Heteroscedasticity | R programming language

Preliminaries

Script Editor

```r
# setup
rm(list = ls())
directory <- "C:/Users/amalz/OneDrive/Desktop/"

# Install packages
PackageNames <- c("tidyverse", "stargazer", "magrittr", "lmtest", "sandwich")
for(i in PackageNames){
  if(!require(i, character.only = T)){
    install.packages(i, dependencies = T)
    require(i, character.only = T)
  }
}
```

Data exploration

Script Editor

```r
hprice1 <- read.csv(paste0(directory, "hprice1.csv"))

hprice1 %>%
  select(price, lprice, lotsize, sqrft, bdrms) %>%
  str

hprice1 %>%
  select(price, lprice, lotsize, sqrft, bdrms) %>%
  stargazer(type = "text")

hprice1 %>%
  select(price, lprice, lotsize, sqrft, bdrms) %>%
  head(05)
```

The results are as usual as what we have seen for other data files like CEOSAL1.csv and wage1.csv. Look through those notes for reference for how the console would look like

Graphical analysis of heteroscedasticity

Script Editor

```r
# Regression model for price
model_0 <- lm(price ~ lotsize + sqrft + bdrms, hprice1)
summary(model_0)
hprice1 %<>% mutate(uhat = resid(model_0))

# Graph of residuals against independent variable
ggplot(data = hprice1, mapping = aes(x = sqrft, y = uhat)) +
  theme_bw() +
  geom_point() +
  geom_hline(yintercept = 0, col = 'red') +
  labs(y = 'Residuals', x = 'Square feet, sqrft')

# Graph of residuals against fitted values
hprice1 %<>% mutate(yhat = fitted(model_0))
ggplot(data = hprice1, mapping = aes(x = yhat, y = uhat)) +
  theme_bw() +
  geom_point() +
  geom_hline(yintercept = 0, col = 'red') +
  labs(y = 'Residuals', x = 'Fitted values')
```

Heteroscedasticity | R programming language

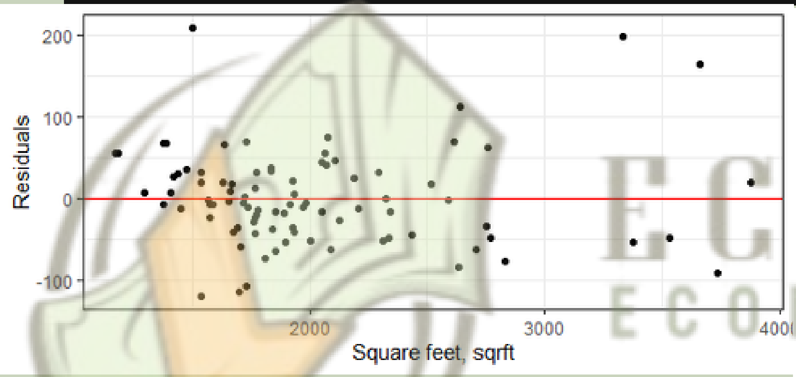
Console and Plots pane

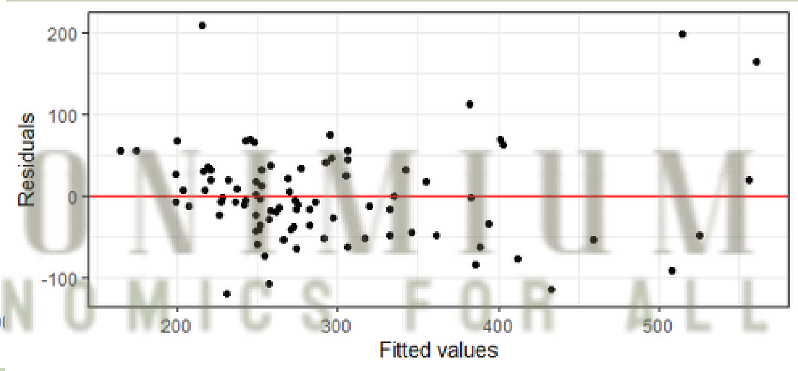
```
Call:
lm(formula = price ~ lotsize + sqrft + bdrms, data = hprice1)

Residuals:
     Min       1Q   Median       3Q      Max 
-120.026  -38.530   -6.555   32.323  209.376 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -2.177e+01  2.948e+01  -0.739  0.46221    
lotsize      2.068e-03  6.421e-04   3.220  0.00182 ** 
sqrft        1.228e-01  1.324e-02   9.275 1.66e-14 ***
bdrms        1.385e+01  9.010e+00   1.537  0.12795    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 59.83 on 84 degrees of freedom
Multiple R-squared:  0.6724,	Adjusted R-squared:  0.6607 
F-statistic: 57.46 on 3 and 84 DF,  p-value: < 2.2e-16
```





Script Editor

```r
# Regression model for lprice
model_1 <- lm(lprice ~ llotsize + lsqrft + bdrms, hprice1)
summary(model_1)
hprice1 %<>% mutate(uhat1 = resid(model_1))

# Graph of residuals against independent variable
ggplot(hprice1) +
  theme_bw() +
  geom_point(aes(x = lsqrft, y = uhat1)) +
  geom_hline(yintercept = 0, col = 'red') +
  labs(y = 'Residuals', x = 'Log square feet, lsqrft')

# Graph of residuals against fitted values
hprice1 %<>% mutate(yhat1 = fitted(model_1))
ggplot(data = hprice1, mapping = aes(x = yhat1, y = uhat1)) +
  theme_bw() +
  geom_point() +
  geom_hline(yintercept = 0, col = 'red') +
  labs(y = 'Residuals', x = 'Fitted values')
```

Heteroscedasticity | R programming language

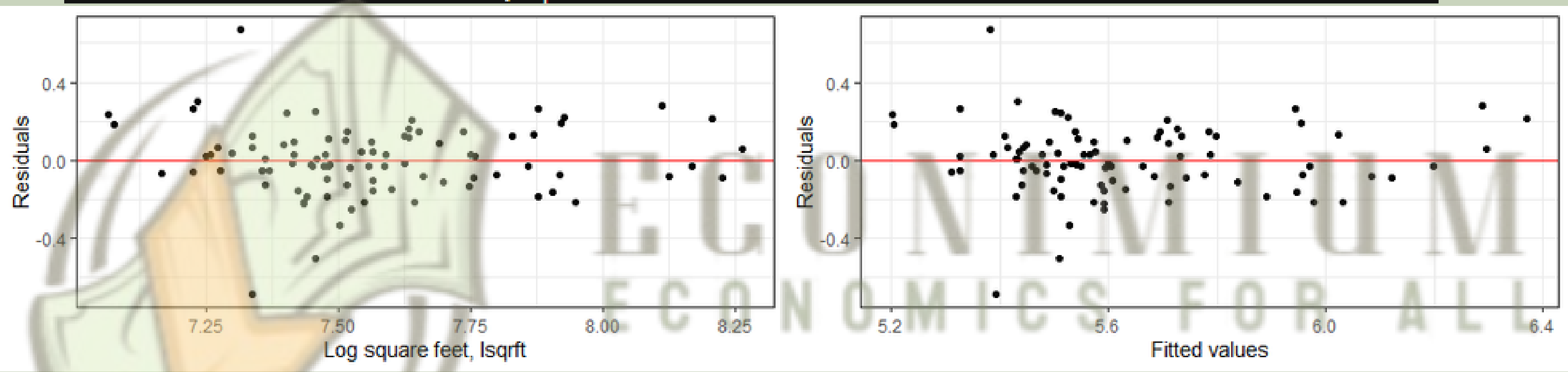
Console and Plots pane

```
Call:
lm(formula = lprice ~ llotsize + lsqrft + bdrms, data = hprice1)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.68422 -0.09178 -0.01584  0.11213  0.66899 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -1.29704    0.65128  -1.992   0.0497 *  
llotsize     0.16797    0.03828   4.388 3.31e-05 ***
lsqrft       0.70023    0.09287   7.540 5.01e-11 ***
bdrms        0.03696    0.02753   1.342   0.1831    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1846 on 84 degrees of freedom
Multiple R-squared:  0.643,	Adjusted R-squared:  0.6302 
F-statistic: 50.42 on 3 and 84 DF,  p-value: < 2.2e-16
```



(studentized) Breusch-Pagan test

Script Editor

```
install.packages("lmtest")
library(lmtest)
model_BP <- lm(price ~ lotsize + sqrft + bdrms, hprice1)
bptest(model_BP)
bptest(model_BP, studentize = FALSE)
```

Console

```
	studentized Breusch-Pagan test

data:  model_BP
BP = 14.092, df = 3, p-value = 0.002782
	Breusch-Pagan test

data:  model_BP
BP = 30.023, df = 3, p-value = 1.365e-06
```

If the p-value is less than 0.05, reject the null hypothesis. This indicates heteroskedasticity (variance of residuals is not constant).

Heteroscedasticity | R programming language

White test

Script Editor

```
# Generate squares and interaction of independent variables
hprice1 %<>% mutate(lotsizesq = lotsize^2,
                    sqrftsq = sqrft^2,
                    bdrmssq = bdrms^2,
                    lotsizeXsqrft = lotsize*sqrft,
                    lotsizeXbdrms = lotsize*bdrms,
                    sqrftXbdrms = sqrft*bdrms)
# Regression for White test
model_White <- lm(uhatsq ~ lotsize + sqrft + bdrms + lotsizesq + sqrftsq +
                    bdrmssq + lotsizeXsqrft + lotsizeXbdrms + sqrftXbdrms, hprice1)
summary(model_White)

# Number of independent variables k2
(k2 <- model_White$rank - 1)

# F-test and LM-test for heteroscedasticity
(r2 <- summary(model_White)$r.squared) # R-squared
(n <- nobs(model_White)) # number of observations

( F_stat <- (r2/k2) / ((1-r2)/(n-k2-1)) ) # F-statistic
( F_pval <- pf(F_stat, k2, n-k2-1, lower.tail = FALSE) ) # p-value

( LM_stat <- n * r2 ) # LM-statistic
( LM_pval <- pchisq(q = LM_stat, df = k2, lower.tail = FALSE)) # p-value
```

Console

```
lotsizesq     -4.978e-07  4.631e-06  -0.107  0.91467
sqrftsq        3.523e-04  1.840e-03   0.191  0.84864
bdrmssq        2.898e+02  7.588e+02   0.382  0.70362
lotsizeXsqrft  4.568e-04  2.769e-04   1.650  0.10303
lotsizeXbdrms  3.146e-01  2.521e-01   1.248  0.21572
sqrftXbdrms   -1.021e+00  1.667e+00  -0.612  0.54210
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5884 on 78 degrees of freedom
Multiple R-squared:  0.3833,	Adjusted R-squared:  0.3122
F-statistic: 5.387 on 9 and 78 DF,  p-value: 1.013e-05

>
> # Number of independent variables k2
> (k2 <- model_White$rank - 1)
[1] 9
>
> # F-test and LM-test for heteroscedasticity
> (r2 <- summary(model_White)$r.squared) # R-squared
[1] 0.3833143
> (n <- nobs(model_White)) # number of observations
[1] 88
>
> ( F_stat <- (r2/k2) / ((1-r2)/(n-k2-1)) ) # F-statistic
[1] 5.386953
> ( F_pval <- pf(F_stat, k2, n-k2-1, lower.tail = FALSE) ) # p-value
[1] 1.012939e-05
>
> ( LM_stat <- n * r2 ) # LM-statistic
[1] 33.73166
> ( LM_pval <- pchisq(q = LM_stat, df = k2, lower.tail = FALSE)) # p-value
[1] 9.95294e-05
> |
```

Heteroscedasticity | R programming language

Heteroscedasticity robust standard errors

Script Editor

```
summary(model_0)
# Regression model for price with robust standard errors
coeftest(model_0, vcov. = vcovHC(model_0, type = "HC1"))
```

Console

```
Call:
lm(formula = price ~ lotsize + sqrft + bdrms, data = hprice1)

Residuals:
     Min       1Q   Median       3Q      Max 
-120.026  -38.530   -6.555   32.323  209.376 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -2.177e+01  2.948e+01  -0.739  0.46221    
lotsize      2.068e-03  6.421e-04   3.220  0.00182 ** 
sqrft        1.228e-01  1.324e-02   9.275 1.66e-14 ***
bdrms        1.385e+01  9.010e+00   1.537  0.12795    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 59.83 on 84 degrees of freedom
Multiple R-squared:  0.6724,	Adjusted R-squared:  0.6607 
F-statistic: 57.46 on 3 and 84 DF,  p-value: < 2.2e-16

> # Regression model for price with robust standard errors
> coeftest(model_0, vcov. = vcovHC(model_0, type = "HC1"))

t test of coefficients:

               Estimate  Std. Error t value  Pr(>|t|)    
(Intercept) -21.7703081  37.1382106 -0.5862    0.5593    
lotsize       0.0020677   0.0012514  1.6523    0.1022    
sqrft         0.1227782   0.0177253  6.9267 8.096e-10 ***
bdrms        13.8525217   8.4786250  1.6338    0.1060    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Same coefficients, but heteroscedasticity consistent standard errors

Weighted Least Squares (WLS)

If the heteroskedasticity form is known, the Weighted Least Squares (WLS) can be used to estimate the model.

- *If the heteroscedasticity is known say is of a multiplicative relationship i.e. Var(u|x)= σ^2 . h(x) = σ^2 .hi, then, multiplying the errors ui by 1/ sqrt(hi) will make the errors homoscedastic.*
- *The transformed model will have variables and the constant multiplied by 1/sqrt(hi).*
- *OLS of the transformed model multiplying the variables by 1/sqrt(hi)is the same as estimating the original model using WLS with weight =1/hi*

Steps for WLS estimation from the datafile hprice1.csv

The heteroscedasticity form is known var(u|x)=(sigma^2)(sqrft), use WLS with weight=1/sqrft.*

[1] *WLS: estimate model with weight=1/sqrft*
[2] *Multiply all variables and the constant by 1/sqrt(sqrft)*
[3] *WLS: estimate model with transformed variables by OLS*

Script Editor

```
model_WLS1 <- lm(formula = price ~ lotsize + sqrft + bdrms,
                 data = hprice1, weights = 1/sqrft)                [1]
summary(model_WLS1)


hprice1 %<>% mutate(pricestar = price/sqrt(sqrft),
                    lotsizestar = lotsize/sqrt(sqrft),             [2]
                    sqrftstar = sqrft/sqrt(sqrft),
                    bdrmsstar = bdrms/sqrt(sqrft),
                    constantstar = 1/sqrt(sqrft))
model_WLS2 <- lm(pricestar ~ 0 + constantstar+ lotsizestar + sqrftstar + bdrmsstar,hprice1) [3]
summary(model_WLS2)
```

Console

```
Call:
lm(formula = price ~ lotsize + sqrft + bdrms, data = hprice1,
    weights = 1/sqrft)

Weighted Residuals:
    Min      1Q  Median      3Q     Max
-3.1464 -0.8548 -0.2247  0.6864  5.2457

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 4.199e+00  2.970e+01   0.141  0.88790
lotsize     1.588e-03  5.914e-04   2.686  0.00872 **
sqrft       1.178e-01  1.400e-02   8.412 9.06e-13 ***
bdrms       1.061e+01  8.659e+00   1.225  0.22398
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.308 on 84 degrees of freedom
Multiple R-squared:  0.5911,    Adjusted R-squared:  0.5765
F-statistic: 40.47 on 3 and 84 DF,  p-value: 2.799e-16
Call:
lm(formula = pricestar ~ 0 + constantstar + lotsizestar + sqrftstar +
    bdrmsstar, data = hprice1)

Residuals:
    Min      1Q  Median      3Q     Max
-3.1464 -0.8548 -0.2247  0.6864  5.2457

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
constantstar 4.199e+00  2.970e+01   0.141  0.88790
lotsizestar  1.588e-03  5.914e-04   2.686  0.00872 **
sqrftstar    1.178e-01  1.400e-02   8.412 9.06e-13 ***
bdrmsstar    1.061e+01  8.659e+00   1.225  0.22398
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.308 on 84 degrees of freedom
Multiple R-squared:  0.9634,    Adjusted R-squared:  0.9617
F-statistic: 552.8 on 4 and 84 DF,  p-value: < 2.2e-16
```

Feasible GLS (FGLS)

- *If the form of heteroscedasticity is unknown, the Feasible Generalized Least Squares (FGLS) method transforms the variables to achieve homoscedasticity. Here, h is estimated as h^ and used in the FGLS procedure. The procedure is as shown below*

The heteroscedasticity form is expressed as:

$$\mathrm{Var}(u_i \mid x_i) = \sigma^2 h(x_i) = \sigma^2 \exp(\delta_0 + \delta_1 x_{i1} + \cdots + \delta_k x_{ik}),$$

where u_i is the error term, x_i represents the regressors, and $\delta_0, \delta_1, \ldots, \delta_k$ are parameters.

Estimate the regression model:

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

Obtain the residuals $\hat{u}_i$ from the regression and calculate $\log(\hat{u}_i^2)$.

Estimate the regression model:

$$\log(\hat{u}_i^2) = \delta_0 + \delta_1 x_{i1} + \cdots + \delta_k x_{ik} + e_i,$$

where e_i is the error term.

Obtain the fitted values $\hat{g}_i$ from this regression and compute $\hat{h}_i = \exp(\hat{g}_i)$.

Use Weighted Least Squares (WLS) to estimate the original regression model:

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

with weights $w_i = \frac{1}{\hat{h}_i}$.

Alternatively, transform all variables (including the dependent variable y_i, the regressors x_{ij}, and the constant term) by multiplying them with $\frac{1}{\sqrt{\hat{h}_i}}$. The transformed model is then estimated using Ordinary Least Squares (OLS).

Script Editor

```r
# When the heteroscedasticity form is not known,
# var(u|x) = sigma^2*(delta0 + delta1*lotsize + delta2*sqrft + delta3*bdrms)
# estimate hhat and use WLS with weight=1/hhat.

# Heteroscedasticity form, estimate hhat
# model_0 <- lm(price ~ lotsize + sqrft + bdrms, hprice1)
summary(model_0)
hprice1 %<>% mutate(u = resid(model_0),
                    g = log(u^2))
model_g <- lm(g ~ lotsize + sqrft + bdrms, hprice1)
hprice1 %<>% mutate(ghat = fitted(model_g),
                    hhat = exp(ghat))

# FGLS: estimate model using WLS with weight=1/hhat
model_FGLS1 <- lm(formula = price ~ lotsize + sqrft + bdrms,
                  data = hprice1,
                  weights = 1/hhat)
summary(model_FGLS1)

# Multiply all variables and the constant by 1/sqrt(hhat)
hprice1 %<>% mutate(pricestar1 = price/sqrt(hhat),
                    lotsizestar1 = lotsize/sqrt(hhat),
                    sqrftstar1 = sqrft/sqrt(hhat),
                    bdrmsstar1 = bdrms/sqrt(hhat),
                    constantstar1 = 1/sqrt(hhat))

# FGLS: estimate model with transformed variables by OLS
model_FGLS2 <- lm(pricestar1 ~ 0 + constantstar1 + lotsizestar1 + sqrftstar1 +
                    bdrmsstar1, hprice1)
summary(model_FGLS2)
```

Console

```
> summary(model_0)

Call:
lm(formula = price ~ lotsize + sqrft + bdrms, data = hprice1)

Residuals:
     Min       1Q   Median       3Q      Max 
-120.026  -38.530   -6.555   32.323  209.376 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -2.177e+01  2.948e+01  -0.739  0.46221    
lotsize      2.068e-03  6.421e-04   3.220  0.00182 ** 
sqrft        1.228e-01  1.324e-02   9.275 1.66e-14 ***
bdrms        1.385e+01  9.010e+00   1.537  0.12795    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 59.83 on 84 degrees of freedom
Multiple R-squared:  0.6724,    Adjusted R-squared:  0.6607 
F-statistic: 57.46 on 3 and 84 DF,  p-value: < 2.2e-16
> summary(model_FGLS1)

Call:
lm(formula = price ~ lotsize + sqrft + bdrms, data = hprice1, 
    weights = 1/hhat)

Weighted Residuals:
    Min      1Q  Median      3Q     Max 
-4.9267 -1.1313 -0.2846  1.1836  9.8270 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 45.911604  30.823535   1.489  0.14010    
lotsize      0.004135   0.001426   2.901  0.00475 ** 
sqrft        0.092462   0.014866   6.220 1.86e-08 ***
bdrms        6.175451   8.893592   0.694  0.48937    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.974 on 84 degrees of freedom
Multiple R-squared:  0.4675,    Adjusted R-squared:  0.4485 
F-statistic: 24.58 on 3 and 84 DF,  p-value: 1.633e-11
> summary(model_FGLS2)

Call:
lm(formula = pricestar1 ~ 0 + constantstar1 + lotsizestar1 + 
    sqrftstar1 + bdrmsstar1, data = hprice1)

Residuals:
    Min      1Q  Median      3Q     Max 
-4.9267 -1.1313 -0.2846  1.1836  9.8270 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
constantstar1 45.911604  30.823535   1.489  0.14010    
lotsizestar1   0.004135   0.001426   2.901  0.00475 ** 
sqrftstar1     0.092462   0.014866   6.220 1.86e-08 ***
bdrmsstar1     6.175451   8.893592   0.694  0.48937    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.974 on 84 degrees of freedom
Multiple R-squared:  0.9681,    Adjusted R-squared:  0.9665 
F-statistic: 636.3 on 4 and 84 DF,  p-value: < 2.2e-16
```