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):
  \[ \text{Var}(u_i|x_{i1}, \ldots, x_{ik}) = \sigma^2. \]

- **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
  \[ \widehat{\text{Var}}(\hat{\beta}_{OLS}) = (X'X)^{-1} X'\hat{\Omega}X (X'X)^{-1}, \]
  
  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 \( \sigma_1^2 \));
  2. the second group of the last \( n_2 \) observations (with variance \( \sigma_2^2 \));
  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
  
  \[
  H_0 : \quad \sigma_1^2 = \sigma_2^2, \\
  H_1 : \quad \sigma_1^2 < \sigma_2^2.
  \]

- 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:
  
  \[
  \frac{SSR_j}{\sigma_j^2} \sim \chi^2_{n_j-k}, \quad j = 1, 2,
  \]
  
  and these two statistics are independent.

- Therefore:
  
  \[
  \frac{SSR_2}{(n_2-k)\sigma_2^2} = \frac{s_2^2}{\sigma_2^2} \sim F(n_2 - k, n_1 - k).
  \]

- 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
  \[
  \text{Var}(u_i|x_i) = \sigma^2 \exp (\delta_0 + \delta_1 x_{i1} + \cdots + \delta_k x_{ik}) ;
  \]

- **additive** heteroskedasticity model
  \[
  \text{Var}(u_i|x_i) = \delta_0 + \delta_1 x_{i1} + \cdots + \delta_k x_{ik}.
  \]

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

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

\[
\begin{align*}
\text{Var}(u_i|x_i) & \overset{E(u_i|x_i)=0}{=} E(u_i^2|x_i) \\
& = \sigma^2 \exp (\delta_0 + \delta_1 x_{i1} + \cdots + \delta_k x_{ik}) ,
\end{align*}
\]

so it is equivalent with

\[
\begin{align*}
u_i^2 &= \sigma^2 \exp (\delta_0 + \delta_1 x_{i1} + \cdots + \delta_k x_{ik}) v_i , \\
v_i &= \frac{u_i^2}{E(u_i^2|x_i)} \quad (⇐ \text{mean 1 random variable})
\end{align*}
\]

Hence, we consider

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

where \( \eta_i \) is the error term

\[
\eta_i = \log(v_i) - E(\log(v_i))
\]

and \( \alpha_0 \) is a constant term

\[
\alpha_0 = \log(\sigma^2) + \delta_0 + E(\log(v_i)).
\]

Hence, the coefficient \( \delta_0 \) 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 \( \delta_0 \) is then estimated by

\[
\hat{\delta}_0 + a ,
\]

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

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

We will see how this works in Computer Exercise 2(i)\(^1\).

\(^1\)Note, however, that a consistent estimator of \( \delta_0 \) 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 $\text{Var}(u|X) = \Omega \neq \sigma^2 I$ we have for $\hat{\beta}_{OLS} = \beta + (X'X)^{-1} X'u$:

$$\text{plim} \left( \hat{\beta}_{OLS} \right) = \beta + \text{plim} \left( \frac{1}{n} \sum_{i=1}^{n} x_i x_i' \right)^{-1} \text{plim} \left( \frac{1}{n} \sum_{i=1}^{n} x_i u_i \right)$$

$$= \beta + \mathbb{E}(X'X)^{-1} \frac{\mathbb{E}(X'u)}{\mathbb{E}(X'E(u|X))=0},$$

so the OLS estimator is still consistent.

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

Now, we have

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

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:

$$\text{beer} = \beta_0 + \beta_1 \text{inc} + \beta_2 \text{price} + \beta_3 \text{educ} + \beta_4 \text{female} + u,$$

$$\mathbb{E}(u|\text{inc}, \text{price}, \text{educ}, \text{female}) = 0,$$

$$\text{Var}(u|\text{inc}, \text{price}, \text{educ}, \text{female}) = \sigma^2 \text{inc}^2.$$

Write the transformed equation that has a homoskedastic error term.

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

$$\frac{\text{beer}}{\text{inc}} = \frac{\beta_0}{\text{inc}} + \frac{\beta_1}{\text{inc}} \text{inc} + \frac{\beta_2}{\text{inc}} \text{price} \text{inc} + \frac{\beta_3}{\text{inc}} \text{educ} \text{inc} + \frac{\beta_4}{\text{inc}} \text{female} \text{inc} + \frac{u}{\text{inc}}$$

$$= \beta_0 \frac{1}{\text{inc}} + \beta_1 + \beta_2 \frac{\text{price}}{\text{inc}} + \beta_3 \frac{\text{educ}}{\text{inc}} + \beta_4 \frac{\text{female}}{\text{inc}} + \frac{u}{\text{inc}}.$$

Notice that $\beta_1$, which is the slope on $\text{inc}$ 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* run the regression

\[ y_i = \beta_1 d_{1i} + \beta_2 d_{2i} + \beta_3 d_{3i} + u_i \]  

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 nR² 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 heteroskedasticity test (Breusch-Pagan-Godfrey) in EViews:

- 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 \( \chi^2_2 = \chi^2_3 \), 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 \( \beta \):

\[
\begin{align*}
  b_1 &= \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}, \\
  b_2 &= \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i}, \\
  b_3 &= \frac{1}{n} \sum_{i=1}^{n} \frac{y_i}{x_i}.
\end{align*}
\]

\(^2\)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 $\sigma_h^2$ for which this estimator is the best linear unbiased estimator of $\beta$.

Recall that when we have a model for heteroskedasticity, i.e. in $\text{Var}(u_i|x_i) = \sigma^2 h(x_i)$ 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

$$y_i = \beta x_i + \varepsilon_i,$$

$$\frac{y_i}{\sqrt{h_i}} = \frac{\beta x_i}{\sqrt{h_i}} + \frac{\varepsilon_i}{\sqrt{h_i}},$$

so that the corresponding OLS estimator is

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

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 $\text{Var}(u_i|x_i) = \sigma^2$. Notice that this is simply the OLS estimator for the homoskedastic case.

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

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

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

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

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

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

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

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

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 $\eta_i$ be a random drawing from the distribution $N(0, x_i^2)$, 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 $\beta$ by OLS. Compute the homoskedasticity-only standard error of $\hat{\beta}_{OLS}$ and the White heteroskedasticity-robust standard error of $\hat{\beta}_{OLS}$.

(ii) Estimate $\beta$ by WLS using the knowledge that $\sigma^2_i = \sigma^2 x_i^2$. 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:

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

so that now the transformed error terms $\varepsilon_i^*$ 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.

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.
(iii) Now estimate \( \beta \) by WLS using the (incorrect) heteroskedasticity model \( \sigma_i^2 = x_i^2 \). 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:

\[
y_i^{**} := y_i x_i, \quad x_i^{**} := x_i x_i = x_i^2, \quad \varepsilon_i^{**} := \varepsilon_i x_i,
\]

so that now the transformed error terms \( \varepsilon_i^{**} \) 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.

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

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

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

\[
\hat{\beta}_{WLS} = \frac{\sum_{i=1}^{n} x_i^3 y_i}{\sum_{i=1}^{n} x_i^4} = \frac{\sum_{i=1}^{n} x_i^2 (\beta x_i + \varepsilon_i)}{\sum_{i=1}^{n} x_i^4} = \beta + \frac{\sum_{i=1}^{n} x_i^2 \varepsilon_i}{\sum_{i=1}^{n} x_i^4}.
\]
Because $\hat{\beta}_{WLS}$ is unbiased, i.e. $E(\hat{\beta}_{WLS} \mid x) = \beta$, the variance of $\hat{\beta}_{WLS}$ is

$$\text{Var}(\hat{\beta}_{WLS} \mid x) = E \left[ \left( \hat{\beta}_{WLS} - E(\hat{\beta}_{WLS} \mid x) \right)^2 \right]$$

$$= E \left[ \left( \beta + \frac{\sum_{i=1}^{n} x_i^3 \varepsilon_i}{\sum_{i=1}^{n} x_i^4} - \beta \right)^2 \right]$$

$$= E \left[ \left( \frac{\sum_{i=1}^{n} x_i^3 \varepsilon_i}{\sum_{i=1}^{n} x_i^4} \right)^2 \right]$$

$$= \sum_{i=1}^{n} x_i^4 E \left[ \varepsilon_i^2 \mid x_i \right]$$

$$= \sum_{i=1}^{n} x_i^4 \text{Var}[\varepsilon_i \mid x_i]$$

$$= \sum_{i=1}^{n} x_i^8$$,

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

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

$$\text{Var}(\hat{\beta}_{WLS} \mid x) = \frac{9962.1182}{(318.3814)^2} = 0.0983,$$

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 $\eta_i$ are different drawings from the $N(0, x_i^2)$ 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 $\beta$ 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 $\beta$ using the required three methods are as follows:

<table>
<thead>
<tr>
<th>Method</th>
<th>Homosked. only</th>
<th>Heterosked. robust</th>
<th>Simulation st. deviations</th>
</tr>
</thead>
<tbody>
<tr>
<td>OLS</td>
<td>0.0956</td>
<td>0.1597</td>
<td>0.1799</td>
</tr>
<tr>
<td>WLS correct</td>
<td>0.0989</td>
<td>0.0989</td>
<td>0.0972</td>
</tr>
<tr>
<td>WLS incorrect</td>
<td>0.0895</td>
<td>0.2296</td>
<td>0.3155</td>
</tr>
</tbody>
</table>

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 $\beta$ 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.
Figure 1: EViews code example for Computer Exercise 1.
Exercise 2

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

\[ y_i = \beta_1 + \beta_2 x_i + \beta_3 D_g + \beta_4 D_m + \beta_5 D_2 + \beta_6 D_3 + \epsilon_i, \]

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:

\[ \sigma^2_i = E[\epsilon^2_i] = \epsilon_1 + \gamma_2 D_2 + \gamma_3 D_3. \]

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 \( \log(\hat{\epsilon}^2_i) \), where \( \hat{\epsilon}_i \) are the OLS residuals of from the first regression.

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

\[ \hat{\sigma}^2_i = \exp(1.27 + \hat{\gamma}_1 + \hat{\gamma}_2 D_2 + \hat{\gamma}_3 D_3), \]

so that

\[ \hat{\sigma}_1^2 = \exp(1.27 + \hat{\gamma}_1), \]

\[ \hat{\sigma}_2^2 = \exp(1.27 + \hat{\gamma}_1 + \hat{\gamma}_2), \]

\[ \hat{\sigma}_3^2 = \exp(1.27 + \hat{\gamma}_1 + \hat{\gamma}_3). \]

Plugging in the obtained estimates, we obtain:

\[ \hat{\sigma}_1^2 = \exp(1.27 - 4.7332) = 0.0313, \]

\[ \hat{\sigma}_2^2 = \exp(1.27 - 4.7332 - 0.2892) = 0.0235, \]

\[ \hat{\sigma}_3^2 = \exp(1.27 - 4.7332 + 0.4605) = 0.0497, \]

which gives us the required standard deviations per job category:

\[ \hat{\sigma}_1 = \sqrt{\hat{\sigma}_1^2} = 0.1769, \]

\[ \hat{\sigma}_2 = \sqrt{\hat{\sigma}_2^2} = 0.1532, \]

\[ \hat{\sigma}_3 = \sqrt{\hat{\sigma}_3^2} = 0.2228. \]

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.
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:

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

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).

With the additive model we now estimate the variances as

\[ \sigma_i^2 = \hat{\gamma}_1 + \hat{\gamma}_2 D_2 + \hat{\gamma}_3 D_3 + \hat{\gamma}_4 x_i + \hat{\gamma}_5 x_i^2, \]

so that

\[ \sigma_1^2 = \hat{\gamma}_1 + \hat{\gamma}_4 x_i + \hat{\gamma}_5 x_i^2, \]

\[ = 0.0163 + 0.0005x_i + 7e-05x_i^2, \]

\[ \sigma_2^2 = \hat{\gamma}_1 + \hat{\gamma}_2 + \hat{\gamma}_4 x_i + \hat{\gamma}_5 x_i^2, \]

\[ = 0.0163 - 0.0124 + 0.0005x_i + 7e-05x_i^2, \]

\[ \sigma_3^2 = \hat{\gamma}_1 + \hat{\gamma}_3 + \hat{\gamma}_4 x_i + \hat{\gamma}_5 x_i^2, \]

\[ = 0.0163 + 0.0085 + 0.0005x_i + 7e-05x_i^2. \]
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.

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

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 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 $x_i \geq 15$ 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 $n_3 = n - n_1 - n_2 = 8^4$.

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

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

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

4You 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 homoskedasticity 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 heteroskedasticity 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)

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

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 $\chi^2$ 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.

**(v) Also perform the White test on heteroskedasticity.**

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

The LM statistic for the White test without cross terms is equal to $13.0811$ and under the null it follows the $\chi^2_5$ distribution. The corresponding $p$-value is $0.0226$. For the White test with cross terms we obtain $LM = 28.7527$, which follows the $\chi^2_{14}$ 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%.
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.