exam1

Mathias Kold

2024-04-18

Exam 1 - OLS and heteroskedasticity

In a multiple linear regression (MLR) there are 6 assumptions. Assumption 1-5 are called Gauss-Markov assumptions and assumption 6 is called the normality assumption. The first 4 assumptions exist to secure that the model is unbiased, while assumption 5 checks for heteroskedasticity and assumption 6 checks for normality in the model. The assumptions are:

- **MLR1: Linear in Parameters**

The model in the population can be written as

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + u$$

where $\beta_0, \beta_1, \ldots, \beta_k$ are the unknown parameters of interest and u is an unobserved random error or disturbance term.

MLR2: Random Sampling

We have a random sample of n observations, $\{(x_{i1}, x_{i2}, \ldots, x_{ik}, y_i) : i = 1, 2, \ldots, n\}$, following the population model in Assumption MLR.1.

MLR3: No Perfect Collinearity

In the sample (and therefore in the population), none of the independent variables is constant, and there are no exact linear relationships among the independent variables.

MLR4: Zero Conditional Mean

The error u has an expected value of zero given any values of the independent variables. In other words,

$$E(u|x_1, x_2, \ldots, x_k) = 0$$

MLR5: Homoskedasticity

The error u has the same variance given any value of the explanatory variables. In other words,

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

MLR6: Normality

The population error u is independent of the explanatory variables $x_1, x_2, \ldots, x_k$ and is normally distributed with zero mean and variance σ^2: $u \sim \text{Normal}(0, \sigma^2)$.

Look at the following model for bank employees wage:

$$log(salary) = \beta_0 + \beta_1 educ + \beta_2 log(salbegin) + \beta_3 male + \beta_4 minority + u$$

where salary is yearly wage (in 1000 US dollars), educ is education measured in number of years, salbegin is the starting salary (in 1000 US dollars) for the person's first position in the same bank, male is a dummy variable for gender, minority is one dummy variable indicating whether one belongs to a minority.

1 - Estimate the model using OLS. Comment on the output and interpret the results

```
options(scipen = 999)
library(readr)
library(foreign)
library(car)
```

```
## Loading required package: carData
```

```
library(sandwich)
library(lmtest)
```

```
## Loading required package: zoo
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
library(texreg)
```

```
## Version:  1.39.3
## Date:     2023-11-09
## Author:   Philip Leifeld (University of Essex)
##
## Consider submitting praise using the praise or praise_interactive functions.
## Please cite the JSS article in your publications -- see citation("texreg").
```

```
data1 <- read_csv("data1.csv")
```

```
## Rows: 474 Columns: 10
```

```
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## dbl (10): obs, idnumber, salary, lsalary, educ, salbegin, lsalbegin, male, m...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
model <- lm(log(salary) ~ educ + log(salbegin) + male + minority, data = data1)
summary(model)
```

```
##
## Call:
## lm(formula = log(salary) ~ educ + log(salbegin) + male + minority,
```

```
##     data = data1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.45572 -0.11508 -0.00516  0.10765  0.87060
##
## Coefficients:
##               Estimate Std. Error t value             Pr(>|t|)
## (Intercept)    0.84868    0.07512  11.298 < 0.0000000000000002 ***
## educ           0.02327    0.00387   6.013        0.00000000366 ***
## log(salbegin)  0.82180    0.03603  22.808 < 0.0000000000000002 ***
## male           0.04816    0.01991   2.419               0.0160 *
## minority      -0.04237    0.02034  -2.083               0.0378 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1766 on 469 degrees of freedom
## Multiple R-squared:  0.8041, Adjusted R-squared:  0.8024
## F-statistic: 481.3 on 4 and 469 DF,  p-value: < 0.00000000000000022
```

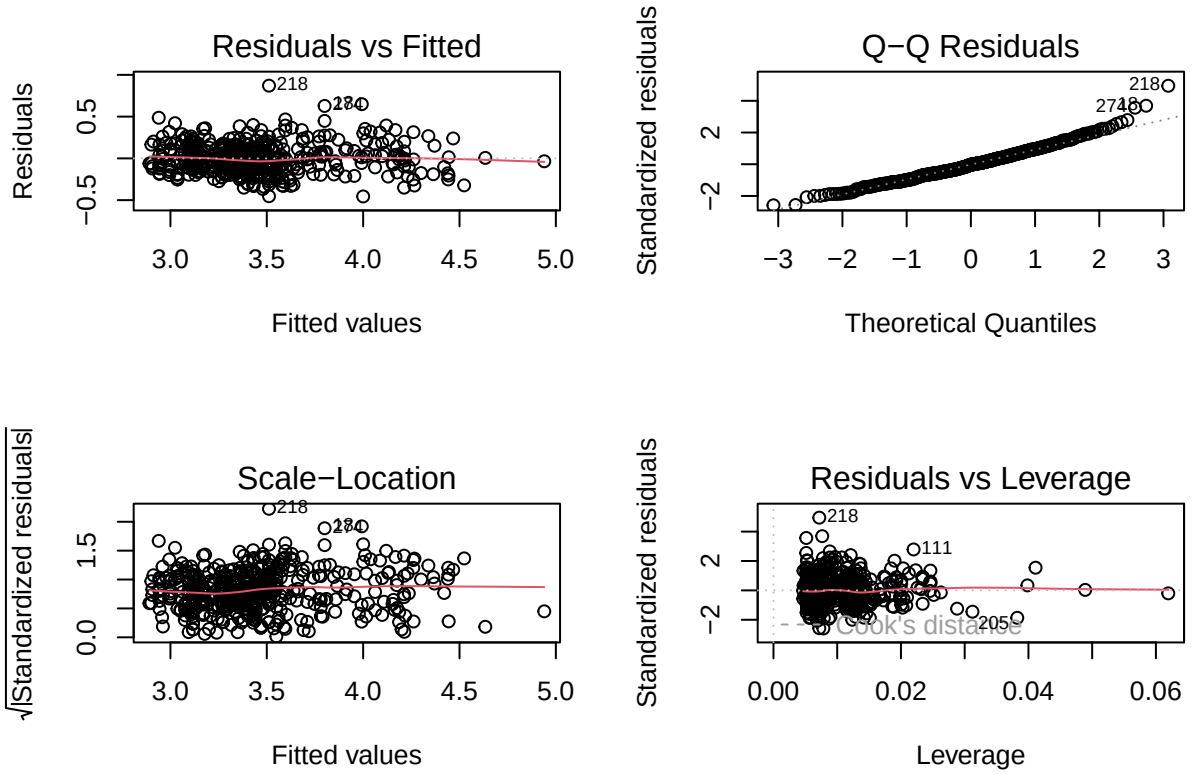
From the estimated model it can be seen that the education variable has a value of 0.02327 which means that one extra year of education will raise the salary by approximately 2.3%. The male variable has a value 0.04816 which means that being a man will raise your salary by approximately 4.8%. The minority variable has a value of -0.04237 which means that minorities approximately will have a 4.2% lower salary than people who are not minorities. The salbegin variable has a different interpretation because it is in log form. It has a value of 0.82180 which means that an increase of 1% in your starting salary will make your salary approximately 0.82% higher.

The intercept is 0.84868 which is the value of the dependent variable, in this case salary, when all other variables have a value of 0. So in this case 0.84868 is the expected value of the salary for a person with zero education, without a starting salary and someone who is not a male or a minority.

It can also be seen that all the variables are statistically significant at a 5% significance level due to p-values < 0.05 but only education and log(salbegin) are statistically significant at a 1% significance level.

2 - Perform graphical model checking.

```
par(mfrow = c(2, 2))
plot(model)
```



The residual vs fitted model shows if the residuals have non-linear patterns. If they are equally spread around a horizontal line without any patterns it indicates that the residuals does not have non-linear patterns. From the plot it can be seen that the spread is quite equal around the horizontal red line, although a few outliers exists, which indicates a linear pattern.

The Q-Q residuals plot shows if the residuals are normally distributed. If the residuals fit the dotted line they are normally distributed. In this case, the residuals fit the dotted line except in the last quantiles where they deviate a bit from the dotted line. This could indicate that the residuals are approximately normally distributed although not perfectly normally distributed.

The scale-location plot shows if homoskedasticity exists within the residuals. If the residuals are spread equally around the predictors, the model fulfills the assumption of homoskedasticity. In this case, it can be argued that there is an increasing pattern in the spread of the residuals which indicates heteroskedasticity.

The residuals vs leverage plot helps detect outliers in the model. It is important because outliers can have a major impact in the model. The residuals can be seen if they have a large Cook's distance, which is the dashed line. In this case, it seems like there might be an outlier with observation 218 but other than that, most of the observations does not have a large Cook's distance.

3 - Test for heteroskedasticity using the Breusch-Pagan test and the special edition of the White test.

First, the test for heteroskedasticity can be carried out by using the Breusch-Pagan test:

```
r=residuals(model)
res = r^2
summary(lm(res~educ + log(salbegin) + male + minority, data = data1))
```

```
##
## Call:
## lm(formula = res ~ educ + log(salbegin) + male + minority, data = data1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.04773 -0.02506 -0.01345  0.00908  0.71750
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    0.0036404  0.0235992   0.154    0.877
## educ           0.0019699  0.0012157   1.620    0.106
## log(salbegin) -0.0008061  0.0113201  -0.071    0.943
## male           0.0094827  0.0062553   1.516    0.130
## minority      -0.0104497  0.0063908  -1.635    0.103
##
## Residual standard error: 0.05548 on 469 degrees of freedom
## Multiple R-squared:  0.02923,    Adjusted R-squared:  0.02095
## F-statistic: 3.531 on 4 and 469 DF,  p-value: 0.007475
```

Because the p-value is 0.007475 which is < 0.05, we reject the null hypothesis meaning that there is heteroskedasticity in the model.

We can now calculate the LM test, where we have 474 observations and $R^2 = 0.1231$:

```
LM = 0.1231*474
LM
```

```
## [1] 58.3494
```

We can then calculate the p-value of chi-square χ^2_k:

```
1-pchisq(LM,4)
```

```
## [1] 0.000000000006445178
```

Both the LM and F test reject the null hypothesis meaning there is heteroskedasticity in the model. The Breusch-Pagan test can also be directly run in R using the bptest function which also gives us the p-value.

```
library(lmtest)
bptest(model)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  model
## BP = 13.857, df = 4, p-value = 0.007767
```

The special White test can also be used to do the test the model for heteroskedasticity.

```r
yhat <- fitted(model)
quadu <- (residuals(model)^2)
model_white <- lm(quadu ~ yhat + I(yhat^2))

whitetest <- c("yhat=0", "I(yhat^2)=0")
linearHypothesis(model_white,whitetest)
```

```
## Linear hypothesis test
##
## Hypothesis:
## yhat = 0
## I(yhat^2) = 0
##
## Model 1: restricted model
## Model 2: quadu ~ yhat + I(yhat^2)
##
##   Res.Df    RSS Df Sum of Sq      F   Pr(>F)
## 1    473 1.4873
## 2    471 1.4531  2  0.034139 5.5327 0.004217 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From both the BP test and the special White test it can be seen that the p-value is < 0.05 which means that there is heteroskedasticity in the model.

4 - Calculate robust standard errors for the model and compare with the results in question 1.

Since there is heteroskedasticity in the model, it is necessary to take this into account, when performing a linear regression. A valid estimator of multiple linear regression in the presence of heteroskedasticity is

$$Var(\hat{\beta}_j) = \frac{\sum_{i=0}^{n} \hat{r}_{ij}^2 \cdot \hat{u}_i^2}{(SSR_j)^2}$$

where $\hat{r}_{ij}$ denotes the ith residual from regressing x_j on all other independent variables. The robust standard error is obtained by taking the square root of the above equation. R can calculate the robust standard errors using coeftest():

```r
coef_model <- coeftest(model, vcov=vcovHC(model,type="HC0"))
screenreg(list(OLS=model,Standard_Robust_Error=coef_model), digits=4)
```

```
##
## ================================================
##                OLS           Standard_Robust_Error
## ------------------------------------------------
## (Intercept)     0.8487 ***    0.8487 ***
##                (0.0751)      (0.0794)
## educ            0.0233 ***    0.0233 ***
##                (0.0039)      (0.0035)
## log(salbegin)   0.8218 ***    0.8218 ***
##                (0.0360)      (0.0374)
```

```
## male             0.0482 *     0.0482 *
##                 (0.0199)     (0.0200)
## minority        -0.0424 *    -0.0424 *
##                 (0.0203)     (0.0177)
## ------------------------------------------------
## R^2              0.8041
## Adj. R^2         0.8024
## Num. obs.      474
## ================================================
## *** p < 0.001; ** p < 0.01; * p < 0.05
```

It can be seen that the estimates from the two models are the same, which is expected, because we try to account for the non-constant variance of the error term in the OLS-model. Hence, it is only the standard errors that change. Furthermore, the significance level does not change, when performing a coeftest(), when adjusting for heteroskedasticity. The robust standard errors can be used to perform hypothesis testing and to calculate confidence intervals. Even when adjusting for heteroskedasticity, the conclusion from the regressions does not change, since the estimates are the same, and there are no significant changes.

5 - Test the hypothesis $H_0 : \beta_2 = 1$ against the alternative $H_1 : \beta_2 \neq 1$.

β_2 is the estimate for the starting salary's effect on the the yearly salary, so when we want to test the null hypothesis $H_0 : \beta_2 = 1$ it means that the starting salary has an effect on the yearly salary. (MAYBE COMMENT ON WHAT KIND OF EFFECT!).

To test our null hypothesis against the alternative hypothesis we use the t-statistics, which is given by:

$$t_{\hat{\beta}_j} = \frac{\hat{\beta}_j - 1}{se(\hat{\beta}_j)}$$

We will perform a two-sided test, so the decision rule will be $|t_{\hat{\beta}_j}| > c$, meaning that the null hypothesis will be rejected if the absolute value of t-statistics is greater than the critical value.

To perform the test we use our estimate and standard error for $\hat{\beta}_2$ from 1.1.

```
bhat2 <- 0.82180
se_bhat2 <- 0.03603

t_stat <- (bhat2 - 1) / se_bhat2
t_stat
```

```
## [1] -4.945878
```

Then we need to calculate the critical values for the two-sided test:

```
alpha <- 0.05
c <- qt(1-alpha/2, 469)
c
```

```
## [1] 1.965035
```

```
abs(t_stat)>c
```

```
## [1] TRUE
```

We can see that the absolute value of the t-statistic is greater than the critical value at a 5% significance level, meaning that we reject the null hypothesis, so we instead accept the alternative hypothesis saying that the starting salary does have a statistically significant effect on the yearly salary.

Another way to test the null hypothesis is by calculating the p-value. The definition of the p-value is the probability of obtaining a t-statistic more or as extreme as the one observed in the sample. We use R to calculate the p-value in the following way:

```
pval <- 2*pt(-abs(t_stat), 469)
pval
```

```
## [1] 0.000001057867
```

So at a 5% significance level the p-value also tells us to reject the null hypothesis, further confirming the rejecting from the t-statistics before.

6 - Test the hypothesis $H_0 : \beta_3 = \beta_4 = 0$

When we want to test multiple hypothesis we turn to the F-statistics, where we use the following formula:

$$F = \frac{R_{ur}^2 - R_r^2}{1 - R_{ur}^2} * \frac{n - k - 1}{q}$$

where R_{ur}^2 is the R-squared from our unrestricted model, R_r^2 is the R-squared from the restricted model, q is the difference in the degrees of freedom between the unrestricted and restricted model, n is the number of observations in the dataset and k is the number of independent variables in the unrestricted model.

In our case the unrestricted model is the one given in the beginning of assignment 1.1 with all the independent variables, and in the case where we want to test the null hypothesis $H_0 : \beta_3 = \beta_4 = 0$, we then get the following restricted model:

$$log(salary) = \beta_0 + \beta_1 educ + \beta_2 log(salbegin) + u$$

So we start by estimating our restricted model:

```
model_r <- lm(log(salary) ~ educ + log(salbegin), data = data1)
```

We then obtain our R^2 from the restricted and unrestricted model:

```
r2_ur <- summary(model)$r.squared
r2_r <- summary(model_r)$r.squared
```

We now have what we need to calculate our F-statistic:

```
F <- (r2_ur-r2_r)/(1-r2_ur) * (474-4-1)/2
F
```

```
## [1] 4.234946
```

We then calculate our critical values of the F-statistics at 5% significance level:

```r
qf(0.95, 2, 469)
```

```
## [1] 3.014949
```

Our decision rule says that if F > c then we reject our null hypothesis and instead accept our alternative hypothesis meaning that β_3 and β_4 does have a statistical impact on yearly salary.

It is also possible to do the test directly in R by using the following code:

```r
myh0 <- c("male=0", "minority=0")
linearHypothesis(model, myh0)
```

```
## Linear hypothesis test
##
## Hypothesis:
## male = 0
## minority = 0
##
## Model 1: restricted model
## Model 2: log(salary) ~ educ + log(salbegin) + male + minority
##
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1    471 14.892
## 2    469 14.627  2   0.26416 4.2349 0.01504 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Using the code in R to perform the F-statistics gives us the same answer as before still confirming the rejecting of our null hypothesis.

7 - Estimate the model using FGLS and comment on the results.

Feasible generalised least squares (FGLS) is a method used to adress heteroskedasticity. It can be difficult to know the form of heteroskedasticity (i.e, h(x_i)). In many cases, it is possible to model the function h and use this data to estimate the unknown parameters. By estimating h_i, an estimated value of h is obtained, $\hat{h_i}$. One method to find the exact form of h_i is as follows:

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

where $x_1, x_2, ..., x_k$ are the independent variables in the regression.

Since the parameters δ is unknown for the population, these are estimated using the given data, where $\hat{h}$ can be used to account for heteroskedasticity. To find the estimates for δ, we can use the following:

$$u^2 = \sigma^2 exp(\delta_0 + \delta_1 x_1 + \delta_2 x_2 + ... + \delta_k x_k)v$$

By taking log, the model can be linearised:

$$log(u^2) = a_0 + \delta_1 x_1 + \delta_2 x_2 + ... + \delta_k x_k + e$$

We already know from 1.3 that our model has heteroskedasticity problems, so we obtain our squared residuals from our OLS model and log them:

```
logu2 <- log(resid(model)^2)
```

Now we run the regression on the form mentioned earlier:

```
varreg <- lm(logu2 ~ educ + log(salbegin) + male + minority, data = data1)
```

After we calculate the weights by exponentiating the fitted values from varreg:

```
w <- exp(fitted(varreg))
```

And then we can estimate the FGLS:

```
FGLS <- lm(log(salary) ~ educ + log(salbegin) + male + minority, weight=1/w, data = data1)
screenreg(list(OLS=model, FLGS=FGLS), digits=4)
```

```
##
## ========================================
##                OLS           FLGS
## ----------------------------------------
## (Intercept)     0.8487 ***    0.8493 ***
##                (0.0751)      (0.0756)
## educ            0.0233 ***    0.0222 ***
##                (0.0039)      (0.0038)
## log(salbegin)   0.8218 ***    0.8270 ***
##                (0.0360)      (0.0358)
## male            0.0482 *      0.0487 *
##                (0.0199)      (0.0196)
## minority       -0.0424 *     -0.0429 *
##                (0.0203)      (0.0187)
## ----------------------------------------
## R^2             0.8041        0.8046
## Adj. R^2        0.8024        0.8029
## Num. obs.      474           474
## ========================================
## *** p < 0.001; ** p < 0.01; * p < 0.05
```

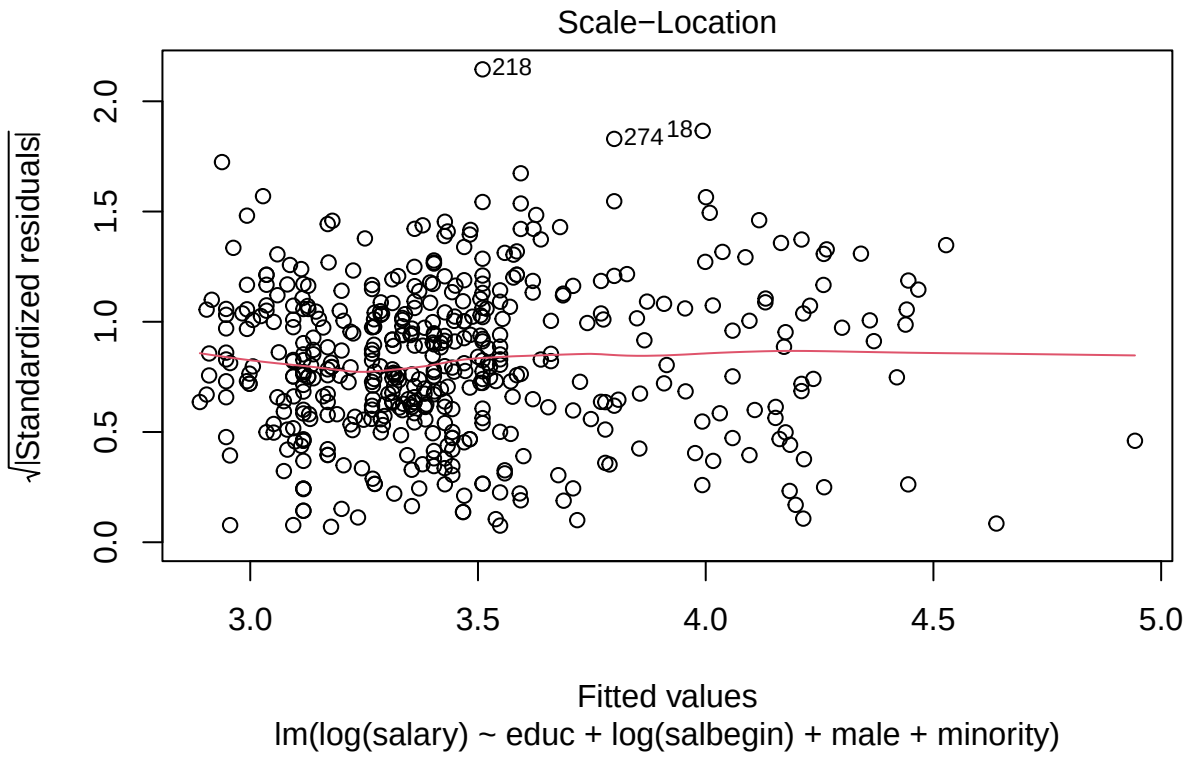
It can be seen that the standard errors for the different variables have been changed a bit, but nothing significant. It is very small changes, which maybe could indicate that the FGLS estimation have not taken all the heteroskedasticity into account. This will be elaborated further in 1.8.

8 - Has the FGLS estimation taken into account all the heteroskedasticity?

There are more ways to check whether the FGLS estimation has taken all the heteroskedasticity into account. One way is to look at it graphically and another is by doing a BP-test.

We start by looking at it graphically in the Scale-Location plot:

```
plot(FGLS, 3)
```



Then we perform a BP-test:

```
bptest(FGLS)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  FGLS
## BP = 60112, df = 4, p-value < 0.00000000000000022
```

Both the plot and BP-test indicates that the FGLS estimation have not taken all of the heteroskedasticity into account. In the plot it looks like there is a slight upward trend and that the spread of the residuals increase with higher fitted values. This suggest that the variance of the residuals is not constant indicating heteroskedasticity. As mentioned in 1.3 the null hypothesis in a BP-test is homoskedasticity, but we reject the null hypothesis here, since p-value is > 0.05.