Topic 7: Heteroskedasticity

Consider the linear regression model

$$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad ; \qquad \boldsymbol{\varepsilon} \sim N[\mathbf{0}, \sigma^2\Omega]$$

where

$$\sigma^2\Omega = \sigma^2 \begin{bmatrix} \omega_{11} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \ldots & \omega_{nn} \end{bmatrix} = \begin{bmatrix} \sigma_1^2 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \ldots & \sigma_n^2 \end{bmatrix} = diag.(\sigma_i^2)$$

Then the errors exhibit Heteroskedasticity, but they are still uncorrelated.

- We know, from Topic 6, that in this case the OLS estimator of $\boldsymbol{\beta}$ is unbiased and consistent, but it is *inefficient*.
- We know that we can use White's modified estimator for the covariance matrix of $\boldsymbol{\beta}$ to ensure that the standard errors of the b_i's are *consistent* estimators for the true s.e.($\boldsymbol{b_i}$)'s.
- We also know that we can use GLS to obtain the BLU estimator of $\boldsymbol{\beta}$ if Ω *is known*.
- If $$\Omega = \begin{bmatrix} \omega_{11} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \ldots & \omega_{nn} \end{bmatrix},$$

  then $$P = \begin{bmatrix} \omega_{11}^{-1/2} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \ldots & \omega_{nn}^{-1/2} \end{bmatrix},$$

  so that $$P'P = \Omega^{-1}$$

- So, in this particular case, GLS estimation involves transforming the data:

  $$\boldsymbol{y}^* = P\boldsymbol{y} \quad ; \quad X^* = PX$$

- Just multiply the model by the matrix, P, or simply scale the i^{th} observation of all variables by $\omega_{ii}^{-1/2}$:

  $$\omega_{ii}^{-1/2} y_i = \beta_1 \omega_{ii}^{-1/2} + \beta_2 \left(\omega_{ii}^{-\frac{1}{2}} x_{i2} \right) + \cdots + \left(\omega_{ii}^{-\frac{1}{2}} \varepsilon_i \right)$$

- This particular variant of GLS is often referred to as "**Weighted Least Squares**" estimation. It is just OLS applied using "weighted" data.

Example:

$$y_i = \beta_1 + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i$$

$$E[\varepsilon_i] = 0 \quad ; \quad var.[\varepsilon_i] \propto x_{i2}^2$$

So, we can write:

$$var.[\varepsilon_i] = \sigma^2 x_{i2}^2 \quad ; \quad \omega_{ii} = x_{i2}^2 \quad ; \quad {\omega_{ii}}^{-1/2} = 1/x_{i2}$$

$$(y_i/x_{i2}) = \beta_1(1/x_{i2}) + \beta_2 + \cdots + \beta_k(x_{ik}/x_{i2}) + {\varepsilon_i}^*$$

where $\quad {\varepsilon_i}^* = (\frac{\varepsilon_i}{x_{i2}}) \quad ; \quad E[{\varepsilon_i}^*] = 0 \quad$ (assumption?)

$$var.[{\varepsilon_i}^*] = (1/x_{i2})^2 var.[\varepsilon_i] = (1/x_{i2})^2\, \sigma^2 x_{i2}^2 = \sigma^2$$

Example:

$$y_i = \beta_1 + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i$$

$$E[\varepsilon_i] = 0 \quad ; \quad var.[\varepsilon_i] \propto z_i^p$$

$$var.[\varepsilon_i] = \sigma^2 z_i^p \quad ; \quad \omega_{ii} = z_i^p \quad ; \quad {\omega_{ii}}^{-1/2} = z_i^{-p/2}$$

$$(y_i z_i^{-p/2}) = \beta_1(z_i^{-p/2}) + \beta_2(x_{i2} z_i^{-p/2}) + \cdots + \beta_k(x_{ik} z_i^{-p/2}) + {\varepsilon_i}^*$$

where $\quad {\varepsilon_i}^* = (\varepsilon_i z_i^{-p/2}) \quad ; \quad E[{\varepsilon_i}^*] = 0 \quad$ (assumption?)

$$var.[{\varepsilon_i}^*] = \left(z_i^{-p/2}\right)^2 var.[\varepsilon_i] = z_i^{-p} \sigma^2 z_i^p = \sigma^2$$

Note that in this case we end up with a fitted model with no intercept, but we are still estimating the original parameters of interest.

- In some cases we will actually ***know*** the form of the heteroskedasticity, so we can apply WLS directly.

Example:

$$y_i = \beta_1 + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i$$

$$E[\varepsilon_i] = 0 \quad ; \quad var.[\varepsilon_i] = \sigma^2 \quad ; \quad i.i.d$$

However, suppose that we only observe "grouped" data, rather than the observations on the individual agents.

This happens frequently in practice, when data are released in this way to preserve confidentiality.

Suppose there are m groups (*e.g.*, income groups), with n_j observations in the j^{th} group; $j = 1, 2, \ldots, m$.

The model we can *actually estimate* is of the form:

$$\bar{y}_j = \beta_1 + \beta_2 \bar{x}_{j2} + \cdots + \beta_k \bar{x}_{jk} + \bar{\varepsilon}_j \quad ; \quad j = 1, 2, ..., m$$

and clearly,

$$E[\bar{\varepsilon}_j] = E\left[\frac{1}{n_j}\sum_{i=1}^{n_j} \varepsilon_i\right] = \left[\frac{1}{n_j}\sum_{i=1}^{n_j} E(\varepsilon_i)\right] = 0$$

$$var.[\bar{\varepsilon}_i] = var.\left[\frac{1}{n_j}\sum_{i=1}^{n_j} \varepsilon_i\right] = \left[\frac{1}{{n_j}^2}\sum_{i=1}^{n_j} var.(\varepsilon_i)\right]$$

$$= \left(n_j \sigma^2 / n_j^2\right)$$

$$= \left(\sigma^2 / n_j\right) \ .$$

The n_j values are generally reported, so we know the error covariance matrix:

$$\sigma^2 \Omega = \sigma^2 \begin{bmatrix} 1/n_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \ldots & 1/n_m \end{bmatrix}.$$

Because Ω is *known*, we can compute the GLS estimator of the coefficient vector immediately:

$$\hat{\beta} = [X'\Omega^{-1}X]^{-1}X'\Omega^{-1}\boldsymbol{y} .$$

However, in many other applications, we *won't know* the values of the elements of Ω , and we'll have to use Feasible GLS estimation.

FGLS Example:

- Estimate β by OLS (b is at least consistent)
- Obtain the OLS residuals, $\boldsymbol{e}$
- Estimate Ω by: $\widehat{\Omega}_{OLS} = diag(e_1^2, \dots, e_n^2)$
- Estimate $\hat{\beta}_{FGLS1} = \left[X'\widehat{\Omega}_{OLS}^{-1}X\right]^{-1}X'\widehat{\Omega}_{OLS}^{-1}\boldsymbol{y}$

The procedure can be iterated, until estimation of $\widehat{\Omega}$ converges. Note that the benefit of iterating is questionable, as each estimator for β past the first iteration is consistent.

Testing for Homoskedasticity

- Clearly, it would be very useful to have a test of the hypothesis that the errors in our regression model are *homoscedastic*, against the alternative that they exhibit some sort of *heteroskedasticity*.
- Recall that heteroskedasticity reduces the efficiency of the OLS estimator of $\boldsymbol{\beta}$ and has serious implications for the properties of the associated standard errors, confidence intervals, and tests.
- Because OLS is still a *consistent* estimator of $\boldsymbol{\beta}$ even if the errors are heteroskedastic, this means that we can use the OLS residuals to construct tests that will still be (at least) asymptotically valid.
- In particular, we can use these residuals to construct asymptotically valid tests for homoskedasticity.

White's Test

$$y_i = \beta_1 + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i$$

$$E[\varepsilon_i] = 0 \quad ; \quad var.[\varepsilon_i] = \sigma_i^2 \quad ; \quad i.i.d$$

Consider the following null and alternative hypotheses:

$$H_0: \sigma_i^2 = \sigma^2 \quad ; \quad i = 1, 2, \ldots, n \qquad\qquad H_A: Not\ H_0$$

- The Alternative Hypothesis is very general.
- No specific form of heteroskedasticity is declared.
- To implement the test –

1. Estimate the model by OLS, and get the residuals, e_i ; $i = 1, 2, \ldots, n$.
2. Using OLS, regress the e_i^2 values on each of the x's in the original model; their squared values; all of the cross-products of the regressors; and an intercept.
3. nR^2 from the regression in Step 2 is *asymptotically* $\chi^2_{(p)}$ if H_0 is true; where p is the number of parameters that are estimated at Step 2.
4. Reject H_0 in favour of H_A if $nR^2 > c(\alpha)$.

- Note the *limitations* of this test:
  1. It is valid only asymptotically.
  2. The test is "non-constructive", in the sense that if we reject H_0, we don't know what form of heteroskedasticity we may have.
  3. This means that it won't be clear what form the GLS estimator should take.
- However, this may be enough information to alert us to the fact that we should probably use White's "heteroskedasticity-consistent" estimator of $V(\boldsymbol{b})$.
- In fact, there is little, if anything, to be lost in using this covariance matrix estimator, anyway, as long as the sample is large.

Example

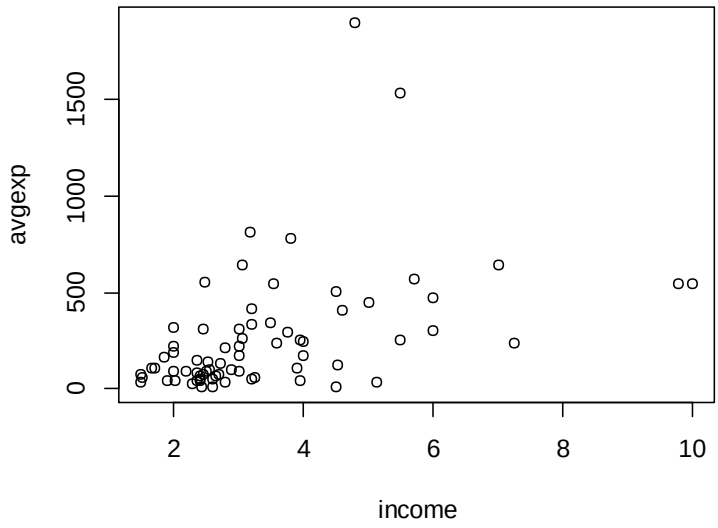
Data is on average monthly credit card expenditure (*avgexp*). The explanatory variables are *age*, *ownrent* (= 1 if homeowner, = 0 if renter), and *income* (in $10,000). Produce a scatter plot of *avgexp* against *income*.

```
ccard=read.csv("http://home.cc.umanitoba.ca/~godwinrt/7010/credi
  tcard.csv")

attach(ccard)

plot(income,avgexp)
```

Does it look like heteroskedasticity is apparent?



Estimate the following model by OLS:

$$avgexp = \beta_1 + \beta_2 age + \beta_3 ownrent + \beta_4 income + \beta_5 income^2 + \varepsilon$$

```
income2 = income^2
res = lm(avgexp ~ age + ownrent + income + income2)
summary(res)
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -237.147    199.352  -1.190  0.23841
age           -3.082      5.515  -0.559  0.57814
ownrent       27.941     82.922   0.337  0.73721
income       234.347     80.366   2.916  0.00482 **
income2      -14.997      7.469  -2.008  0.04870 *
```

White's heteroskedasticity consistent standard errors can be calculated using standard econometric software (e.g. Eviews, Stata). However, we can easily write R code to estimate the appropriate variance-covariance matrix.

Recall that in the presence of heteroskedasticity, White's estimator for the *var-cov* matrix of b is:

$$\hat{V}^* = n[(X'X)^{-1}S_0(X'X)^{-1}]$$

where

$$S_0 = \frac{1}{n}\sum_{i=1}^{n} e_i^2 \boldsymbol{x}_i \boldsymbol{x}_i'$$

To code this into R:

```
resids2 = res$residuals^2
```

Get the squared resids. from 1st regression

```
n = length(avgexp)
```

Read the sample size from the data

```
X = matrix(c(rep(1,n),age,ownrent,income,income2),n,5)
```

Create *X* matrix

```
S = matrix(0,5,5)
```

Create "empty" *S* matrix

```
for(i in 1:n){
S = S + (resids2[i]) * X[i,] %*% t(X[i,])
}
```

This is a "for" loop. In each iteration, $e_i^2 \boldsymbol{x}_i \boldsymbol{x}_i'$ will be added to the *S* matrix.

```
S = S/n
diag((n*solve(t(X) %*% X) %*% S %*% solve(t(X) %*% X))^.5)
```

Finally, this reports the diagonal elements of the $\hat{V}^*$ matrix.

```
212.990530   3.301661  92.187777  88.866352   6.944563
```

How do these compare to the previous standard errors?

White's Heteroskedasticity Test - Example

We'll regress the squared residuals from the OLS regression on all explanatory variables, and squared and cross-products of the explanatory variables. If the R^2 from this auxiliary regression is high enough, we'll reject the null of homoscedasticity.

First, create all the variables needed for the auxiliary regression, then run OLS:

```
age2 = age^2
income4 = income^4
age_own = age*ownrent
age_inc = age*income
age_inc2 = age*income2
own_inc = ownrent*income
own_inc2 = ownrent*income2
inc_inc2 = income^3
summary(lm(resids2 ~ age + ownrent + income + income2 + age2 +
  income4 + age_own + age_inc + age_inc2 + own_inc + own_inc2 +
  inc_inc2))

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1637390.4  1290979.7   1.268   0.2097
age             5366.2    48893.8   0.110   0.9130
ownrent       812036.8   991630.2   0.819   0.4161
income      -2021697.6  1053559.1  -1.919   0.0598 .
income2       669055.3   365666.7   1.830   0.0724 .
age2             -424.1      627.5  -0.676   0.5018
income4          3762.7     2277.4   1.652   0.1038
age_own          4661.7    14424.6   0.323   0.7477
age_inc         11499.9    15614.3   0.736   0.4643
age_inc2        -1093.3     1568.1  -0.697   0.4884
own_inc       -510192.3   469792.6  -1.086   0.2819
own_inc2        51835.1    61799.8   0.839   0.4050
inc_inc2       -86805.3    51162.6  -1.697   0.0950 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 274600 on 59 degrees of freedom
Multiple R-squared:  0.199,     Adjusted R-squared:  0.0361
F-statistic: 1.222 on 12 and 59 DF,  p-value: 0.2905
```

- Can variation in $\boldsymbol{e}'\boldsymbol{e}$ be explained?
- Should we use the F-test reported in the regression results?
- ```
  > 1 - pchisq(n*0.199,12)
  [1] 0.280255
  ```

So, even though regression seems apparent from the plot of *avgexp* against *income*, we cannot reject the null of homoskedasticity using White's test.

What would be the safe thing to do in this case?
  ```