Econometrics II
Tutorial Problems No. 4

Lennart Hoogerheide & Agnieszka Borowska

08.03.2017

1 Summary

- **Gauss-Markov assumptions (for multiple linear regression model):**

  MLR.1 (linearity in parameters): The model is

  $$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + u_i,$$

  where $\beta_0, \ldots, \beta_k$ are unknown parameters (constants) and u_i is an unobserved random error term.

  MLR.2 (random sampling): We have a random sample of n independent observations

  $$\{(x_{i1}, \ldots, x_{ik}, y_i) : i = 1, \ldots, n\}.$$

  MLR.3 (no perfect collinearity): No exact linear relationships between variables (and none of the independent variables is constant).

  MLR.4 (zero conditional mean): $\mathbb{E}(u_i | x_{i1}, \ldots, x_{ik}) = 0$.

  MLR.5 (homoskedasticity): $\mathbb{V}\text{ar}(u_i | x_{i1}, \ldots, x_{ik}) = \sigma^2$.

- **Heteroskedasticity of Unknown Form:** Heteroskedasticity that may depend on the explanatory variables in an unknown, arbitrary fashion.

- **Heteroskedasticity-Robust Standard Error:** (White standard errors) A standard error that is (asymptotically) robust to heteroskedasticity of unknown form. Can be obtained as the square root of a diagonal element of

  $$\widehat{\mathbb{V}\text{ar}}(\hat{\beta}_{OLS}) = (X'X)^{-1} X'\hat{\Omega}X (X'X)^{-1},$$

  where $\hat{\Omega} = \text{diag}(\hat{u}_1^2, \ldots, \hat{u}_n^2)$, the diagonal matrix with squared OLS residuals on the diagonal.

- **Heteroskedasticity-Robust Statistic:** A statistic that is (asymptotically) robust to heteroskedasticity of unknown form. E.g. t, F, LM statistics.

- **Breusch-Pagan Test:** (LM test) A test for heteroskedasticity where the squared OLS residuals are regressed on exogenous variables – often (a subset of) the explanatory variables in the model, their squares and/or cross terms.

- **White Test (without cross terms):** A special case of Breusch-Pagan Test, which involves regressing the squared OLS residuals on the squared explanatory variables.

- **Weighted Least Squares (WLS) Estimator:** An estimator used to adjust for a known form of heteroskedasticity, where each squared residual is weighted by the inverse of the variance of the error.

- **Feasible WLS (FWLS) Estimator:** An estimator used to adjust for an unknown form of heteroskedasticity, where variance parameters are unknown and therefore must first be estimated.

2 Extra Topics

2.1 Goldfeld–Quandt (1965) test

In a nutshell

- **Idea:** If the error variances are homoskedastic (equal across observations), then the variance for one part of the sample will be the same as the variance for another part of the sample.
- Based on the ratio of variances.
- Test for the equality of error variances using an F-test on the ratio of two variances.
- **Key assumption:** independent and normally distributed error terms.
- Divide the sample of into three parts, then discard the middle observations.
- Estimate the model for each of the two other sets of observations and compute the corresponding residual variances.

- It requires that the data can be ordered with nondecreasing variance.
- The ordered data set is split in three groups:
  1. the first group consists of the first n_1 observations (with variance σ_1^2);
  2. the second group of the last n_2 observations (with variance σ_2^2);
  3. the third group of the remaining $n_3 = n - n_1 - n_2$ observations in the middle. This last group is left out of the analysis, to obtain a sharper contrast between the variances in the first and second group.
- The null hypothesis is that the variance is constant for all observations, and the alternative is that the variance *increases*.
- Hence, the null and alternative hypotheses are

$$\begin{aligned} H_0 : &\quad \sigma_1^2 = \sigma_2^2, \\ H_1 : &\quad \sigma_1^2 < \sigma_2^2. \end{aligned}$$

- Apply OLS to groups 1 and 2 separately, with resulting sums of squared residuals SSR_1 and SSR_2 respectively and estimated variances $s_1^2 = \frac{SSR_1}{n_1-k}$ and $s_2^2 = \frac{SSR_2}{n_2-k}$.
- Under the assumption of *independently and normally distributed* error terms:

$$\frac{SSR_j}{\sigma_j^2} \sim \chi^2_{n_j-k}, \qquad j = 1, 2,$$

  and these two statistics are independent.
- Therefore:

$$\frac{\frac{SSR_2}{(n_2-k)\sigma_2^2}}{\frac{SSR_1}{(n_1-k)\sigma_1^2}} = \frac{\frac{s_2^2}{\sigma_2^2}}{\frac{s_1^2}{\sigma_1^2}} \sim F(n_2 - k, n_1 - k).$$

- So, *under the null* hypothesis of equal variances, the test statistic

$$F = \frac{s_2^2}{s_1^2} \sim F(n_2 - k, n_1 - k).$$

  The null hypothesis is rejected in favour of the alternative if F takes large values.
- There exists no generally accepted rule to choose the number n_3 of excluded middle observations.
  - If the variance changes only at a single break-point, then it would be optimal to select the two groups accordingly and to take $n_3 = 0$.
  - On the other hand, if nearly all variances are equal and only a few first observations have smaller variance and a few last ones have larger variance, then it would be best to take n_3 large.
  - In practice one uses rules of thumb: e.g. $n_3 = \frac{n}{5}$ if the sample size n is small and $n_3 = \frac{n}{3}$ if n is large.

2.2 Correction factor for multiplicative models

Recall that we distinguish two models for heteroskedasticity in the context of FWLS:

- **multiplicative** heteroskedasticity model

$$\mathbb{V}\text{ar}(u_i|x_i) = \sigma^2 \exp\left(\delta_0 + \delta_1 x_{i1} + \cdots + \delta_k x_{ik}\right);$$

- **additive** heteroskedasticity model

$$\mathbb{V}\text{ar}(u_i|x_i) = \delta_0 + \delta_1 x_{i1} + \cdots + \delta_k x_{ik}.$$

The latter has, however, a disadvantage that (estimate of) $\mathbb{V}\text{ar}(u_i|x_i)$ can be negative, so we mainly focus on the former one.

Notice that in the **multiplicative** model we have

$$\begin{aligned}\mathbb{V}\text{ar}(u_i|x_i) &\overset{\mathbb{E}(u_i|x_i)=0}{=} \mathbb{E}(u_i^2|x_i) \\ &= \sigma^2 \exp\left(\delta_0 + \delta_1 x_{i1} + \cdots + \delta_k x_{ik}\right),\end{aligned}$$

so it is equivalent with

$$\begin{aligned}u_i^2 &= \sigma^2 \exp\left(\delta_0 + \delta_1 x_{i1} + \cdots + \delta_k x_{ik}\right) v_i, \\ v_i &= \frac{u_i^2}{\mathbb{E}(u_i^2|x_i)} \qquad (\Leftarrow \text{mean 1 random variable})\end{aligned}$$

Hence, we consider

$$\log(u_i^2) = \alpha_0 + \delta_1 x_{i1} + \cdots + \delta_k x_{ik} + \eta_i,$$

where η_i is the error term

$$\eta_i = \log(v_i) - \mathbb{E}(\log(v_i))$$

and α_0 is a constant term

$$\alpha_0 = \log(\sigma^2) + \delta_0 + \mathbb{E}(\log(v_i)).$$

Hence, the coefficient δ_0 of the constant term is **not** consistently estimated by $\hat{\alpha}_0$ from OLS. To obtain its consistent estimate a **correction factor** is needed so δ_0 is then estimated by

$$\hat{\delta}_0 + a,$$

where, if the errors are normally distributed ($u_i|x_i \sim \mathcal{N}(0, \sigma_i^2)$),

$$a = -\mathbb{E}[\log(\chi_1^2)] \approx 1.27.$$

We will see how this works in Computer Exercise 2*(i)*[1].

[1] Note, however, that a consistent estimator of δ_0 is not needed, because $\exp(\hat{\delta}_0)$ is merely a constant scaling factor that does not affect the FWLS estimator.

3 Warm-up Exercises

3.1 W8/1

Which of the following are consequences of heteroskedasticity?

(i) *The OLS estimators, $\hat{\beta}_j$, are inconsistent.*

The homoskedasticity assumption played no role in showing that the OLS estimator is consistent. Indeed, even with $\mathbb{V}\text{ar}(u|X) = \Omega \neq \sigma^2 \mathbb{I}$ we have for $\hat{\beta}_{OLS} = \beta + (X'X)^{-1} X'u$:

$$\begin{aligned} \text{plim}\left(\hat{\beta}_{OLS}\right) &= \beta + \text{plim}\left(\frac{X'X}{n}\right)^{-1} \text{plim}\left(\frac{X'u}{n}\right) \\ &= \beta + \text{plim}\left(\frac{1}{n}\sum_{i=1}^{n} x_i x_i'\right)^{-1} \text{plim}\left(\frac{1}{n}\sum_{i=1}^{n} x_i u_i\right) \\ &= \beta + \mathbb{E}(X'X)^{-1} \underbrace{\mathbb{E}(X'u)}_{=\mathbb{E}(X\mathbb{E}(u|X))=0}, \end{aligned}$$

so the OLS estimator is still consistent.

(ii) *The usual (homoskedasticity-only) F statistic no longer has an F distribution.*

Now, we have

$$\mathbb{V}\text{ar}(\hat{\beta}_{OLS}) = (X'X)^{-1} X'\Omega X (X'X)^{-1},$$

so the usual expression

$$\sigma^2 (X'X)^{-1}$$

for the variance does not apply anymore. The latter expression is biased, which makes the standard (homoskedasticity-only) F test (and t test) invalid. One should use a heteroskedasticity-robust F (and t) statistic, based on heteroskedasticity-robust standard errors.

(iii) *The OLS estimators are no longer BLUE.*

As heteroskedasticity is a violation of the Gauss-Markov assumptions, the OLS estimator is no longer BLUE: it is still linear, unbiased, but not "best" in a sense that it is not efficient. Intuitively, the inefficiency of the OLS estimator under heteroskedasticity can be contributed to the fact that observations with low variance are likely to convey more information about the parameters than observations with high variance, and so the former should be given more weight in an efficient estimator (but all are weighted equally).

3.2 W8/2

Consider a linear model to explain monthly beer consumption:

$$\begin{aligned} beer &= \beta_0 + \beta_1 inc + \beta_2 price + \beta_3 educ + \beta_4 female + u, \\ \mathbb{E}(u|inc, price, educ, female) &= 0, \\ \mathbb{V}ar(u|inc, price, educ, female) &= \sigma^2 inc^2. \end{aligned}$$

Write the transformed equation that has a homoskedastic error term.

With $\mathbb{V}\text{ar}(u|inc, price, educ, female) = \sigma^2 inc^2$ we have $h(x) = inc^2$, where $h(x)$ is a function of the explanatory variables that determines the heteroskedasticity (defined as $\mathbb{V}\text{ar}(u|x) = \sigma^2 h(x)$). Therefore, $\sqrt{h(x)} = inc$, and so the transformed equation is obtained by dividing the original equation by inc:

$$\begin{aligned} \frac{beer}{inc} &= \beta_0 \frac{1}{inc} + \beta_1 \frac{inc}{inc} + \beta_2 \frac{price}{inc} + \beta_3 \frac{educ}{inc} + \beta_4 \frac{female}{inc} + \frac{u}{inc} \\ &= \beta_0 \frac{1}{inc} + \beta_1 + \beta_2 \frac{price}{inc} + \beta_3 \frac{educ}{inc} + \beta_4 \frac{female}{inc} + \frac{u}{inc}. \end{aligned}$$

Notice that β_1, which is the slope on inc in the original model, is now a constant in the transformed equation. This is simply a consequence of the form of the heteroskedasticity and the functional forms of the explanatory variables in the original equation.

3.3 Small computer exercise

Using the data in the file `earnings.wf1`[2] *run the regression*

$$y_i = \beta_1 d_{1i} + \beta_2 d_{2i} + \beta_3 d_{3i} + u_i \tag{1}$$

where d_{ki}, $k = 1, 2, 3$, *are dummy variables for three age groups. Then test the null hypothesis that* $\mathbb{E}(u_i^2) = \sigma^2$ *against the alternative that*

$$\mathbb{E}(u_i^2) = \gamma_1 d_{1i} + \gamma_2 d_{2i} + \gamma_3 d_{3i}.$$

Report p-values for both F *and* nR^2 *tests.*

Recall that tests for homoskedasticity are constructed as follows:

$$H_0 : \text{homoskedasticity,}$$
$$H_1 : \text{not } H_0, \text{ i.e. heteroskedasticity.}$$

The easiest way to perform the required test is simply to regress the squared residuals from (1) on a constant and two of the three (to prevent collinearity) dummy variables. Notice that this gives us the same results as running the built-in heteroskedastisity test (Breusch-Pagan-Godfrey) in EViews:

Dependent Variable: RESID^2
Method: Least Squares
Sample: 1 4266
Included observations: 4266

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	2.72E+08	11436983	23.79601	0.0000
GROUP1	-57210408	17471405	-3.274517	0.0011
GROUP2	-38452071	15687465	-2.451133	0.0143

R-squared	0.002747	Mean dependent var	2.42E+08
Adjusted R-squared	0.002280	S.D. dependent var	4.40E+08
S.E. of regression	4.40E+08	Akaike info criterion	42.64243
Sum squared resid	8.25E+20	Schwarz criterion	42.64690
Log likelihood	-90953.30	Hannan-Quinn criter.	42.64401
F-statistic	5.872230	Durbin-Watson stat	0.019275
Prob(F-statistic)	0.002839		

Heteroskedasticity Test: Breusch-Pagan-Godfrey

F-statistic	5.872230	Prob. F(2,4263)	0.0028
Obs*R-squared	11.72044	Prob. Chi-Square(2)	0.0029
Scaled explained SS	19.34589	Prob. Chi-Square(2)	0.0001

Test Equation:
Dependent Variable: RESID^2
Method: Least Squares
Sample: 1 4266
Included observations: 4266
Collinear test regressors dropped from specification

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	2.72E+08	11436983	23.79601	0.0000
GROUP1	-57210408	17471405	-3.274517	0.0011
GROUP2	-38452071	15687465	-2.451133	0.0143

R-squared	0.002747	Mean dependent var	2.42E+08
Adjusted R-squared	0.002280	S.D. dependent var	4.40E+08
S.E. of regression	4.40E+08	Akaike info criterion	42.64243
Sum squared resid	8.25E+20	Schwarz criterion	42.64690
Log likelihood	-90953.30	Hannan-Quinn criter.	42.64401
F-statistic	5.872230	Durbin-Watson stat	0.019275
Prob(F-statistic)	0.002839		

- The F statistic from this regression for the hypothesis that the coefficients of the dummy variables are zero is 5.872. It is asymptotically distributed as $F(k, n-k-1) = F(2, 4263)$, and the p-value is 0.0028.
- An alternative statistic is nR^2, which is equal to 11.72. It is asymptotically distributed as $\chi_k^2 = \chi_2^2$, and the p value is 0.0029. (Recall from the lecture that this is worse than F test in finite samples).

The two test statistics yield identical inferences, namely, that the null hypothesis should be rejected at any conventional significance level.

4 Problem on heteroskedasticity modelling

Consider the model $y_i = \beta x_i + \varepsilon_i$ *(without constant term and with* $k = 1$*), where* $x_i > 0$ *for all observations,* $\mathbb{E}(\varepsilon_i) = 0$, $\mathbb{E}(\varepsilon_i \varepsilon_j) = 0$, $i \neq j$, *and* $\mathbb{E}(\varepsilon_i^2) = \sigma_i^2$. *Consider the following three estimators of* β:

$$b_1 = \frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2},$$
$$b_2 = \frac{\sum_{i=1}^n y_i}{\sum_{i=1}^n x_i},$$
$$b_3 = \frac{1}{n} \sum_{i=1}^n \frac{y_i}{x_i}.$$

[2] Average annual earnings in 1988 and 1989, in 1982 US dollars, for individuals in three age groups.

For each estimator, derive a model for the variances σ_i^2 for which this estimator is the best linear unbiased estimator of β.

Recall that when we have a model for heteroskedasticity, i.e. in $\mathbb{V}\text{ar}(u_i|x_i) = \sigma^2 h(x_i)$ the function $h_i = h(x_i)$ is known, then transforming the original data by dividing them by $\sqrt{h_i}$ results in a linear regression where all Gauss-Markov assumptions are satisfied, which means that the corresponding OLS estimator is BLUE.

Consider

$$y_i = \beta x_i + \varepsilon_i, \qquad \mathbb{V}\text{ar}(u_i|x_i) = \sigma^2 h_i,$$

$$\underbrace{\frac{y_i}{\sqrt{h_i}}}_{=:y_i^*} = \beta \underbrace{\frac{x_i}{\sqrt{h_i}}}_{=:x_i^*} + \underbrace{\frac{\varepsilon_i}{\sqrt{h_i}}}_{=:\varepsilon_i*}, \qquad \mathbb{V}\text{ar}\left(\frac{u_i}{\sqrt{h_i}}\middle| x_i\right) = \sigma^2,$$

so that the corresponding OLS estimator is

$$\begin{aligned}\hat{\beta}_{OLS} &= \frac{\sum_{i=1}^n x_i^* y_i^*}{\sum_{i=1}^n (x_i^*)^2} \\ &= \frac{\sum_{i=1}^n \frac{x_i}{\sqrt{h_i}} \frac{y_i}{\sqrt{h_i}}}{\sum_{i=1}^n \left(\frac{x_i}{\sqrt{h_i}}\right)^2} \\ &= \frac{\sum_{i=1}^n \frac{x_i y_i}{h_i}}{\sum_{i=1}^n \frac{x_i^2}{h_i}}.\end{aligned}$$

Hence, we simply need to find what functions h_i have led to the three given WLS estimators b_1–b_3.

1. To have $\hat{\beta}_{OLS} = b_1$ we need

$$\frac{\sum_{i=1}^n \frac{x_i y_i}{h_i}}{\sum_{i=1}^n \frac{x_i^2}{h_i}} = \frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2},$$

which means that $h_i = 1$, $i = 1, \ldots, n$ (or $h_i = C$ for any other positive constant C, since this would simply drop out in the numerator and the denominator), and $\mathbb{V}\text{ar}(u_i|x_i) = \sigma^2$. Notice that this is simply the OLS estimator for the homoskedastic case.

2. To have $\hat{\beta}_{OLS} = b_2$ we need

$$\frac{\sum_{i=1}^n \frac{x_i y_i}{h_i}}{\sum_{i=1}^n \frac{x_i^2}{h_i}} = \frac{\sum_{i=1}^n y_i}{\sum_{i=1}^n x_i},$$

which means that $h_i = x_i$, $i = 1, \ldots, n$ (or $h_i = Cx_i$ for any other positive constant C), and $\mathbb{V}\text{ar}(u_i|x_i) = \sigma^2 x_i$. Notice that this is a valid expression for the variance due to the assumption that $x_i > 0$, $i = 1, \ldots, n$.

3. To have $\hat{\beta}_{OLS} = b_3$ we need

$$\frac{\sum_{i=1}^n \frac{x_i y_i}{h_i}}{\sum_{i=1}^n \frac{x_i^2}{h_i}} = \frac{1}{n}\sum_{i=1}^n \frac{y_i}{x_i} = \frac{\sum_{i=1}^n \frac{y_i}{x_i}}{n} = \frac{\sum_{i=1}^n \frac{x_i}{x_i}\frac{y_i}{x_i}}{\sum_{i=1}^n \frac{x_i^2}{x_i^2}},$$

which means that $h_i = x_i^2$, $i = 1, \ldots, n$ (or $h_i = Cx_i^2$ for any other positive constant C), and $\mathbb{V}\text{ar}(u_i|x_i) = \sigma^2 x_i^2$.

5 Computer Exercises

Exercise 1

Simulate $n = 100$ data points as follows. Let x_i consist of 100 random drawings from the standard normal distribution, let η_i be a random drawing from the distribution $\mathcal{N}(0, x_i^2)$, and let $y_i = x_i + \eta_i$ (i.e. the true value is $\beta = 1$). We will estimate the model $y_i = \beta x_i + \varepsilon_i$.

(i) *Estimate β by OLS. Compute the homoskedasticity-only standard error of $\hat{\beta}_{OLS}$ and the White heteroskedasticity-robust standard error of $\hat{\beta}_{OLS}$.*

Dependent Variable: Y
Method: Least Squares
Date: 03/07/17 Time: 15:20
Sample: 1 100
Included observations: 100

Variable	Coefficient	Std. Error	t-Statistic	Prob.
X	0.979034	0.095976	10.20087	0.0000

R-squared	0.499684	Mean dependent var	-0.218837
Adjusted R-squared	0.499684	S.D. dependent var	1.359004
S.E. of regression	0.961264	Akaike info criterion	2.768815
Sum squared resid	91.47887	Schwarz criterion	2.794867
Log likelihood	-137.4407	Hannan-Quinn criter.	2.779358
Durbin-Watson stat	2.100710		

Dependent Variable: Y
Method: Least Squares
Date: 03/07/17 Time: 15:20
Sample: 1 100
Included observations: 100
White heteroskedasticity-consistent standard errors & covariance

Variable	Coefficient	Std. Error	t-Statistic	Prob.
X	0.979034	0.159735	6.129109	0.0000

R-squared	0.499684	Mean dependent var	-0.218837
Adjusted R-squared	0.499684	S.D. dependent var	1.359004
S.E. of regression	0.961264	Akaike info criterion	2.768815
Sum squared resid	91.47887	Schwarz criterion	2.794867
Log likelihood	-137.4407	Hannan-Quinn criter.	2.779358
Durbin-Watson stat	2.100710		

(ii) *Estimate β by WLS using the knowledge that $\sigma_i^2 = \sigma^2 x_i^2$. Compare the estimate and the homoskedasticity-only and heteroskedasticity-robust standard errors obtained for this WLS estimator with the results for OLS in (i).*

We start with constructing the (correctly) transformed series:

$$y_i^* := \frac{y_i}{x_i}, \qquad x_i^* := \frac{x_i}{x_i} = 1, \qquad \varepsilon_i^* := \frac{\varepsilon_i}{x_i},$$

so that now the transformed error terms ε_i^* are homoskedastic. We then run two OLS regressions on the transformed series (one with the homoskedasticity-only standard errors and one with the White heteroskedasticity-robust standard errors). Not surprisingly, both give us the same results.

Dependent Variable: Y_STAR
Method: Least Squares
Date: 03/07/17 Time: 15:20
Sample: 1 100
Included observations: 100

Variable	Coefficient	Std. Error	t-Statistic	Prob.
X_STAR	1.026907	0.098879	10.38549	0.0000

R-squared	0.000000	Mean dependent var	1.026907
Adjusted R-squared	0.000000	S.D. dependent var	0.988790
S.E. of regression	0.988790	Akaike info criterion	2.825281
Sum squared resid	96.79293	Schwarz criterion	2.851333
Log likelihood	-140.2640	Hannan-Quinn criter.	2.835824
Durbin-Watson stat	1.834168		

Dependent Variable: Y_STAR
Method: Least Squares
Date: 03/07/17 Time: 15:20
Sample: 1 100
Included observations: 100
White heteroskedasticity-consistent standard errors & covariance

Variable	Coefficient	Std. Error	t-Statistic	Prob.
X_STAR	1.026907	0.098879	10.38549	0.0000

R-squared	0.000000	Mean dependent var	1.026907
Adjusted R-squared	0.000000	S.D. dependent var	0.988790
S.E. of regression	0.988790	Akaike info criterion	2.825281
Sum squared resid	96.79293	Schwarz criterion	2.851333
Log likelihood	-140.2640	Hannan-Quinn criter.	2.835824
Durbin-Watson stat	1.834168		

Next, we run two WLS regressions on the original series, using the correct weights, $h_i = x_i^2$ (again, one with the homoskedasticity-only standard errors and one with the White heteroskedasticity-robust standard errors). Notice that because now x_i can be negative we need to take their absolute values for weighting. As expected, the results are exactly the same as in the previous 'transformed' case.

Dependent Variable: Y
Method: Least Squares
Date: 03/07/17 Time: 16:25
Sample: 1 100
Included observations: 100
Weighting series: @ABS(X)
Weight type: Standard deviation (average scaling)

Variable	Coefficient	Std. Error	t-Statistic	Prob.
X	1.026907	0.098879	10.38549	0.0000

Weighted Statistics

R-squared	0.511002	Mean dependent var	-0.015143
Adjusted R-squared	0.511002	S.D. dependent var	0.111590
S.E. of regression	0.077913	Akaike info criterion	-2.256509
Sum squared resid	0.600966	Schwarz criterion	-2.230458
Log likelihood	113.8255	Hannan-Quinn criter.	-2.245966
Durbin-Watson stat	2.074588	Weighted mean dep.	0.016349

Unweighted Statistics

R-squared	0.498427	Mean dependent var	-0.218837
Adjusted R-squared	0.498427	S.D. dependent var	1.359004
S.E. of regression	0.962471	Sum squared resid	91.70878
Durbin-Watson stat	2.103202		

Dependent Variable: Y
Method: Least Squares
Date: 03/07/17 Time: 16:26
Sample: 1 100
Included observations: 100
Weighting series: @ABS(X)
Weight type: Standard deviation (average scaling)
White heteroskedasticity-consistent standard errors & covariance

Variable	Coefficient	Std. Error	t-Statistic	Prob.
X	1.026907	0.098879	10.38549	0.0000

Weighted Statistics

R-squared	0.511002	Mean dependent var	-0.015143
Adjusted R-squared	0.511002	S.D. dependent var	0.111590
S.E. of regression	0.077913	Akaike info criterion	-2.256509
Sum squared resid	0.600966	Schwarz criterion	-2.230458
Log likelihood	113.8255	Hannan-Quinn criter.	-2.245966
Durbin-Watson stat	2.074588	Weighted mean dep.	0.016349

Unweighted Statistics

R-squared	0.498427	Mean dependent var	-0.218837
Adjusted R-squared	0.498427	S.D. dependent var	1.359004
S.E. of regression	0.962471	Sum squared resid	91.70878
Durbin-Watson stat	2.103202		

(iii) Now estimate β by WLS using the (incorrect) heteroskedasticity model $\sigma_i^2 = \frac{\sigma^2}{x_i^2}$. Compute the standard error of this estimate in three ways: by the WLS expression corresponding to this (incorrect) model, by the White method for OLS on the (incorrectly) weighted data, and also by deriving the correct formula for the standard deviation of WLS with this incorrect model for the variance.

We start with constructing the (incorrectly) transformed series:

$$y_i^{**} := y_i x_i, \qquad x_i^{**} := x_i x_i = x_i^2, \qquad \varepsilon_i^{**} := \varepsilon_i x_i,$$

so that now the transformed error terms ε_i^{**} are heteroskedastic. To have a reference to the previous subpoint, we run four regressions: two OLS ones and two WLS ones, each time with one with the homoskedasticity-only standard errors and one with the White heteroskedasticity-robust standard errors. Now the not-heteroskedasticity-robustified regressions (OLS and WLS) give the same results, and so do both (OLS and WLS) with the White correction.

Dependent Variable: Y_STAR2
Method: Least Squares
Date: 03/07/17 Time: 15:20
Sample: 1 100
Included observations: 100

Variable	Coefficient	Std. Error	t-Statistic	Prob.
X_STAR2	0.913154	0.089559	10.19616	0.0000

R-squared	0.400692	Mean dependent var	0.982115
Adjusted R-squared	0.400692	S.D. dependent var	2.064220
S.E. of regression	1.598016	Akaike info criterion	3.785353
Sum squared resid	252.8120	Schwarz criterion	3.811405
Log likelihood	-188.2676	Hannan-Quinn criter.	3.795897
Durbin-Watson stat	2.032602		

Dependent Variable: Y_STAR2
Method: Least Squares
Date: 03/07/17 Time: 15:20
Sample: 1 100
Included observations: 100
White heteroskedasticity-consistent standard errors & covariance

Variable	Coefficient	Std. Error	t-Statistic	Prob.
X_STAR2	0.913154	0.229583	3.977439	0.0001

R-squared	0.400692	Mean dependent var	0.982115
Adjusted R-squared	0.400692	S.D. dependent var	2.064220
S.E. of regression	1.598016	Akaike info criterion	3.785353
Sum squared resid	252.8120	Schwarz criterion	3.811405
Log likelihood	-188.2676	Hannan-Quinn criter.	3.795897
Durbin-Watson stat	2.032602		

Dependent Variable: Y
Method: Least Squares
Date: 03/07/17 Time: 16:38
Sample: 1 100
Included observations: 100
Weighting series: 1/@ABS(X)
Weight type: Standard deviation (average scaling)

Variable	Coefficient	Std. Error	t-Statistic	Prob.
X	0.913154	0.089559	10.19616	0.0000

Weighted Statistics

R-squared	0.496523	Mean dependent var	-0.271670
Adjusted R-squared	0.496523	S.D. dependent var	2.268110
S.E. of regression	1.595508	Akaike info criterion	3.782211
Sum squared resid	252.0189	Schwarz criterion	3.808263
Log likelihood	-188.1106	Hannan-Quinn criter.	3.792755
Durbin-Watson stat	2.135345	Weighted mean dep.	-0.401395

Unweighted Statistics

R-squared	0.497303	Mean dependent var	-0.218837
Adjusted R-squared	0.497303	S.D. dependent var	1.359004
S.E. of regression	0.963549	Sum squared resid	91.91425
Durbin-Watson stat	2.094606		

Dependent Variable: Y
Method: Least Squares
Date: 03/07/17 Time: 16:38
Sample: 1 100
Included observations: 100
Weighting series: 1/@ABS(X)
Weight type: Standard deviation (average scaling)
White heteroskedasticity-consistent standard errors & covariance

Variable	Coefficient	Std. Error	t-Statistic	Prob.
X	0.913154	0.229583	3.977439	0.0001

Weighted Statistics

R-squared	0.496523	Mean dependent var	-0.271670
Adjusted R-squared	0.496523	S.D. dependent var	2.268110
S.E. of regression	1.595508	Akaike info criterion	3.782211
Sum squared resid	252.0189	Schwarz criterion	3.808263
Log likelihood	-188.1106	Hannan-Quinn criter.	3.792755
Durbin-Watson stat	2.135345	Weighted mean dep.	-0.401395

Unweighted Statistics

R-squared	0.497303	Mean dependent var	-0.218837
Adjusted R-squared	0.497303	S.D. dependent var	1.359004
S.E. of regression	0.963549	Sum squared resid	91.91425
Durbin-Watson stat	2.094606		

What is left is to derive the correct formula for the standard deviation of WLS under the incorrect model for the variance. Recall that in the one-variable (and without a constant term) setting we have

$$\hat{\beta}_{WLS} = \frac{\sum_{i=1}^n \frac{x_i y_i}{h_i}}{\sum_{i=1}^n \frac{x_i^2}{h_i}},$$

so with the weights $h_i = \frac{1}{x_i^2}$ and using $y_i = \beta x_i + \varepsilon_i$, we arrive at

$$\begin{aligned} \hat{\beta}_{WLS} &= \frac{\sum_{i=1}^n x_i^3 y_i}{\sum_{i=1}^n x_i^4} \\ &= \frac{\sum_{i=1}^n x_i^3(\beta x_i + \varepsilon_i)}{\sum_{i=1}^n x_i^4} \\ &= \beta + \frac{\sum_{i=1}^n x_i^3 \varepsilon_i}{\sum_{i=1}^n x_i^4}. \end{aligned}$$

Because $\hat{\beta}_{WLS}$ is unbiased, i.e. $\mathbb{E}\left(\hat{\beta}_{WLS}\middle| x\right) = \beta$, the variance of $\hat{\beta}_{WLS}$ is

$$\begin{aligned}
\mathbb{V}\text{ar}\left(\hat{\beta}_{WLS}\middle| x\right) &= \mathbb{E}\left[\left(\hat{\beta}_{WLS} - \mathbb{E}\left(\hat{\beta}_{WLS}\middle| x\right)\right)^2 \middle| x\right] \\
&= \mathbb{E}\left[\left(\beta + \frac{\sum_{i=1}^n x_i^3 \varepsilon_i}{\sum_{i=1}^n x_i^4} - \beta\right)^2 \middle| x\right] \\
&= \mathbb{E}\left[\frac{\left(\sum_{i=1}^n x_i^3 \varepsilon_i\right)^2}{\left(\sum_{i=1}^n x_i^4\right)^2} \middle| x\right] \\
&\overset{(*)}{=} \frac{\sum_{i=1}^n x_i^6 \mathbb{E}\left[\varepsilon_i^2 | x_i\right]}{\left(\sum_{i=1}^n x_i^4\right)^2} \\
&\overset{(**)}{=} \frac{\sum_{i=1}^n x_i^6 \mathbb{V}\text{ar}\left[\varepsilon_i | \, x_i\right]}{\left(\sum_{i=1}^n x_i^4\right)^2} \\
&\overset{(***)}{=} \frac{\sum_{i=1}^n x_i^8}{\left(\sum_{i=1}^n x_i^4\right)^2},
\end{aligned}$$

where in $(*)$ we use the conditioning on x and the fact that ε_i are mutually independent, in $(**)$ the fact that $\mathbb{E}(\varepsilon_i|x_i) = 0$ and in $(* * *)$ that $\mathbb{V}\text{ar}(\varepsilon_i|x_i) = \sigma^2 x_i^2 = x_i^2$.

For the simulated x_i we obtain $\sum_{i=1}^n x_i^4 = 318.3814$ and $\sum_{i=1}^n x_i^8 = 9962.1182$, hence

$$\widehat{\mathbb{V}\text{ar}}\left(\hat{\beta}_{WLS}\middle| x\right) = \frac{9962.1182}{(318.3814)^2} = 0.0983,$$

so that the standard deviation of $\hat{\beta}_{WLS}$ is $\sqrt{0.0983} \approx 0.3135$. This shows that the standard error from the heteroskedasticity-robust regressions of 0.22 is still estimated with some error.

(iv) Perform 1000 *simulations, where the* $n = 1000$ *values of* x_i *remain the same over all simulations but the* 100 *values of* η_i *are different drawings from the* $\mathcal{N}(0, x_i^2)$ *distributions and where the values of* $y_i = x_i + \eta_i$ *differ accordingly between the simulations. Determine the sample standard deviations over the* 1000 *simulations of the three estimators of* β *in (i)-(iii), that is, OLS, WLS (with correct weights), and WLS (with incorrect weights).*

Figure 1 present an EViews code used for this simulation experiment (and also for the previous computations). The standard deviations of the obtained series of 1000 estimates for β using the required three methods are as follows:

$$\begin{aligned}
\text{St.dev}(\hat{\beta}_{OLS}) &= 0.1799, \\
\text{St.dev}(\hat{\beta}_{WLS,correct}) &= 0.0972, \\
\text{St.dev}(\hat{\beta}_{WLS,incorrect}) &= 0.3155.
\end{aligned}$$

Notice that the last value is almost identical to the theoretical one, obtained in *(iii)*.

(v) Compare the three sample standard deviations in (iv) with the estimated standard errors in (i)–(iii), and comment on the outcomes. Which standard errors are reliable, and which ones are not?

The table below summarises the required results. Clearly, WLS with the correctly specified model for the variances gives reliable standard errors. OLS and WLS with the incorrect weighting greatly underestimate the variability of the estimator for β when the heteroskedasticity-robust standard errors are not used. When the latter are applied the standard error for both methods improve considerably, but still are estimated with some error.

	Single estimation st. errors		
Method	Homosked. only	Heterosked. robust	Simulation st. deviations
OLS	0.0956	0.1597	0.1799
WLS correct	0.0989	0.0989	0.0972
WLS incorrect	0.0895	0.2296	0.3155

```
wfcreate(wf=C1) u 100
rndseed 123456
!N = 100

series x = nrnd
series u = nrnd
series eta = x*u
series y = x + eta

' (i)
equation eq.ls y x
equation eq_white.ls(cov=white) y x

'(ii)
series x_star = x/x
series y_star = y/x
equation eq_wls_true.ls y_star x_star
equation eq_wls_true_white.ls(cov=white) y_star x_star

'standard deviation (average scaling), weighting series x
equation eq_wls_true_model.ls(w=@abs(x), wtype=stdev, wscale=avg) y x
equation eq_wls_true_model_white.ls(w=@abs(x), wtype=stdev, wscale=avg,cov=white) y x

' (iii)
series x_star2 = x*x
series y_star2 = y*x
equation eq_wls_false.ls y_star2 x_star2
equation eq_wls_false_white.ls(cov=white) y_star2 x_star2
equation eq_wls_false_model.ls(w=1/@abs(x), wtype=stdev, wscale=avg) y x
equation eq_wls_false_model_white.ls(w=1/@abs(x), wtype=stdev, wscale=avg,cov=white) y x

series x4 = x^4
scalar sum_x4 = @sum(x4)
series x8 = x^8
scalar sum_x8 = @sum(x8)

' (iv)
!M = 1000
matrix(!M,1) betas_ols
matrix(!M,1) betas_wls_true
matrix(!M,1) betas_wls_false

for !i=1 to !M
    series u = nrnd
    series eta = x*u
    series y = x + eta

    equation eq_sim.ls y x
    betas_ols(!i,1) = eq_sim.@coefs(1)

    series y_star = y/x
    equation eq_wls_true_sim.ls y_star x_star
    betas_wls_true(!i,1) = eq_wls_true_sim.@coefs(1)

    series y_star2 = y*x
    equation eq_wls_false_sim.ls y_star2 x_star2
    betas_wls_false(!i,1) = eq_wls_false_sim.@coefs(1)
next

scalar sd_ols = @stdev(betas_ols)
scalar sd_wls_true = @stdev(betas_wls_true)
scalar sd_wls_false = @stdev(betas_wls_false)
```

Figure 1: EViews code example for Computer Exercise 1.

Exercise 2

Consider the bank wages data `bankwages.wf1` *with the regression model*

$$y_i = \beta_1 + \beta_2 x_i + \beta_3 D_{gi} + \beta_4 D_{mi} + \beta_5 D_{2i} + \beta_6 D_{3i} + \varepsilon_i,$$

where y_i is the logarithm of yearly wage, x_i is the number of years of education, D_g is a gender dummy (1 for males, 0 for females), and D_m is a minority dummy (1 for minorities, 0 otherwise). Administration is taken as reference category and D_2 and D_3 are dummy variables ($D_2 = 1$ for individuals with a custodial job and $D_2 = 0$ otherwise, and $D_3 = 1$ for individuals with a management position and $D_3 = 0$ otherwise).

(i) *Consider the following multiplicative model for the variances:*

$$\sigma_i^2 = \mathbb{E}[\varepsilon_i^2] = e^{\gamma_1 + \gamma_2 D_2 + \gamma_3 D_3}.$$

Estimate the nine parameters (six regression parameters and three variance parameters) by (two-step) FWLS. Obtain the estimates of the standard deviations per job category and interpret the results.

To apply (two-step) FWLS, we start by estimating the regression and the model for variances by OLS. For the latter we consider as the explained variable $\log(\hat{\varepsilon}_i^2)$, where $\hat{\varepsilon}_i$ are the OLS residuals of from the first regression.

Dependent Variable: LOGSALARY
Method: Least Squares
Sample: 1 474
Included observations: 474

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	9.574694	0.054218	176.5965	0.0000
EDUC	0.044192	0.004285	10.31317	0.0000
GENDER	0.178340	0.020962	8.507685	0.0000
MINORITY	-0.074858	0.022459	-3.333133	0.0009
DUMJCAT2	0.170360	0.043494	3.916891	0.0001
DUMJCAT3	0.539075	0.030213	17.84248	0.0000

R-squared	0.760775	Mean dependent var	10.35679
Adjusted R-squared	0.758219	S.D. dependent var	0.397334
S.E. of regression	0.195374	Akaike info criterion	-0.415222
Sum squared resid	17.86407	Schwarz criterion	-0.362549
Log likelihood	104.4077	Hannan-Quinn criter.	-0.394507
F-statistic	297.6627	Durbin-Watson stat	1.886057
Prob(F-statistic)	0.000000		

Dependent Variable: LOG_RES_OLD2
Method: Least Squares
Sample: 1 474
Included observations: 474

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-4.733237	0.123460	-38.33819	0.0000
DUMJCAT2	-0.289197	0.469221	-0.616335	0.5380
DUMJCAT3	0.460492	0.284800	1.616892	0.1066

R-squared	0.006882	Mean dependent var	-4.668104
Adjusted R-squared	0.002665	S.D. dependent var	2.355372
S.E. of regression	2.352231	Akaike info criterion	4.554914
Sum squared resid	2606.038	Schwarz criterion	4.581251
Log likelihood	-1076.515	Hannan-Quinn criter.	4.565272
F-statistic	1.632002	Durbin-Watson stat	1.944100
Prob(F-statistic)	0.196641		

Keeping in mind the correction factor for multiplicative models (assuming that ε_i has a normal distribution), we estimate the variances as

$$\hat{\sigma}_i^2 = \exp(1.27 + \hat{\gamma}_1 + \hat{\gamma}_2 D_{2i} + \hat{\gamma}_3 D_{3i}),$$

so that

$$\begin{aligned}
\hat{\sigma}_1^2 &= \exp(1.27 + \hat{\gamma}_1), \\
\hat{\sigma}_2^2 &= \exp(1.27 + \hat{\gamma}_1 + \hat{\gamma}_2), \\
\hat{\sigma}_3^2 &= \exp(1.27 + \hat{\gamma}_1 + \hat{\gamma}_3).
\end{aligned}$$

Plugging in the obtained estimates, we obtain:

$$\begin{aligned}
\hat{\sigma}_1^2 &= \exp(1.27 - 4.7332) = 0.0313, \\
\hat{\sigma}_2^2 &= \exp(1.27 - 4.7332 - 0.2892) = 0.0235, \\
\hat{\sigma}_3^2 &= \exp(1.27 - 4.7332 + 0.4605) = 0.0497,
\end{aligned}$$

which gives us the required standard deviations per job category:

$$\begin{aligned}
\hat{\sigma}_1 &= \sqrt{\hat{\sigma}_1^2} = 0.1769, \\
\hat{\sigma}_2 &= \sqrt{\hat{\sigma}_2^2} = 0.1532, \\
\hat{\sigma}_3 &= \sqrt{\hat{\sigma}_3^2} = 0.2228.
\end{aligned}$$

As expected, the standard deviation is smallest for custodial jobs and it is largest for management jobs. Notice, however, that the estimates $\hat{\gamma}_2$ and $\hat{\gamma}_3$ are not significant, indicating that the homoskedasticity of the error cannot be rejected.

Next, we run WLS with weights equal to the inverse of the fitted standard deviation.

Dependent Variable: LOGSALARY
Method: Least Squares
Sample: 1 474
Included observations: 474
Weighting series: 1/STDEV_FITTED
Weight type: Inverse standard deviation (EViews default scaling)

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	9.594902	0.052131	184.0539	0.0000
EDUC	0.042693	0.004123	10.35597	0.0000
GENDER	0.178160	0.020345	8.757099	0.0000
MINORITY	-0.078365	0.021330	-3.674013	0.0003
DUMJCAT2	0.167288	0.037542	4.456083	0.0000
DUMJCAT3	0.545052	0.032882	16.57581	0.0000

Weighted Statistics

R-squared	0.716557	Mean dependent var	10.33140
Adjusted R-squared	0.713529	S.D. dependent var	0.778134
S.E. of regression	0.191905	Akaike info criterion	-0.451050
Sum squared resid	17.23537	Schwarz criterion	-0.398377
Log likelihood	112.8989	Hannan-Quinn criter.	-0.430334
F-statistic	236.6254	Durbin-Watson stat	1.886442
Prob(F-statistic)	0.000000	Weighted mean dep.	10.31027

Unweighted Statistics

R-squared	0.760690	Mean dependent var	10.35679
Adjusted R-squared	0.758133	S.D. dependent var	0.397334
S.E. of regression	0.195409	Sum squared resid	17.87038
Durbin-Watson stat	1.891828		

We can see that the outcomes are quite close to those of OLS, so that the effect of heteroskedasticity is relatively small (which is in line with the fact that we did not reject the null of homoskedastic error term).

(ii) Next, adjust the model for the variances as follows:

$$\mathbb{E}[\varepsilon_i^2] = \gamma_1 + \gamma_2 D_2 + \gamma_3 D_3 + \gamma_4 x_i + \gamma_5 x_i^2,$$

i.e. the model for the variances is additive and contains also effects of the level of education. Estimate the eleven parameters (six regression parameters and five variance parameters) by (two-step) FWLS and compare the outcomes with the results in (i).

Dependent Variable: RES_OLD2
Method: Least Squares
Sample: 1 474
Included observations: 474

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.016276	0.053297	0.305388	0.7602
DUMJCAT2	-0.012381	0.013621	-0.908991	0.3638
DUMJCAT3	0.008538	0.011506	0.742033	0.4584
EDUC	0.000506	0.008329	0.060741	0.9516
EDUC^2	7.24E-05	0.000325	0.223071	0.8236

R-squared	0.025792	Mean dependent var	0.037688
Adjusted R-squared	0.017483	S.D. dependent var	0.065791
S.E. of regression	0.065213	Akaike info criterion	-2.611815
Sum squared resid	1.994549	Schwarz criterion	-2.567921
Log likelihood	624.0003	Hannan-Quinn criter.	-2.594552
F-statistic	3.104203	Durbin-Watson stat	1.902122
Prob(F-statistic)	0.015377		

Dependent Variable: LOGSALARY
Method: Least Squares
Sample: 1 474
Included observations: 474
Weighting series: 1/STDEV_FITTED_EDU
Weight type: Inverse standard deviation (EViews default scaling)

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	9.632344	0.047967	200.8111	0.0000
EDUC	0.039311	0.003885	10.11958	0.0000
GENDER	0.181978	0.020253	8.985090	0.0000
MINORITY	-0.067395	0.020538	-3.281424	0.0011
DUMJCAT2	0.178342	0.032217	5.535650	0.0000
DUMJCAT3	0.559036	0.032881	17.00192	0.0000

Weighted Statistics

R-squared	0.720268	Mean dependent var	10.32242
Adjusted R-squared	0.717280	S.D. dependent var	1.568529
S.E. of regression	0.188043	Akaike info criterion	-0.491719
Sum squared resid	16.54849	Schwarz criterion	-0.439045
Log likelihood	122.5373	Hannan-Quinn criter.	-0.471003
F-statistic	241.0064	Durbin-Watson stat	1.908193
Prob(F-statistic)	0.000000	Weighted mean dep.	10.29357

Unweighted Statistics

R-squared	0.759814	Mean dependent var	10.35679
Adjusted R-squared	0.757248	S.D. dependent var	0.397334
S.E. of regression	0.195766	Sum squared resid	17.93579
Durbin-Watson stat	1.901450		

With the additive model we now estimate the variances as

$$\hat{\sigma}_i^2 = \hat{\gamma}_1 + \hat{\gamma}_2 D_{2i} + \hat{\gamma}_3 D_{3i} + \hat{\gamma}_4 x_i + \hat{\gamma}_5 x_i^2,$$

so that

$$\begin{aligned}
\hat{\sigma}_1^2 &= \hat{\gamma}_1 + \hat{\gamma}_4 x_i + \hat{\gamma}_5 x_i^2, \\
&= 0.0163 + 0.0005 x_i + 7\text{e-}05 x_i^2, \\
\hat{\sigma}_2^2 &= \hat{\gamma}_1 + \hat{\gamma}_2 + \hat{\gamma}_4 x_i + \hat{\gamma}_5 x_i^2 \\
&= 0.0163 - 0.0124 + 0.0005 x_i + 7\text{e-}05 x_i^2, \\
\hat{\sigma}_3^2 &= \hat{\gamma}_1 + \hat{\gamma}_3 + \hat{\gamma}_4 x_i + \hat{\gamma}_5 x_i^2 \\
&= 0.0163 + 0.0085 + 0.0005 x_i + 7\text{e-}05 x_i^2.
\end{aligned}$$

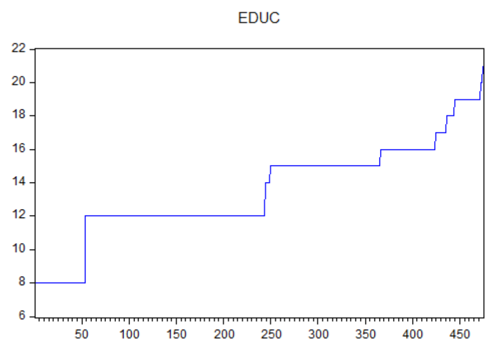
Notice that this time we cannot obtain standard deviations per job category, because the estimates of standard deviation are individual specific (depending on the education level). However, the estimates $\hat{\gamma}_2$–$\hat{\gamma}_5$ are not significant, indicating that again the homoskedasticity of the error cannot be rejected.

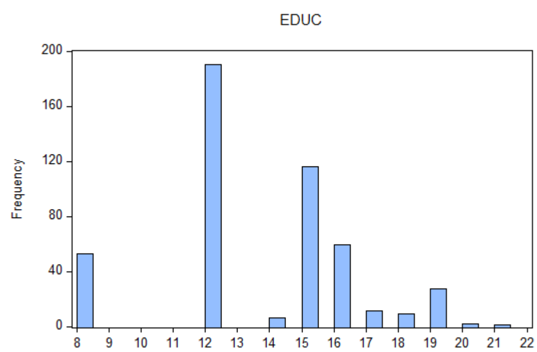
Below we sum up the three sets of standard errors.

		Standard errors		
Variable	$\hat{\beta}_k$	OLS	FWLS no x_i	FWLS with x_i
C	9.574694	0.054218	0.052131	0.047967
EDUC	0.044192	0.004285	0.004123	0.003885
GENDER	0.178340	0.020962	0.020345	0.020253
MINORITY	-0.074858	0.022459	0.021330	0.020538
DUMJCAT2	0.170360	0.043494	0.037542	0.032217
DUMJCAT3	0.539075	0.030213	0.032882	0.032881

We can see that changing of the model for heteroskedasticity does not have a big impact on the results, which are similar to those from *(i)*. Nevertheless, the "additive" FWLS estimator including the education effect is somewhat more accurate than the "multiplicative", job-category-only FWLS estimator, which is a bit more accurate than the OLS one.

(iii) *Check that the data in the data file are sorted with increasing values of x_i. Inspect the histogram of x_i and choose two subsamples to perform the Goldfeld–Quandt test*[3] *on possible heteroskedasticity due to the variable x_i.*





Based on the plots of x_i above we choose $x_i \leq 12$ as the first group and $x_i >= 15$ as the second group, so that both groups are large enough and so that there are some observations dropped with $12 < x_i < 15$ (a few ones with $x_i = 14$). This results in $n_1 = 241$, $n_2 = 225$ and $n_3 = n - n_1 - n_2 = 8$[4].

Dependent Variable: LOGSALARY
Method: Least Squares
Sample: 1 474 IF EDUC<=12
Included observations: 243

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	9.766853	0.075173	129.9256	0.0000
EDUC	0.026684	0.006587	4.050890	0.0001
GENDER	0.172143	0.025703	6.697433	0.0000
MINORITY	-0.069209	0.024714	-2.800447	0.0055
DUMJCAT2	0.172763	0.039865	4.333729	0.0000
DUMJCAT3	0.802059	0.166775	4.809218	0.0000

R-squared	0.379846	Mean dependent var	10.12726
Adjusted R-squared	0.366762	S.D. dependent var	0.206856
S.E. of regression	0.164608	Akaike info criterion	-0.746115
Sum squared resid	6.421724	Schwarz criterion	-0.659867
Log likelihood	96.65298	Hannan-Quinn criter.	-0.711375
F-statistic	29.03258	Durbin-Watson stat	2.023838
Prob(F-statistic)	0.000000		

Dependent Variable: LOGSALARY
Method: Least Squares
Sample: 1 474 IF EDUC>=15
Included observations: 225

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	8.986274	0.213577	42.07501	0.0000
EDUC	0.083984	0.014012	5.993834	0.0000
GENDER	0.163522	0.034615	4.723966	0.0000
MINORITY	-0.080841	0.040897	-1.976695	0.0493
DUMJCAT2	-0.230489	0.221845	-1.038965	0.3000
DUMJCAT3	0.448859	0.042767	10.49538	0.0000

R-squared	0.721045	Mean dependent var	10.60491
Adjusted R-squared	0.714676	S.D. dependent var	0.409211
S.E. of regression	0.218583	Akaike info criterion	-0.176998
Sum squared resid	10.46349	Schwarz criterion	-0.085902
Log likelihood	25.91226	Hannan-Quinn criter.	-0.140231
F-statistic	113.2145	Durbin-Watson stat	1.739112
Prob(F-statistic)	0.000000		

Running the original regression (with $k = 5$) on both subsamples yields $SSR_1 = 6.4217$ and $SSR_2 = 10.4635$, so that we obtain

$$F = \frac{\frac{SSR_2}{n_2-k}}{\frac{SSR_1}{n_1-k}} = \frac{10.4635}{6.4217} \cdot \frac{241-5}{225-5} = 1.7627$$

[3] Since the Goldfeld-Quandt has not been in the lecture slides, it will be explained during the tutorial.

[4] You can use the commands `smpl if educ<=12, scalar n1 = @obssmpl` and `smpl if educ>=15, scalar n2 = @obssmpl` in EViews.

(with the exact values in EViews, with the above ones it is 1.7479), which under the null of homosekdasticity follows the $F(n_2 = k, n_1 - k) = F(225 - 5, 241 - 5) = F(220, 236)$ distribution. The corresponding p-value is 9.76E-06 so virtually 0. Hence, at any reasonable significance level we reject the null of homoskedasticity and conclude that there is evidence for heteroskedasticy due to the education level.

(iv) Perform the Breusch–Pagan test on heteroskedasticity, using the specified model for the variances.

We still use the additive model for the variances from *(ii)*, i.e. we consider R^2 from the auxiliary regression from *(ii)*

$$\hat{\varepsilon}_i^2 = \gamma_1 + \gamma_2 D_2 + \gamma_3 D_3 + \gamma_4 x_i + \gamma_5 x_i^2 + \eta_i.$$

With $R^2 = 0.0258$, the obtained value of the LM statistic is

$$LM = nR^2 = 474 \cdot 0.0258 = 12.2255,$$

with the corresponding p-value of 0.0157 (we use the χ^2_4 distribution). Hence, at the standard significance level of 5% we can reject the null of homoskedasticity.

Alternatively, we can run the built-in test in EViews, where we need to adjust the regressors in the test specification box, which leads to the same results.

Heteroskedasticity Test: Breusch-Pagan-Godfrey

F-statistic	3.104203	Prob. F(4,469)	0.0154
Obs*R-squared	12.22552	Prob. Chi-Square(4)	0.0158
Scaled explained SS	18.12103	Prob. Chi-Square(4)	0.0012

Test Equation:
Dependent Variable: RESID^2
Method: Least Squares
Sample: 1 474
Included observations: 474

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.016276	0.053297	0.305388	0.7602
DUMJCAT2	-0.012381	0.013621	-0.908991	0.3638
DUMJCAT3	0.008538	0.011506	0.742033	0.4584
EDUC	0.000506	0.008329	0.060741	0.9516
EDUC^2	7.24E-05	0.000325	0.223071	0.8236

R-squared	0.025792	Mean dependent var	0.037688
Adjusted R-squared	0.017483	S.D. dependent var	0.065791
S.E. of regression	0.065213	Akaike info criterion	-2.611815
Sum squared resid	1.994549	Schwarz criterion	-2.567921
Log likelihood	624.0003	Hannan-Quinn criter.	-2.594552
F-statistic	3.104203	Durbin-Watson stat	1.902122
Prob(F-statistic)	0.015377		

(v) Also perform the White test on heteroskedasticity.

The results for the White test without and with cross terms, respectively, are shown below.

Heteroskedasticity Test: White

F-statistic	2.656429	Prob. F(5,468)	0.0221
Obs*R-squared	13.08118	Prob. Chi-Square(5)	0.0226
Scaled explained SS	19.38931	Prob. Chi-Square(5)	0.0016

Test Equation:
Dependent Variable: RESID^2
Method: Least Squares
Sample: 1 474
Included observations: 474

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.020947	0.009810	2.135169	0.0333
EDUC^2	9.53E-05	5.60E-05	1.702415	0.0893
GENDER^2	-0.001069	0.007027	-0.152050	0.8792
MINORITY^2	-0.006732	0.007498	-0.897810	0.3697
DUMJCAT2^2	-0.010073	0.014419	-0.698559	0.4852
DUMJCAT3^2	0.006985	0.010588	0.659719	0.5098

R-squared	0.027597	Mean dependent var	0.037688
Adjusted R-squared	0.017208	S.D. dependent var	0.065791
S.E. of regression	0.065222	Akaike info criterion	-2.609451
Sum squared resid	1.990853	Schwarz criterion	-2.556777
Log likelihood	624.4398	Hannan-Quinn criter.	-2.588735
F-statistic	2.656429	Durbin-Watson stat	1.910577
Prob(F-statistic)	0.022111		

Heteroskedasticity Test: White

F-statistic	2.117199	Prob. F(14,459)	0.0101
Obs*R-squared	28.75268	Prob. Chi-Square(14)	0.0113
Scaled explained SS	42.61808	Prob. Chi-Square(14)	0.0001

Test Equation:
Dependent Variable: RESID^2
Method: Least Squares
Sample: 1 474
Included observations: 474
Collinear test regressors dropped from specification

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.131454	0.071827	1.830155	0.0679
EDUC^2	0.000884	0.000492	1.797828	0.0729
EDUC*GENDER	0.002489	0.003336	0.746059	0.4560
EDUC*MINORITY	-0.002532	0.003481	-0.727354	0.4674
EDUC*DUMJCAT2	0.004829	0.006653	0.725860	0.4683
EDUC*DUMJCAT3	-0.018342	0.006958	-2.636092	0.0087
EDUC	-0.018886	0.011890	-1.588350	0.1129
GENDER^2	-0.037490	0.044811	-0.836627	0.4032
GENDER*MINORITY	-0.002821	0.016624	-0.169688	0.8653
GENDER*DUMJCAT2	-0.062739	0.074654	-0.840391	0.4011
GENDER*DUMJCAT3	0.021593	0.026254	0.822459	0.4112
MINORITY^2	0.021593	0.044828	0.481694	0.6303
MINORITY*DUMJCAT2	0.022435	0.029892	0.750541	0.4533
MINORITY*DUMJCAT3	0.081214	0.037031	2.193140	0.0288
DUMJCAT3^2	0.273542	0.109895	2.489112	0.0132

R-squared	0.060660	Mean dependent var	0.037688
Adjusted R-squared	0.032009	S.D. dependent var	0.065791
S.E. of regression	0.064729	Akaike info criterion	-2.606068
Sum squared resid	1.923163	Schwarz criterion	-2.474384
Log likelihood	632.6381	Hannan-Quinn criter.	-2.554279
F-statistic	2.117199	Durbin-Watson stat	1.953388
Prob(F-statistic)	0.010144		

The LM statistic for the White test without cross terms is equal to 13.0811 and under the null it follows the χ^2_5 distribution. The corresponding p-value is 0.0226. For the White test with cross terms we obtain $LM = 28.7527$, which follows the χ^2_{14} distribution under the null and yields the p-value of 0.0113. Either way we can reject the null of homoskedasticity at the standard significance level of 5% .

(vi) Comment on the similarities and differences between the test outcomes in (iii)–(v).

The main similarity is that all three tests rejected the null of homoskedasticity, hence we have strong grounds to claim that the variance of the unobserved factors changes across different segments of the analysed data.

A difference is the exact level of the p-value: some tests may have more power to detect heteroskedasticity for this dataset (and reject H_0 more clearly with a lower p-value).

Another difference is that the Goldfeld-Quandt test assumes that the errors are normally distributed, whereas the Breusch-Pagan and White tests do not rely on this assumption.