then suggests a similar procedure to what we did above for implementing FGLS with data from a linear probability model where $VariableName1$ is our dependent variable and $VariableName2$ is our independent variable.

1. Regress $VariableName1$ on $VariableName2$.
2. Take predicted values from the regression and save in a new variable *probabilityhat*.
3. Create a new variable *varhat* by multiplying *probabilityhat* by $1 -$ *probabilityhat*.
4. Take a reciprocal of the values created in step 3 and save as a new variable as *inversevarhat*.

5. Regress $VariableName1$ on $VariableName2$ using *inversevarhat* created in step 5 as an inverse variance weight.

In STATA, we would implement this procedure as follows

```
regress VariableName1 VariableName2

predict probabilityhat, xb

generate varhat = (probabilityhat)*(1-probabilityhat

generate inversevarhat=1/varhat

regress VariableName1 VariableName2 [aweight=inversevarhat]
```

Specification Tests

While we have generally assumed that we have specified the 'correct' model, it may be that that we have used the wrong model, implying that the current model we assume is misspecified. We will consider two possible tests, the Ramsey RESET test and the Davidson-McKinnen test.

Ramsey RESET Test

The Ramsey RESET test works by including powers of the fitted predicted values from a regression as covariates. If we have 'correctly' specified the regression, then including the predicted values from the 'correctly' specified regression squared or taken to the third power should not matter. Then, if we include these fitted values taken to various powers in the true model, we can use an F-test to test the null hypothesis that the coefficients on these fitted values should be 0. Note that this test is flexible in the alternate specifications it might possibly describe, but rejecting the null hypothesis here of 'correct' specification does not point the way to a different specification as there is no clear alternative hypothesis implied by this form of a test. Hence, it is not clear what the power of this test is.

Suppose we are regressing $VariableName1$ on $VariableName2$. Then we would implement this test by typing:

```
regress VariableName1 VariableName2

predict predicted1, xb

generate predicted2=predicted^2

generate predicted3=predicted^3
```

```
regress VariableName1 VariableName2 predicted2 predicted3

test (predicted2=0) (predicted3=0)
```

Note that STATA's hard coded version of this test `estat ovtest` uses four powers of the predicted values instead of three.

Davidson-McKinnen Test

As noted previously, the RESET test does not test against a definitive alternative hypothesis and if a rejection occurs, simply implies that the model is missing something - either omitted variables, missing squares of variables, different functions of the variables, etc.. The Davidson-McKinnen test offers more actionable results because it compares specific models with each other. Let's say we are considering regressions that relate $VariableName1$ to $VariableName2$ and we want to compare the two models to each other:

$$\begin{aligned} VariableName1_{Model1} &= \beta_0 + \beta_1 VariableName2 + \epsilon \\ VariableName1_{Model2} &= \eta_0 + \eta_1 \ln(VariableName2) + \epsilon \end{aligned}$$

We estimate these two models separately and produce fitted values of $\widehat{VariableName1}$. We then add in the fitted values from each model into the regressions we estimate for the other model:

$$\begin{aligned} VariableName1_{Model1} &= \beta_0 + \beta_1 VariableName2 + \beta_2 \widehat{VariableName1}_{Model2} + \epsilon \\ VariableName1_{Model2} &= \eta_0 + \eta_1 \ln(VariableName2) + \eta_2 \widehat{VariableName1}_{Model1} + \epsilon \end{aligned}$$

We then test separately the null hypotheses that $\beta_2 = 0$ and $\eta_2 = 0$. There are four possible outcomes:

1. Fail to reject $\beta_2 = 0$, reject $\eta_2 = 0$: The second model seems to fit better than the first.
2. Reject $\beta_2 = 0$, fail to reject $\eta_2 = 0$: The first model seems to fit better than the first.
3. Reject both $\beta_2 = 0$ and $\eta_2 = 0$: Both models seem to perform badly and we should try other models.
4. Fail to reject both $\beta_2 = 0$ and $\eta_2 = 0$: Both models seem to work well, and we should use some other criteria for comparing the two models to each other

In STATA, we would implement this test in regressing $VariableName1$ on $VariableName2$ as follows;

```
regress VariableName1 VariableName2

predict Model1Predicted, xb

generate lnVariableName2 = log(VariableName2)

regress VariableName1 lnVariableName2

predict Model2Predicted, xb

regress VariableName1 VariableName2 Model2Predicted

test (Model2Predicted=0)

regress VariableName1 lnVariableName2 Model1Predicted

test (Model1Predicted=0)
```

Measurement Error

We usually cannot detect measurement error in a dataset by itself, although we can measure how it is expressed in our data by looking at 'validation' studies where we compare results from administrative datasets with survey data. Measurement error can either occur in the dependent variable or in the independent variable.

Error in the Dependent Variable

Suppose our true model is:

$$VariableName1^* = \beta_0 + \beta_1 VariableName2^* + \epsilon$$

But suppose we only observe a measurement of $VariableName1^*$ with error: $VariableName1 = VariableName1^* + \eta$ where the η error term is *iid*. Note if we regress $VariableName1$ on $VariableName2^*$, then the *plim* of $\hat{\beta}_1$ would be:

$$
\begin{aligned}
plim(\beta_1) &= \frac{cov(VariableName1, VariableName2^*)}{V(VariableName2^*)} \\
&= \frac{cov(VariableName1^* + \eta, VariableName2^*)}{V(VariableName2^*)} \\
&= \frac{cov(VariableName1^*, VariableName2^*)}{V(VariableName2^*)} \\
&= \beta_1
\end{aligned}
$$

And

$$
V(\beta_1) = \frac{\sigma_\eta^2 + \sigma_\epsilon^2}{\sum_{i=1}^{n}(x_i^* - \bar{x}^*)^2}
$$

Hence, $\hat{\beta}$ using the error-laden variable is unbiased and consistent, but the error just increases the variance of the error term.

Error in the Independent Variable

Suppose now that we are regression the exact value of $VariableName1^*$ on $VariableName2$ where $VariableName2 = VariableName2^* + \eta$ and where the error term is *iid*. Now, if we regress $VariableName1^*$ on $VariableName2$ then the*plim* of $\hat{\beta}_1$ would be:

$$
\begin{aligned}
plim(\hat{\beta}_1) &= \frac{cov(VariableName1^*, VariableName2)}{V(VariableName2)} \\
&= \beta_1 \frac{V(VariableName2^*)}{V(VariableName2^*) + V(\eta)}
\end{aligned}
$$

$$
\begin{aligned}
plim(\beta_1) &= \frac{cov(VariableName1^*, VariableName2)}{V(VariableName2)} \\
&= \frac{\beta_1 V(VariableName2^*)}{V(VariableName2^*) + V(\eta)} \\
&= \beta_1 \frac{V(VariableName2^*)}{V(VariableName2^*) + V(\eta)}
\end{aligned}
$$

Note that the numerator is necessarily less than the denominator so the probability limit is necessarily less than the true value of β_1.