Topic 7 Continued: Heteroskedasticity

Goldfeld-Quandt Test

- Suppose that we have two samples of data. That is, we have sampled from two *potentially different* populations.
- We want to test if the variance of the error term for our regression model is the same for both populations.
- We'll assume that we know that the coefficient vector *is the same* for both populations.
- So:

$$\boldsymbol{y}_1 = X_1\boldsymbol{\beta} + \boldsymbol{\varepsilon}_1 \quad ; \quad \boldsymbol{\varepsilon}_1 \sim N[0\,, \sigma_1^2 I_{n_1}]$$

$$\boldsymbol{y}_2 = X_2\boldsymbol{\beta} + \boldsymbol{\varepsilon}_2 \quad ; \quad \boldsymbol{\varepsilon}_2 \sim N[0\,, \sigma_2^2 I_{n_2}]$$

(*Subscripts denote samples*)

- We want to test $H_0: \sigma_1^2 = \sigma_2^2$ *vs.* $H_A: \sigma_1^2 > \sigma_2^2$ (*say*)

The Goldfeld-Quandt test for homoscedasticity is constructed as follows:

1. Fit the model, using OLS, over each of the two samples, *separately*.
2. Let the two residual vectors be $\boldsymbol{e_1}$ and $\boldsymbol{e_2}$.
3. If the errors are Normally distributed, then the statistics:

$$(e_i{}'e_i)/(\sigma_i^2) \sim \chi^2_{(n_i - k)} \quad ; \quad i = 1, 2.$$

4. The two regressions are fitted quite separately, so these two statistics are *statistically independent*.
5. Consider the statistic:

$$F = (e_1{}'e_1)/(\sigma_1^2(n_1 - k))/(e_2{}'e_2)/(\sigma_2^2(n_2 - k))$$

6. If $H_0: \sigma_1^2 = \sigma_2^2$ is true, then $F = \left(\frac{s_1^2}{s_2^2}\right) \sim F_{(n_1 - k\,;\, n_2 - k)}$.
7. We would **reject $\boldsymbol{H_0}$** if $F > c(\alpha)$.

- If we *do not reject* H_0, then we would estimate the (common) coefficient vector, $\boldsymbol{\beta}$, by "pooling" both samples together, and applying OLS.
- On the other hand, if we reject H_0, then we would estimate the (common) coefficient vector, $\boldsymbol{\beta}$, by GLS.

- Let's see what form the latter estimator takes in this particular case.
- Recall that we have:

$$\boldsymbol{y}_1 = X_1\boldsymbol{\beta} + \boldsymbol{\varepsilon}_1 \quad ; \quad \boldsymbol{\varepsilon}_1 \sim N[0\,,\sigma_1^2 I_{n_1}] \qquad (n_1)$$

$$\boldsymbol{y}_2 = X_2\boldsymbol{\beta} + \boldsymbol{\varepsilon}_2 \quad ; \quad \boldsymbol{\varepsilon}_2 \sim N[0\,,\sigma_2^2 I_{n_2}] \qquad (n_2)$$

  - Let $\phi = (\sigma_1/\sigma_2)$; and let $\hat{\phi} = (s_1/s_2)$;

  where $s_i^2 = (\boldsymbol{e}_i{}'\boldsymbol{e}_i)/(n_i - k)$; i = 1, 2.

  - Note that $\hat{\phi}$ is a *consistent* estimator of ϕ.
  - If we knew the value of ϕ, we could use it to scale the model for the second sub-sample, as follows:

$$\phi\boldsymbol{y}_2 = \phi X_2\boldsymbol{\beta} + \phi\boldsymbol{\varepsilon}_2 \qquad (n_2)$$

  where $E[\phi\boldsymbol{\varepsilon}_2] = 0$ and

$$V[\phi\boldsymbol{\varepsilon}_2] = \phi^2 V[\boldsymbol{\varepsilon}_2] = \left(\frac{\sigma_1^2}{\sigma_2^2}\right)\sigma_2^2 I_{n_2} = \sigma_1^2 I_{n_2}$$

  - That is, the full error vector, $\boldsymbol{\varepsilon}' = (\boldsymbol{\varepsilon}_1', \phi\boldsymbol{\varepsilon}_2{}')'$, is *homoscedastic*.
  - GLS estimation then amounts to applying OLS to the "pooled" data, but where the data associated with the second sub-sample have been transformed in the above way.
  - Typically, we won't know the value of $\phi = (\sigma_1/\sigma_2)$, but we can use $\hat{\phi} = (s_1/s_2)$ instead to implement *feasible* GLS estimation.
  - Because $\hat{\phi}$ is a consistent estimator of ϕ, this feasible GLS estimator will be consistent for $\boldsymbol{\beta}$.

Example

- Investment data for 2 companies – General Electric & Westinghouse
- 20 years of annual data for each company – 1935 to 1954
- I = Gross investment, in 1947 dollars
- V = Market value of company as of 31 December, in 1947 dollars
- K = Stock of plant & equipment, in 1947 dollars
- "Pool" the data – first 20 observations are for General Electric; second 20 observations are for Westinghouse

First, take a look at the data:

```
fglsdata=read.csv("http://home.cc.umanitoba.ca/~godwinrt/7010/fgls.csv")
attach(fglsdata)
fglsdata
```

```
   Year   Ige    Vge   Kge    Iw     Vw    Kw
1  1935  33.1 1170.6  97.8 12.93  191.5   1.8
2  1936  45.0 2015.8 104.4 25.90  516.0   0.8
3  1937  77.2 2803.3 118.0 35.05  729.0   7.4
4  1938  44.6 2039.7 156.2 22.89  560.4  18.1
5  1939  48.1 2256.2 172.6 18.84  519.9  23.5
6  1940  74.4 2132.2 186.6 28.57  628.5  26.5
7  1941 113.0 1834.1 220.9 48.51  537.1  36.2
8  1942  91.9 1588.0 287.8 43.34  561.2  60.8
9  1943  61.3 1749.4 319.9 37.02  617.2  84.4
10 1944  56.8 1687.2 321.3 37.81  626.7  91.2
11 1945  93.6 2007.7 319.6 39.27  737.2  92.4
12 1946 159.9 2208.3 346.0 53.46  760.5  86.0
13 1947 147.2 1656.7 456.4 55.56  581.4 111.1
14 1947 146.3 1604.4 543.4 49.56  662.3 130.6
15 1949  98.3 1431.8 618.3 32.04  583.8 141.8
16 1950  93.5 1610.5 647.4 32.24  635.2 136.7
17 1951 135.2 1819.4 671.3 54.38  723.8 129.7
18 1952 157.3 2079.7 726.1 71.78  864.1 145.5
19 1953 179.5 2371.6 800.3 90.08 1193.5 174.8
20 1954 189.6 2759.9 888.9 68.60 1188.9 213.5
```

Estimate the "pooled" regression:

```
I = c(Ige,Iw)
V = c(Vge,Vw)
```

```
K = c(Kge,Kw)
res = lm(I ~ V + K)
summary(res)

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 17.872001   7.024081   2.544   0.0153 *
V            0.015193   0.006196   2.452   0.0191 *
K            0.143579   0.018601   7.719 3.19e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 21.16 on 37 degrees of freedom
Multiple R-squared: 0.8098,     Adjusted R-squared: 0.7995
F-statistic: 78.75 on 2 and 37 DF,  p-value: 4.641e-14
```

Perform White's Heteroskedasticity test:

```
resids2 = res$residuals^2
V2 = V^2
K2 = K^2
VK = V*K
summary(lm(resids2 ~ V + K + V2 + K2 + VK))

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.643e+02  4.553e+02  -0.361   0.7204
V           -1.591e-01  1.053e+00  -0.151   0.8808
K            5.238e+00  2.592e+00   2.021   0.0512 .
V2           6.041e-06  3.413e-04   0.018   0.9860
K2          -8.899e-03  3.860e-03  -2.305   0.0274 *
VK           1.233e-03  1.381e-03   0.893   0.3781
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 586.8 on 34 degrees of freedom
Multiple R-squared: 0.337,      Adjusted R-squared: 0.2395
F-statistic: 3.457 on 5 and 34 DF,  p-value: 0.01242

1 - pchisq(40*0.337,5)
0.01927276
```

Now let's try the Goldfeld-Quandt Test:

```
resGE = lm(Ige ~ Vge + Kge)

summary(resGE)


Coefficients:
```

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -9.95631   31.37425  -0.317    0.755
Vge          0.02655    0.01557   1.706    0.106
Kge          0.15169    0.02570   5.902 1.74e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 27.88 on 17 degrees of freedom
Multiple R-squared: 0.7053,     Adjusted R-squared: 0.6706
F-statistic: 20.34 on 2 and 17 DF,  p-value: 3.088e-05
```

$$\frac{e_1{}'e_1}{n_1 - k} = 27.88^2 = 777.45$$

```
resW = lm(Iw ~ Vw + Kw)
summary(resW)
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.50939    8.01529  -0.064  0.95007
Vw           0.05289    0.01571   3.368  0.00365 **
Kw           0.09241    0.05610   1.647  0.11787
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10.21 on 17 degrees of freedom
Multiple R-squared: 0.7444,     Adjusted R-squared: 0.7144
F-statistic: 24.76 on 2 and 17 DF,  p-value: 9.196e-06
```

$$\frac{e_2{}'e_2}{n_2 - k} = 10.21^2 = 104.31$$

In this example, there is more variability in the error term over the first sub-sample (General Electric) than there is over the second sub-sample (Westinghouse): $s_1^2 = 777.45$; $s_2^2 = 104.31$

- $H_0: \sigma_1^2 = \sigma_2^2$ *vs.* $H_A: \sigma_1^2 > \sigma_2^2$
- $F = (777.45/104.31) = 7.45$
- If H_0 is true, $F \sim F_{(n_1-k\,;\,n_2-k)} = F_{(17\,;\,17)}$
- 5% critical value = 2.4 ; 1% critical value = 3.5
- `1 - pf(7.45,17,17)`
  `7.172914e-05`
- **Reject H_0**
- So, leave the data for the *first sub-sample unchanged*, but multiply the data (including the intercept) for the *second sub-sample* by $\hat{\phi} = \frac{s_1}{s_2} = \frac{27.88}{10.21} = 2.73$
- This means that instead of using a constant term in our regression, we must create a vector that consists of 20 1's, followed by 20 values of 2.73 (Cstar), and use this vector as the first term in our regression.

```
Istar = c(Ige, 2.73 * Iw)
Cstar = c(rep(1,20), rep(2.73,20))
Vstar = c(Vge, 2.73 * Vw)
Kstar = c(Kge, 2.73 * Kw)
summary(lm(Istar ~ Cstar + Vstar + Kstar -1)
```

```
Coefficients:
       Estimate Std. Error t value Pr(>|t|)
Cstar 16.747017   4.785409   3.500  0.00123 **
Vstar  0.020391   0.007245   2.814  0.00778 **
Kstar  0.133713   0.024144   5.538 2.65e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 29.74 on 37 degrees of freedom
Multiple R-squared: 0.9436,     Adjusted R-squared: 0.939
F-statistic: 206.3 on 3 and 37 DF,  p-value: < 2.2e-16
```