HETEROSCEDASTICITY

Hüseyin Taştan[1]

[1]Yıldız Technical University
Department of Economics

These presentation notes are based on
Introductory Econometrics: A Modern Approach (2nd ed.)
by J. Wooldridge.

14 Aralık 2012

Heteroscedasticity

- One of the Gauss-Markov assumptions, MLR.5: homoscedasticity, states that

$$\text{Var}(u|x_1, x_2, \ldots, x_k) = \sigma^2$$

- According to this assumption, error variance is not related to the explanatory variables.
- If the variability of the unobserved factors is different for different slices of the population, then this assumption fails.
- The constant error variance implies that the conditional variance of the dependent variable is also constant:

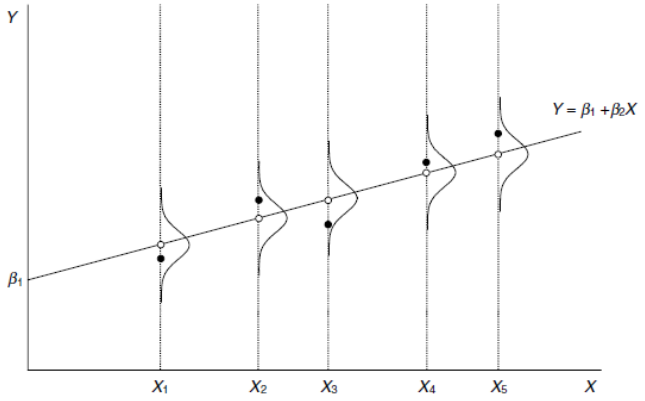
$$\text{Var}(y|x_1, x_2, \ldots, x_k) = \sigma^2$$

- It is important to remember that the failure of MLR.5 does not cause bias or inconsistency in the OLS estimators. But, they are now inefficient and standard errors are not valid.

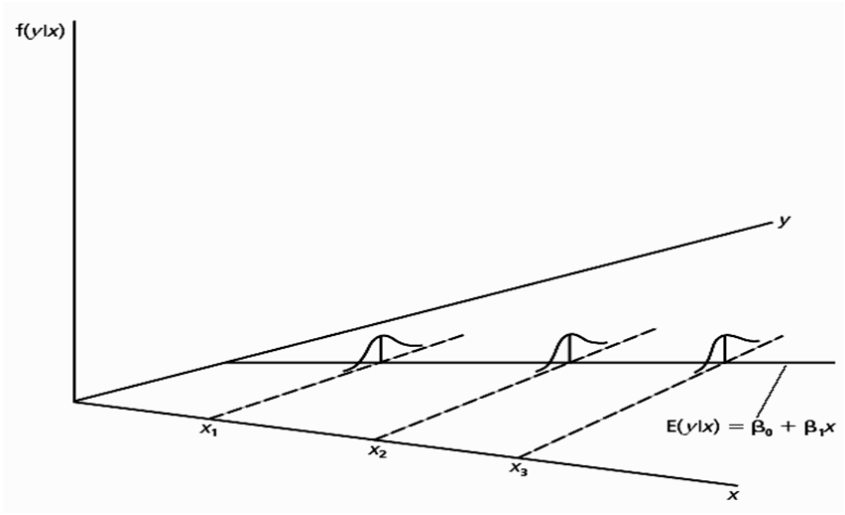
Heteroscedasticity

- Heteroscedasticity is frequently encountered in cross-sectional regression models.
- One reason for this is that the conditional distribution of y has different degree of variability at different levels or layers of population.
- For example, in a model relating household savings to income, household savings may have more variance as income increases. The variance of savings may be low at low levels of income, whereas it may be high at high levels of income
- Similarly, household consumption may have smaller variance at low levels of income.
- Wages may also display varying dispersion at different levels of education.

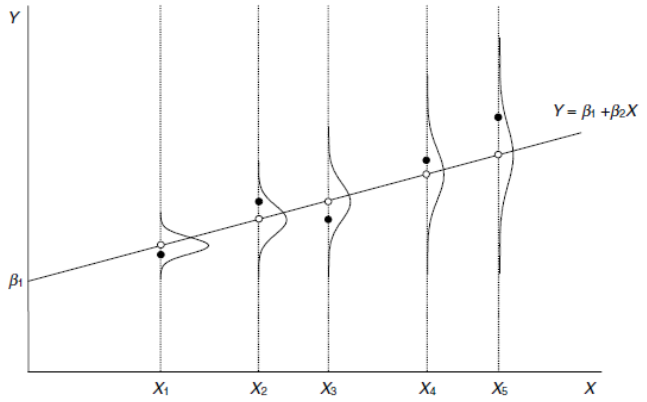
Simple Regression under Homoscedasticity



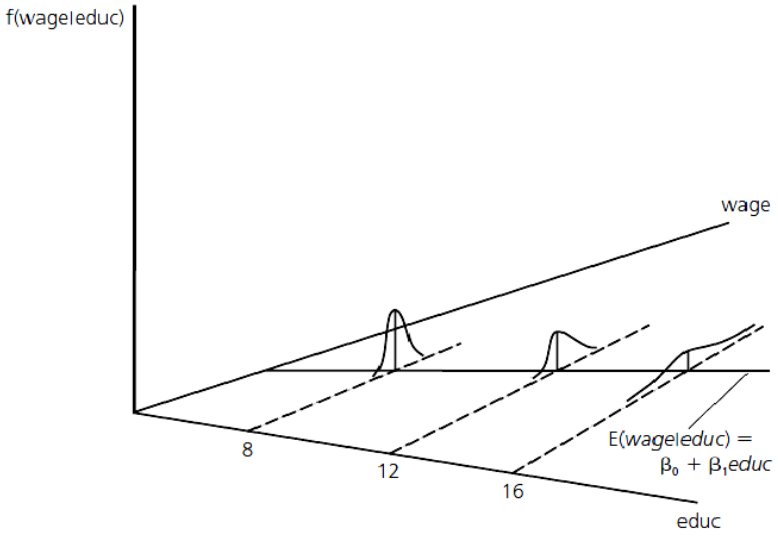
Simple Regression under Homoscedasticity



Heteroscedasticity



Heteroscedasticity



Heteroscedasticity

- Consider the following simple linear regression:

$$y_i = \beta_0 + \beta_1 x_i + u_i$$

- Suppose that the error variance is heteroscedastic:

$$\text{Var}(u_i|x_i) = \sigma_i^2, \quad i = 1, 2, \ldots, n$$

- We know that OLS slope estimator can be writte as:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x}) y_i}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

- Substituting $y = \beta_0 + \beta_1 x + u$ and rearranging we obtain

$$\hat{\beta}_1 = \beta_1 + \frac{\sum_{i=1}^n (x_i - \bar{x}) u_i}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

Heteroscedasticity

- OLS slope estimator in the simple regression is

$$\hat{\beta}_1 = \beta_1 + \frac{\sum_{i=1}^{n}(x_i - \bar{x})u_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

- Under the assumptions MLR.1 through MLR. 4, OLS estimator is unbiased and consistent : $\mathrm{E}(\hat{\beta}_1) = \beta_1$. Since $\mathrm{Var}(\hat{\beta}_1) = \mathrm{E}[(\hat{\beta}_1 - \beta_1)^2]$ the variance of $\hat{\beta}_1$ is given by

$$\mathrm{Var}(\hat{\beta}_1) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2\sigma_i^2}{\left(\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^2}$$

- This is different from the variance formula derived before under the assumption of constant error variance. Notice that, if we had $\sigma_i^2 = \sigma^2$ for each observation (ie, homoscedasticity), then, we would get

$$\mathrm{Var}(\hat{\beta}_1) = \frac{\sigma^2\sum_{i=1}^{n}(x_i - \bar{x})^2}{\left(\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^2} = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Heteroscedasticity

- We just saw that the usual variance formula is no longer valid under heteroscedasticity.
- If the model is heteroscedastic and we continue to use the usual formulas for variances and standard errors then all statistical inference will be invalid and misleading.
- Usual testing procedures, t tests, F tests, LM tests, etc. will **not** be valid.
- If the error term is heteroscedastic then one solution is to use correct formulas for variances and standard errors (so-called heteroscedasticity-robust standard errors).

Heteroscedasticity Robust Standard Errors

- How to adjust standard errors (hence t statistics) so that they are valid in the presence of heteroscedasticity of unknown form.
- Also called White or White-Huber-Ecker standard errors
- In multiple linear regression analysis White heteroscedasticity-robust standard errors are defined as square roots of the following formula:

$$\widehat{\mathrm{Var}(\hat{\beta}_1)} = \frac{\sum_{i=1}^{n}\hat{r}_{ij}^2\hat{u}_i^2}{SSR_j^2}$$

  $\hat{r}_{ij}$: ith residual obtained from the regression of x_j on all other x variables. SSR_j^2: sum of squared residuals from this regression.
- Once we obtained robust standard errors we can calculate heteroscedasticity-robust t statistics.
- Econometric packages, such as GRETL, Eviews, STATA, etc., have special commands to compute these robust standard errors.

Testing for Heteroscedasticity

- Heteroscedasticity-robust standard errors are valid asymptotically. F and LM test statistics can also be made robust to heteroscedasticity.
- To use these robust test statistics we do not need to know whether heteroscedasticity is present.
- However, in small samples we need to implement a more efficient estimation procedure than OLS. We know that OLS is no longer BLUE under heteroscedasticity.
- But first we need to test the presence of heteroscedasticity.
- There are many tests for heteroscedasticity. We will only focus modern tests which detect the kind of heteroscedasticity that invalidates the usual OLS statistics.
- In particular we will learn two tests: Breusch-Pagan and White tests.

Testing for Heteroscedasticity

- We assume that assumptions MLR.1 through MLR.4 are still valid so that OLS is unbiased and consistent.
- We wish to test if the assumption MLR.5 constant variance still holds. Thus, the null hypothesis is:

$$H_0: \ \text{Var}(u|x_1, x_2, \ldots, x_k) = \sigma^2 \qquad \text{constant variance}$$

- Alternative hypothesis is:

$$H_1: \ \text{Var}(u|x_1, x_2, \ldots, x_k) \neq \sigma^2 \qquad \text{heteroscedasticity}$$

- Null hypothesis can also be written as

$$H_0: \ \text{E}(u^2|x_1, x_2, \ldots, x_k) = \text{E}(u^2) = \sigma^2$$

- This hypothesis says that the variance is not related to x_j. Heteroscedasticity tests detect the presence of this relationship.

Testing for Heteroscedasticity

- If the null hypothesis is not correct, the expected value of u^2 can be any function of x_j.
- Assume that this relationship is linear:

$$u^2 = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \ldots + \alpha_k x_k + \nu$$

- The null hypothesis of constant variance can be written as

$$H_0: \ \alpha_1 = \alpha_2 = \ldots = \alpha_k = 0$$

- Under H_0, $\text{E}(u^2|x_1, \ldots, x_k) = \alpha_0$, which is a constant.
- We cannot observe u but we can estimate them. Thus, we can use $\hat{u}$ and carry out an F or LM test.

Testing for Heteroscedasticity

- After estimating the model using OLS, we regress squared residuals on all x variables:

$$\hat{u}^2 = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \ldots + \alpha_k x_k + \text{hata}$$

  And test the joint significance of $x_1, x_2, \ldots, x_k$ using the standard $F(k, n-k-1)$ or $LM = nR^2_{\hat{u}} \sim \chi^2_k$ test procedures.
- If the test statistics are greater than the critical value then we reject the null hypothesis of constant variance in favor of heteroscedasticity.
- The LM version of this test is called **Breusch-Pagan heteroscedasticity test**.

Breusch-Pagan Test for Heteroscedasticity

STEPS

1. Estimate the model using OLS as usual, obtain squared residuals $\hat{u}^2$.
2. Run the following test regression (or auxiliary regression):

$$\hat{u}^2 = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \ldots + \alpha_k x_k + \text{error}$$

   and save the coefficient of determination $R^2_{\hat{u}}$.
3. Using $R^2_{\hat{u}}$ compute either $F \sim F(k, n-k-1)$ or $LM \sim \chi^2_k$ test statistic. (If the test statistic is greater than the critical value reject H_0 : constant variance. This means that there is evidence of heteroscedasticity in the model. Or use p-value: if the p-value is smaller than the chosen significance level, eg 0.05, then we reject the null.)

Example: House Prices, hprice1.gdt

1st STEP: Estimating the Model using OLS:

$$\widehat{\text{price}} = \underset{(29.475)}{-21.77} + \underset{(0.0006)}{0.0021}\,\text{lotsize} + \underset{(0.013)}{0.123}\,\text{sqrft} + \underset{(9.010)}{13.853}\,\text{bdrms}$$

$$n = 88 \quad R^2 = 0.672$$

2nd STEP: Test regression:

$$\widehat{\text{u}}^2 = \underset{(3259.5)}{-5522.79} + \underset{(0.071)}{0.202}\,\text{lotsize} + \underset{(1.464)}{1.691}\,\text{sqrft} + \underset{(996.38)}{1041.76}\,\text{bdrms}$$

$$n = 88 \quad \boldsymbol{R^2 = 0.1601}$$

3rd STEP: Using $R^2 = 0.1601$ compute test statistics: $F = 5.34$ with p-value=0.002, and $LM = 88 \times 0.1601 = 14.09$ (Chi-square with 3 dof, χ^2_3) with p-value=0.0028. **Strong evidence against homoscedasticity**.

Example: House Prices, hprice1.gdt

As we mentioned earlier, using logarithmic transformation may reduce heteroscedasticity. Let us estimate a log-log specification (except rooms) for the house prices:

$$\widehat{\text{logprice}} = \underset{(0.651)}{-1.30} + \underset{(0.038)}{0.17}\,\text{loglotsize} + \underset{(0.093)}{0.70}\,\text{logsqrft} + \underset{(0.028)}{0.04}\,\text{bdrms}$$

$$n = 88 \quad R^2 = 0.643$$

Test statistics and associated p-values are

$$F = 1.141, \quad p - value = 0.245, \qquad LM = 4.22, \quad pvalue = 0.239$$

Obviously p-values are not small enough to reject the null. Therefore, we fail to reject the null hypothesis of homoscedasticity in the model with the logarithmic functional forms. In practice, using log transformation may ease the problem of heteroscedasticity.

White Test for Heteroscedasticity

- In chapter 5 we learned that under the Gauss-Markov assumptions OLS standard errors and test statistics are asymptotically valid.
- This implies that the constant variance assumption (MLR.5) can be replaced with a weaker assumption: the squared error, "u^2, is uncorrelated with all the independent variables, x_j, their squares, x_j^2, and all the cross products, $x_j x_h$, $j \neq h$".
- This weaker assumption forms the basis of the White (1980) test for heteroscedasticity.
- The steps of this test is similar to the Breusch-Pagan test. The only difference is that squares and cross products of x variables are added to the test regression in the second step.

White Test for Heteroscedasticity

- For example, the test regression for $k = 3$ variables will be

$$\begin{aligned}\hat{u}^2 = {} & \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \alpha_3 x_3 \\ & + \alpha_4 x_1^2 + \alpha_5 x_2^2 + \alpha_6 x_3^2 \\ & + \alpha_7 x_1 x_2 + \alpha_8 x_1 x_3 + \alpha_9 x_2 x_3 + \nu\end{aligned}$$

- As k increases the degrees of freedom decreases significantly.
- Compared to the test regression of the Breusch-Pagan test, the White test regression contains 6 more parameters.
- The White test for heteroscedasticity uses the LM test statistic to test the following null hypothesis:

$$H_0: \ \alpha_1 = \alpha_2 = \ldots = \alpha_9 = 0$$

White Test for Heteroscedasticity

- This hypothesis can also tested using F test statistic. Both are asymptotically valid.
- When $k = 6$, the White test has 27 restrictions, using many degrees of freedom. This is a weakness of the White test.
- In practice, usually a simpler version of the White test, given below, is used. This does not lead to reduction in degrees of freedom as k increases.
- Instead of including squares and cross products directly in the test regression, one can use fitted values, $\hat{y}$, and their squares:

$$\hat{u}^2 = \alpha_0 + \alpha_1 \hat{y} + \alpha_2 \hat{y}^2 + \nu$$

- The null hypothesis of homoscedasticity becomes

$$H_0 : \ \alpha_1 = \alpha_2 = 0$$

White Test for Heteroscedasticity

- The null hypothesis of homoscedasticity :

$$H_0 : \ \alpha_1 = \alpha_2 = 0$$

- This can be tested using either F or LM test.
- No matter what k is, this test always has 2 restrictions, conserving on degrees of freedom.
- This test is especially useful when the variance is thought to change with the level of the expected value, E$(y|x)$.

White Test: Example

- **1st STEP**: OLS estimation of model

$$\widehat{\text{lprice}} = \underset{(0.651)}{-1.30} + \underset{(0.038)}{0.17}\ \text{llotsize} + \underset{(0.093)}{0.70}\ \text{lsqrft} + \underset{(0.028)}{0.04}\ \text{bdrms}$$

$$n = 88 \quad R^2 = 0.643$$

- **2nd STEP**: Regression of $\hat{u}^2$ on $\hat{y}$ and $\hat{y}^2$

$$\widehat{\text{u}}^2 = \underset{(3.345)}{5.047} - \underset{(1.163)}{1.709}\,\widehat{lprice} + \underset{(0.100)}{0.145}\,\widehat{lprice}^{\,2}$$

$$n = 88 \quad R^2 = 0.03917$$

- **3rd STEP**: Calculate the test statistic: $LM = nR^2_{\hat{u}} = 88 \times 0.03917 = 3.447$, p-value=0.18. Decision: We fail to reject the null hypothesis of homoscedasticity.

Test for Heteroscedasticity

- Note that we assumed that the assumptions MLR.1-MLR.4 are still valid.
- If these assumptions are not satisfied, for example, if the functional form is incorrect, then the tests for heteroscedasticity may reject the null hypothesis even if the variance is constant.
- In other words, the probability of Type 1. Error can be higher than the nominal significance level.
- This has led some economists to view tests for heteroscedasticity as general misspecification tests.
- But, there are better and more direct tests for functional form misspecification. We should first use these tests to rule out functional form misspecification since this is a more serious problem than heteroscedasticity.

Weighted Least Squares (WLS)

- Suppose that we found evidence of heteroscedasticity in our model. What should we do?
- One option is to use heteroscedasticity-robust standard errors and inference procedures, as we discussed previously.
- But we know that robust inference procedures are only valid if the sample size is large enough.
- An alternative is to use Weighted Least Squares (WLS) instead of OLS.
- Under heteroscedasticity WLS estimators are more efficient than OLS estimators.

Weighted Least Squares (WLS)

- WLS method require the knowledge of the form of the heteroscedasticity. In most cases, it is assumed that the heteroscedasticity is known up to a multiplicative constant.
- In the following multiple linear regression model

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} + u_i$$

suppose that the assumption MLR.5 does not hold and the form of heteroscedasticity is given by:

$$\text{Var}(u_i|x_{i1}, x_{i2}, \ldots, x_{ik}) = \sigma^2 h(x)$$

- Here $h(x) > 0$ is a any function of x variables. We assume that $h(x)$ is known.
- Using $h(x)$ we can transform the original model so that the transformed model has a constant variance. Since the transformed model will not have heteroscedasticity OLS estimation will be efficient. This procedure is called WLS.

Weighted Least Squares (WLS)

- We transform the model by multiplying each term by $1/\sqrt{h(x)}$:

$$\frac{y_i}{\sqrt{h(x)}} = \beta_0 \frac{1}{\sqrt{h(x)}} + \beta_1 \frac{x_{i1}}{\sqrt{h(x)}} + \beta_2 \frac{x_{i2}}{\sqrt{h(x)}} + \ldots + \beta_k \frac{x_{ik}}{\sqrt{h(x)}} + \frac{u_i}{\sqrt{h(x)}}$$

- Or

$$y_i^* = \beta_0 x_{i0}^* + \beta_1 x_{i1}^* + \beta_2 x_{i2}^* + \ldots + \beta_k x_{ik}^* + u_i^*, \quad i = 1, 2, \ldots, n$$

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

- The variance of the error term in the transformed model:

$$\text{Var}(u_i^*|x) = \text{E}(u_i^{*2}|x) = \text{E}\left(\left(\frac{u_i}{\sqrt{h(x)}}\right)^2 |x\right) = \frac{1}{h(x)}\text{E}(u_i^2|x)$$

$$= \frac{1}{h(x)}\sigma^2 h(x) = \sigma^2$$

which is constant.

WLS

- Example: y: savings, x: income

$$y_i = \beta_0 + \beta_1 x_i + u_i$$

$$\text{Var}(u_i|x_i) = \sigma^2 x_i$$

- Here, $h(x) = x_i$. The variance of savings increase with the level of income.
- Since income is always positive, the variance will always be positive too.
- WLS transformation involves dividing all terms by $\sqrt{x_i}$.
- Parameters are interpreted in the same way. Eg., β_1 is still the marginal propensity to save out of income.

Generalized Least Squares (GLS)

- GLS method: applying OLS to the transformed model gives the GLS estimators.
- GLS estimators will be different from the OLS estimators in the original regression. Interpretation of parameter estimates are made in the context of the original model.
- GLS estimators are also used in the presence of serial correlation in the regression analysis of time series data.
- The GLS estimators correcting for heteroscedasticity are called WLS estimators.
- The GLS estimators, β_j^*, are BLUE.
- The R^2 of the transformed model cannot be used as a goodness-of-fit measure. But it can be used to compute test statistics.

GLS, WLS, OLS

- When we apply OLS to the transformed model, the objective function can be written as:

$$\sum_{i=1}^{n} \hat{u_i}^2 = \sum_{i=1}^{n} \frac{1}{h(x)}(\hat{y}_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \ldots - \hat{\beta}_k x_{ik})^2$$

- OLS is a special case of WLS. Equal weight is given to each observation in the OLS estimaton.
- In contrast, WLS uses the inverse of the variance as the weight for each observation. Thus, if the variance is high less weight will be put on that observation (and vice versa).

WLS Example: Saving-Income Relationship

Independent Variables	(1) OLS	(2) WLS	(3) OLS	(4) WLS
inc	.147 (.058)	.172 (.057)	.109 (.071)	.101 (.077)
size	———	———	67.66 (222.96)	−6.87 (168.43)
educ	———	———	151.82 (117.25)	139.48 (100.54)
age	———	———	.286 (50.031)	21.75 (41.31)
black	———	———	518.39 (1,308.06)	137.28 (844.59)
intercept	124.84 (655.39)	−124.95 (480.86)	−1,605.42 (2,830.71)	−1,854.81 (2,351.80)
Observations *R*-Squared	100 .0621	100 .0853	100 .0828	100 .1042

Feasible Generalized Least Squares (FGLS)

- FGLS: Feasible Generalized Least Squares
- To implement GLS-WLS we need to know $h(x)$.
- This is usually not known in practice.
- But we can estimate the form of the variance from the data, resulting in $\hat{h}_i$ for each observation.
- There are many ways to model heteroscedasticity but a widely used flexible approach is to assume

$$\text{Var}(u|x) = \sigma^2 \exp(\delta_0 + \delta_1 x_1 + \delta_2 x_2 + \ldots + \delta_k x_k)$$

where

$$h(x) = \exp(\delta_0 + \delta_1 x_1 + \delta_2 x_2 + \ldots + \delta_k x_k)$$

Steps in Feasible Generalized Least Squares (FGLS)

1. Estimate the model using OLS and save the residuals, $\hat{u}$
2. Compute the logarithm of the squared residuals $\Rightarrow \log(\hat{u}^2)$
3. Regress $\log(\hat{u}^2)$ on $x_1, x_2, \ldots, x_k$. Compute and save fitted values: $\hat{g}$
4. Exponentiate $\hat{g} \Rightarrow \hat{h} = \exp(\hat{g})$
5. Use $1/\hat{h}$ as weights and apply WLS.

FGLS Example: Smoke.gdt

$$\widehat{\text{cigs}} = \underset{(24.079)}{-3.64} + \underset{(0.728)}{0.88}\ \text{lincome} - \underset{(5.773)}{0.751}\ \text{lcigpric} - \underset{(0.167)}{0.50}\ \text{educ}$$
$$+ \underset{(0.160)}{0.771}\ \text{age} - \underset{(0.001)}{0.009}\ \text{agesq} - \underset{(1.111)}{2.826}\ \text{restaurn}$$
$$n = 807 \quad R^2 = 0.053 \quad F(6, 800) = 7.4231 \quad \hat{\sigma} = 13.405$$

(standard errors in parentheses)

White test statistic: LM = 36.15, p-value = 0.000079
Strong evidence of heteroscedasticity.

FGLS Example

Regression of $\log(\hat{u}^2)$ on all x variables:

$$\widehat{\log(u^2)} = \underset{(2.563)}{-1.92} + \underset{(0.077)}{0.29}\ \text{lincome} + \underset{(0.614)}{0.195}\ \text{lcigpric} - \underset{(0.017)}{0.079}\ \text{educ}$$
$$+ \underset{(0.017)}{0.20}\ \text{age} - \underset{(0.0001)}{0.002}\ \text{agesq} - \underset{(0.118)}{0.627}\ \text{restaurn}$$
$$n = 807 \quad \bar{R}^2 = 0.2417$$

Fitted values: $\hat{g}$
Weights: $1/\exp(\hat{g})$

FGLS Example

Using $1/\exp(\hat{g})$ as weights in the WLS estimation:

$$\widehat{\text{cigs}} = \underset{(17.803)}{5.64} + \underset{(0.437)}{1.295}\ \text{lincome} - \underset{(4.46)}{2.94}\ \text{lcigpric} - \underset{(0.120)}{0.46}\ \text{educ}$$
$$+ \underset{(0.097)}{0.482}\ \text{age} - \underset{(0.0009)}{0.006}\ \text{agesq} - \underset{(0.795)}{3.46}\ \text{restaurn}$$
$$n = 807 \quad R^2 = 0.113$$

GRETL: Model $\Rightarrow$ Other Linear Models $\Rightarrow$ Weighted Least Squares
weights: $1/\exp(\hat{g})$

Linear Probability Model (LPM) and FGLS

- We saw previously that LPM is heteroscedastic.
- In this case OLS estimators are not BLUE. How can we apply WLS method for LPM? What are the weights?
- Recall that in the LPM the variance can be written as:

$$\text{Var}(y|x) = p(x)(1 - p(x))$$

- Here, $p(x)$ is the probability of success which is a linear function of x variables.
- Obviously, the variance will change as the fitted values (predicted success probabilities) change. Hence, the weight function, $h(x)$, is:

$$\hat{h}_i = \hat{y}_i(1 - \hat{y}_i)$$

- Values of $\hat{y}_i$ outside the $0 - 1$ interval can be given 0.01 and 0.99. Then we can use FGLS procedure as usual.