HANDLING HETEROSCEDASTICITY OF LINEAR REGRESSION MODELS IN R

PAUL KAGORI

2023-11-15

#1. Required Libraries

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.2.3
```

```
library(AER)
```

```
## Warning: package 'AER' was built under R version 4.2.3
```

```
## Loading required package: car
```

```
## Warning: package 'car' was built under R version 4.2.3
```

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

```
## Warning: package 'carData' was built under R version 4.2.3
```

```
## Loading required package: lmtest
```

```
## Warning: package 'lmtest' was built under R version 4.2.3
```

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

```
## Warning: package 'zoo' was built under R version 4.2.3
```

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

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

```
## Loading required package: sandwich
```

```
## Warning: package 'sandwich' was built under R version 4.2.3
```

```
## Loading required package: survival
```

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 4.2.3
```

```
##
## Attaching package: 'dplyr'
```

```
## The following object is masked from 'package:car':
##
##     recode
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(knitr)
```

```
## Warning: package 'knitr' was built under R version 4.2.3
```

```
library(tidyverse) #data wrangling
```

```
## Warning: package 'tidyverse' was built under R version 4.2.3
```

```
## Warning: package 'tibble' was built under R version 4.2.3
```

```
## Warning: package 'tidyr' was built under R version 4.2.3
```

```
## Warning: package 'readr' was built under R version 4.2.3
```

```
## Warning: package 'purrr' was built under R version 4.2.3
```

```
## Warning: package 'stringr' was built under R version 4.2.3
```

```
## Warning: package 'forcats' was built under R version 4.2.3
```

```
## Warning: package 'lubridate' was built under R version 4.2.3
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v forcats   1.0.0     v stringr   1.5.0
## v lubridate 1.9.2     v tibble    3.2.1
## v purrr     1.0.2     v tidyr     1.3.0
## v readr     2.1.4
```

```
## -- Conflicts ---------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## x dplyr::recode() masks car::recode()
## x purrr::some()   masks car::some()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(foreign)   #read.data()
library(MASS)
```

```
##
## Attaching package: 'MASS'
##
## The following object is masked from 'package:dplyr':
##
##     select
```

1) INTRODUCTION

In my previous article, I reviewed heteroscedasticity, its effect on linear regression, and performed three tests to demonstrate how to identify a heteroscedastic model. In this article we will:

-1. Perform Koenker test and Goldfield-Quandt Test to check if a linear model is heteroscedastic.

-2. Transform heteroscedastic linear model to homoscedastic form through generalized Linear Model(GLM) and Weighted Least Squares(WLS).

-3.Perform Feasible GLS algorithm on data.

2) Loading the Dataset

-We use the Journal Data.It consists of Journal Name, Number of subscriptions (subs), Price per charge per article for publication (price), and price charged per citation

```
load("EconData.RData")
head(journals, n = 5)
```

```
##        subs price  citeprice
## APEL     14   123  5.8571429
## SAJoEH   59    20  0.9090909
## CE       17   443 20.1363636
## MEPiTE    2   276 12.5454545
## JoSE     96   295 12.2916667
```

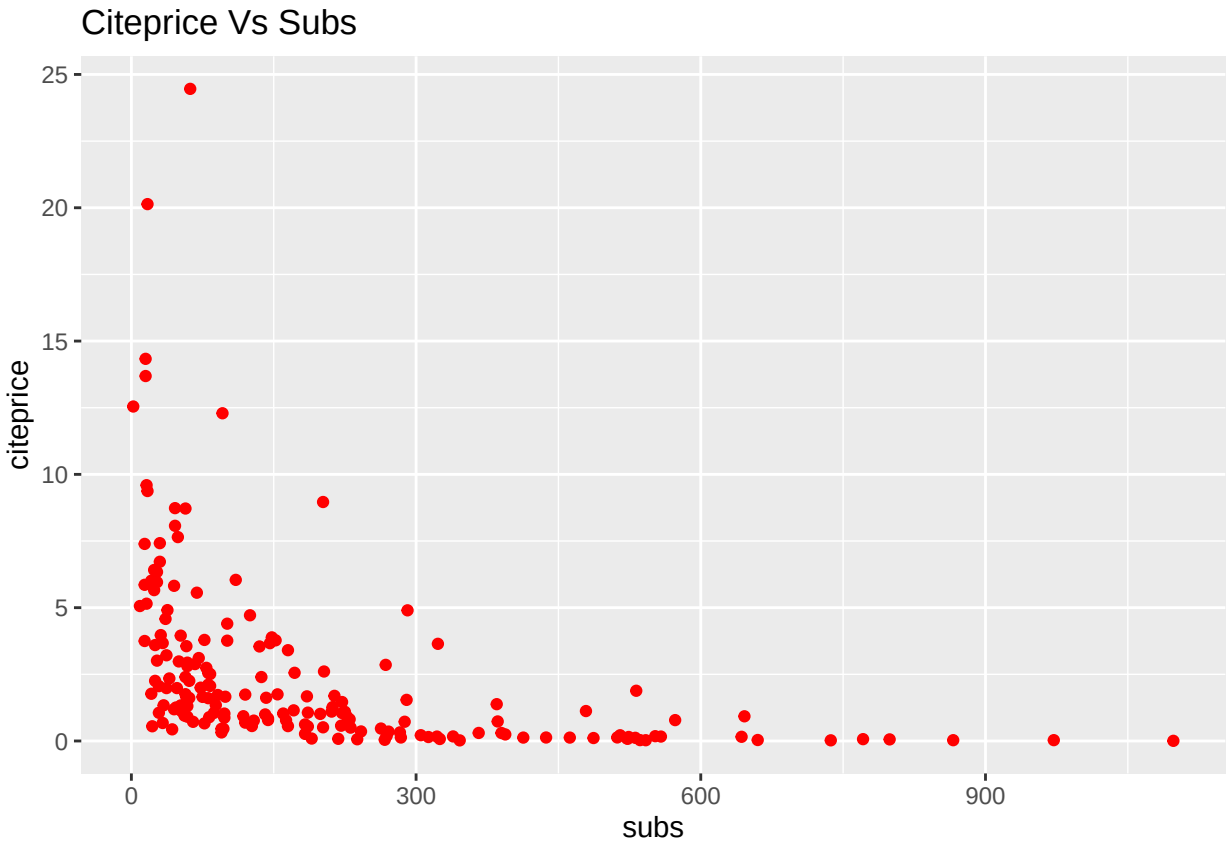
```
dim(journals)
```

```
## [1] 180   3
```

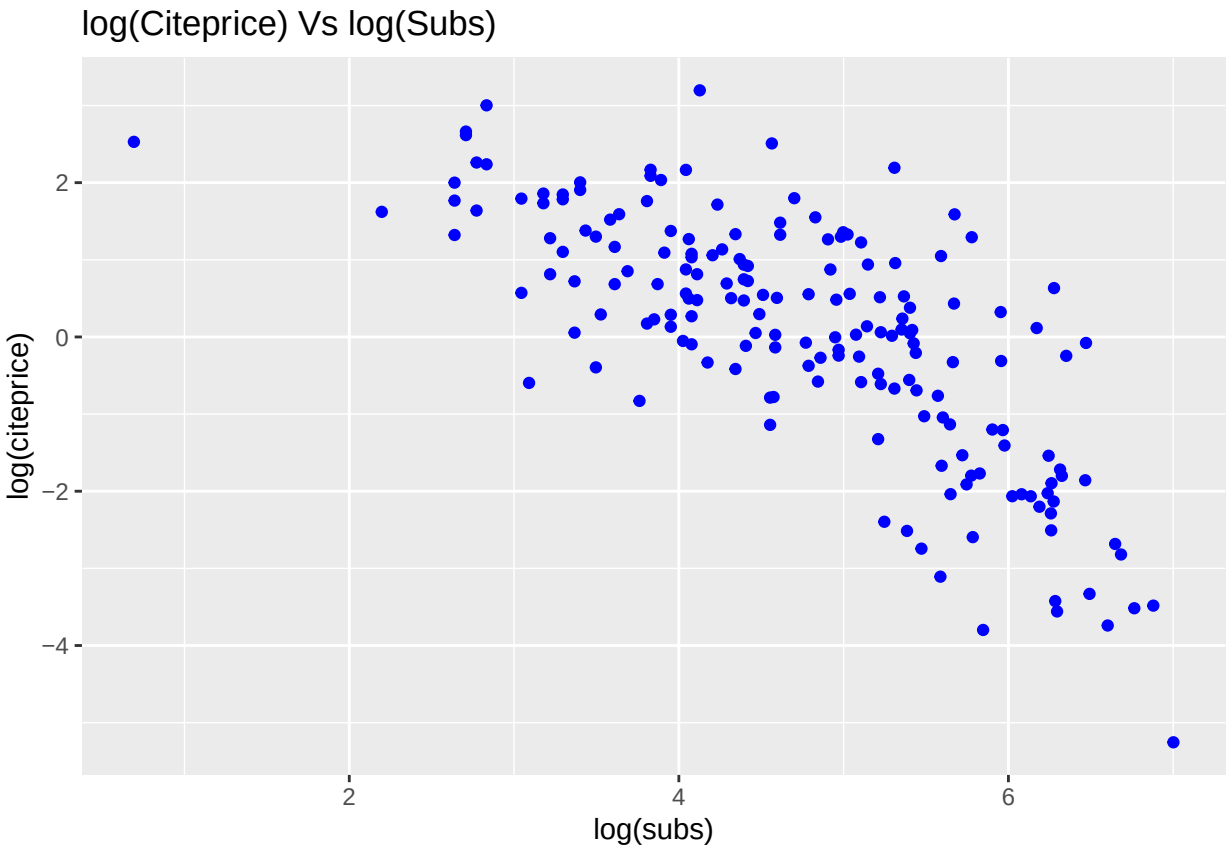
3) Qucik Check on the Scatter Plot

We would like to make prediction on number of subscriptions (subs) based on the cite price(citeprice).

```
library(lmtest)
ggplot(data = journals, aes(subs,citeprice), xlab = "citeprice", ylab = "Subs",
       main = "Citeprice Vs Subs") + geom_point(col = "red") + ggtitle("Citeprice Vs Subs")# Now we hav
```




```
ggplot(data = journals, aes(log(subs),log(citeprice)), xlab = "log(citeprice)",
       ylab = "log(Subs)") + geom_point(col = "blue")+ ggtitle("log(Citeprice) Vs log(Subs)")# Now we h
```



Its now linear, with negative relationship. Just as you can see from the graph

4) Linear Model

```
jlm1 <- lm(log(subs) ~ log(citeprice), data = journals)
summary(jlm1)
```

```
##
## Call:
## lm(formula = log(subs) ~ log(citeprice), data = journals)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.72478 -0.53609  0.03721  0.46619  1.84808
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     4.76621    0.05591   85.25   <2e-16 ***
## log(citeprice) -0.53305    0.03561  -14.97   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7497 on 178 degrees of freedom
## Multiple R-squared:  0.5573, Adjusted R-squared:  0.5548
## F-statistic:   224 on 1 and 178 DF,  p-value: < 2.2e-16
```

```
jlm1$coefficients ## Coefficients of the model
```

```
##    (Intercept) log(citeprice)
##      4.7662121     -0.5330535
```

4(i) Fitted Results

As you can see, both the intercept and the coefficients are significant and so the model is:

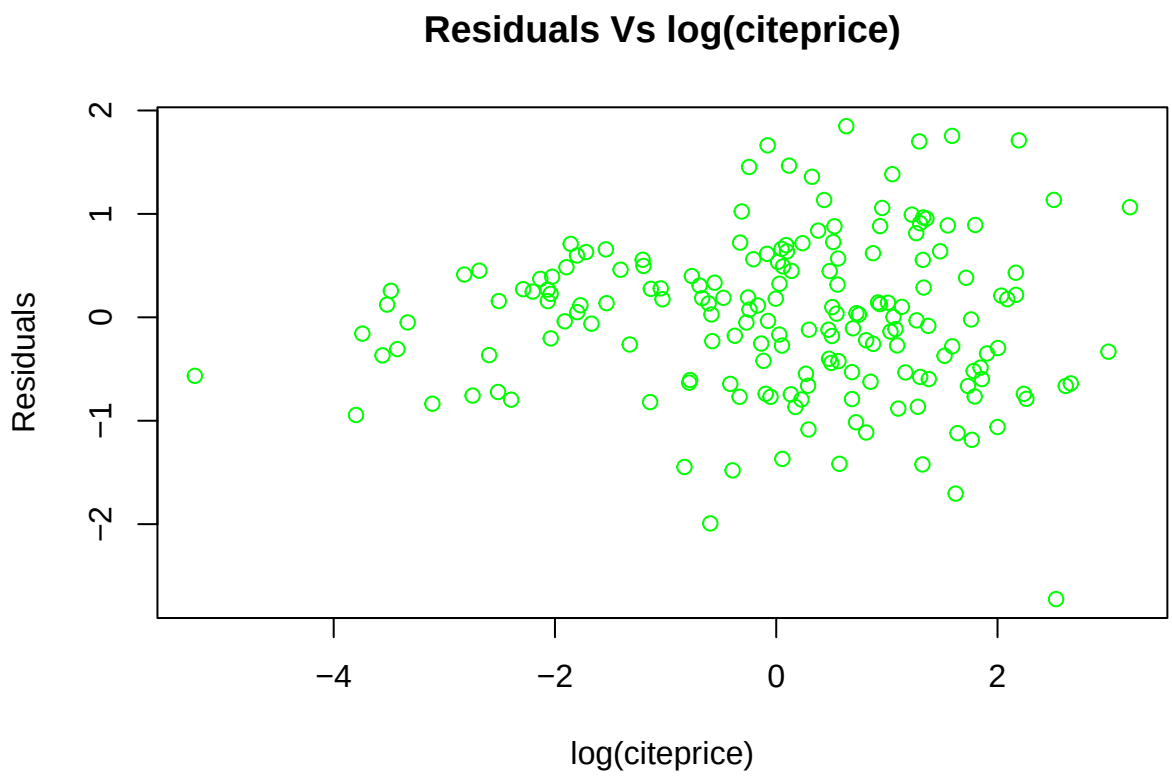
$$subs = 4.766 - 0.533 citeprice$$

In other words, the higher the cite price, the lower the number of subscriptions. In short, citeprice has and subscriptions are inversly related. However we have:

$$R^2 = 0.5573$$

The model is not really a good fit. This makes sense from the plot 1 above. You can see the data are more spread out from a central line.

```
jres <- residuals(jlm1) # We check on the residuals
plot(log(journals$citeprice), jres, xlab = "log(citeprice)", ylab = "Residuals", main = "Residuals Vs l
```

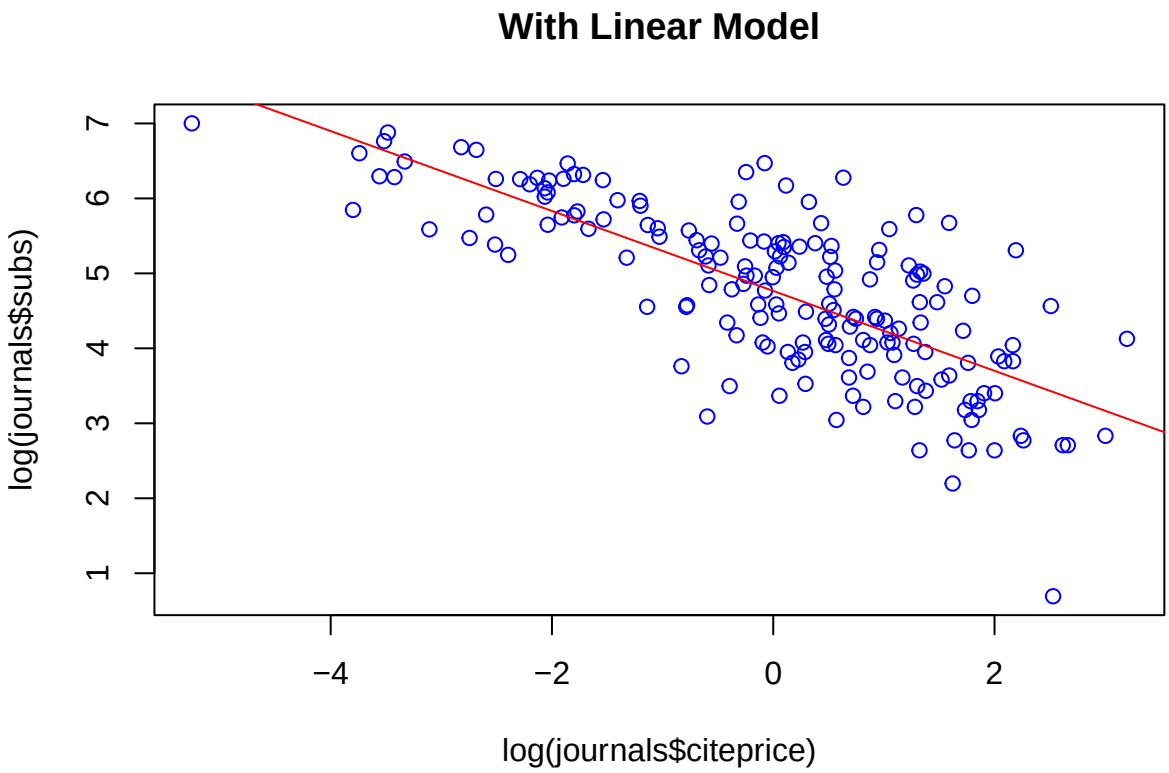


4(ii) The plot of Residuals Vs Cite price

Clearly, as you can see, the model is heteroscedastic. The residuals tend to increase with increase in citeprice

4(ii) Checking the Linear Fit

```
plot(log(journals$subs)~log(journals$citeprice), col= "blue", main = "With Linear Model") + abline(jlm1
```




```
## integer(0)
```

5) Testing for Heteroscedasticity

5(i) Breusch- Pagan Test

Under BP test, We have the following:

Original Model: $y = \beta_o + \beta_1 x_1 + \cdots + \beta_r x_r + \xi$ Test Model: $\hat{\xi^2} = \alpha_o + \alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_r x_r$

Then, we carry out hypothesis test using ANOVA that:

H_o : Original Model is Homoscedastic H_1 : Original Model is Heteroscedastic

Under Ho one has to accept

$$\alpha_1 = \alpha_2 = \cdots = \alpha_r = 0$$

```
bplm2 <-lm(jres^2 ~ log(journals$citeprice))
summary(bplm2)
```

```
##
## Call:
## lm(formula = jres^2 ~ log(journals$citeprice))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.8365 -0.4689 -0.2170  0.0853  6.5401
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)               0.54945    0.06492   8.463 9.33e-15 ***
## log(journals$citeprice)  0.13241    0.04135   3.202  0.00162 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8706 on 178 degrees of freedom
## Multiple R-squared:  0.05446,    Adjusted R-squared:  0.04915
## F-statistic: 10.25 on 1 and 178 DF,  p-value: 0.001617
```

```
anova(bplm2)
```

```
## Analysis of Variance Table
##
## Response: jres^2
##                           Df  Sum Sq Mean Sq F value   Pr(>F)
## log(journals$citeprice)   1   7.771  7.7708  10.252 0.001617 **
## Residuals                178 134.915  0.7579
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

AS you can see, the Pvalue = .001617 and since its less that .05, we REJECT the Null Hypothesis and conclude that the model is heteroscedastic.

5(ii) GoldField -Quandt (GQ)Test

As we have seen, the reggressor causing heteroscedasticity is the Citeprice. To carry out GQ test, we need:

-1. Sort the observations based on asceding value of the regressor causing heteroscedasticity.

-2. Fit to separate reggressor models i.e, one-to-"small" values and one-to-"large" values.

-3. Using F-test, we test the equality of the residual variances of the two models

-4. If the models is heteroscedastic, the theoretical variables in (3) above will be equal.

```
### Goldfeld-Quandt test

gqtest(jlm1, point = .5, fraction = 1/3, order.by = ~log(journals$citeprice))
```

```
##
##  Goldfeld-Quandt test
##
## data:  jlm1
## GQ = 2.136, df1 = 58, df2 = 58, p-value = 0.002226
## alternative hypothesis: variance increases from segment 1 to 2
```

fraction = 1/3 means we drop the middles (1/3) of the data, this gives a clear partition of the two groups. We then order the remaining observations with respect to Citeprice.

5) Results

From the test, we can see that the variances of the two segment increases from first group to second group. Hence, the model is Heteroscedastic.

6) Handling Heteroscedasticity Through Feasible GLS

Since we have already confirmed that our model is Heteroscedastic, and have also identified the variable causing heteroscedasticity, we now apply FGLS to convert it into homoscedastic one. The procedure is as follows:

-1. Regress the response variable(y) on the regressors (x1, x2, . . . , xr) to obtain the OLS residuals.

-2. On the independent variables (citeprice), regress

$$log(\hat{\xi}^2)$$

-3. Calculate prediction.

$$\hat{g} = \hat{\delta_o} + \hat{\delta_1} x_1 + \hat{\delta_2} x_2 + \cdots + \hat{\delta_r} x_r$$

4. Calculate

$$\hat{h} = e^{\hat{g}}$$

.

5. Finally,We then carry out WLS to estimate the regression parameter (y) on (x1, x2, . . . , xr) using the weights

$$\frac{1}{\sqrt{\hat{h}}}$$

```
jlm1fit <- jlm1$fitted.values ### The fitted values from Heteroscedastic model
jlgs <- lm(log(jres^2) ~ jlm1fit + I(jlm1fit^2)) ## Step 2 of FGLS above
summary(jlgs)
```

```
##
## Call:
## lm(formula = log(jres^2) ~ jlm1fit + I(jlm1fit^2))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.8747 -1.1886  0.4791  1.4214  3.5472
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    4.7672     4.2860   1.112    0.268
## jlm1fit       -2.3373     1.7317  -1.350    0.179
## I(jlm1fit^2)   0.1914     0.1713   1.117    0.265
##
## Residual standard error: 2.086 on 177 degrees of freedom
## Multiple R-squared:  0.03385,    Adjusted R-squared:  0.02293
## F-statistic: 3.101 on 2 and 177 DF,  p-value: 0.04748
```

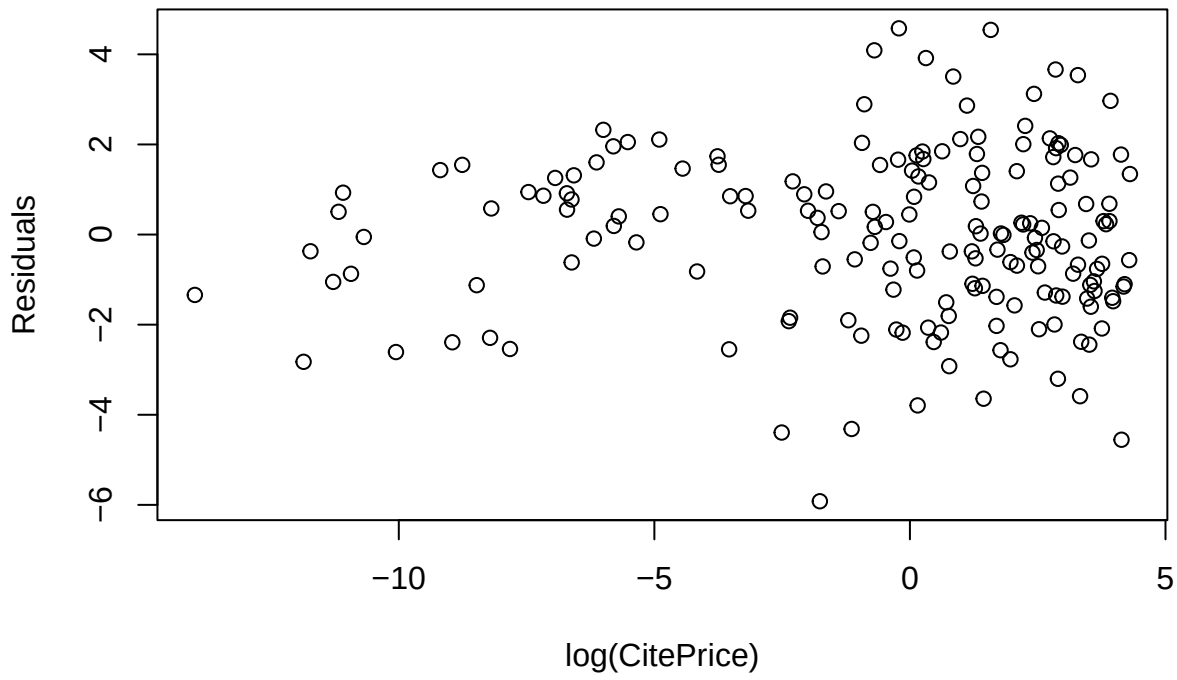
```r
jlgs_w <- 1/exp(fitted.values(jlgs)) ## The weight

j_lmw <- lm(log(subs) ~ log(citeprice), weights = jlgs_w,  data = journals) ### regress step 2above
summary(j_lmw)
```

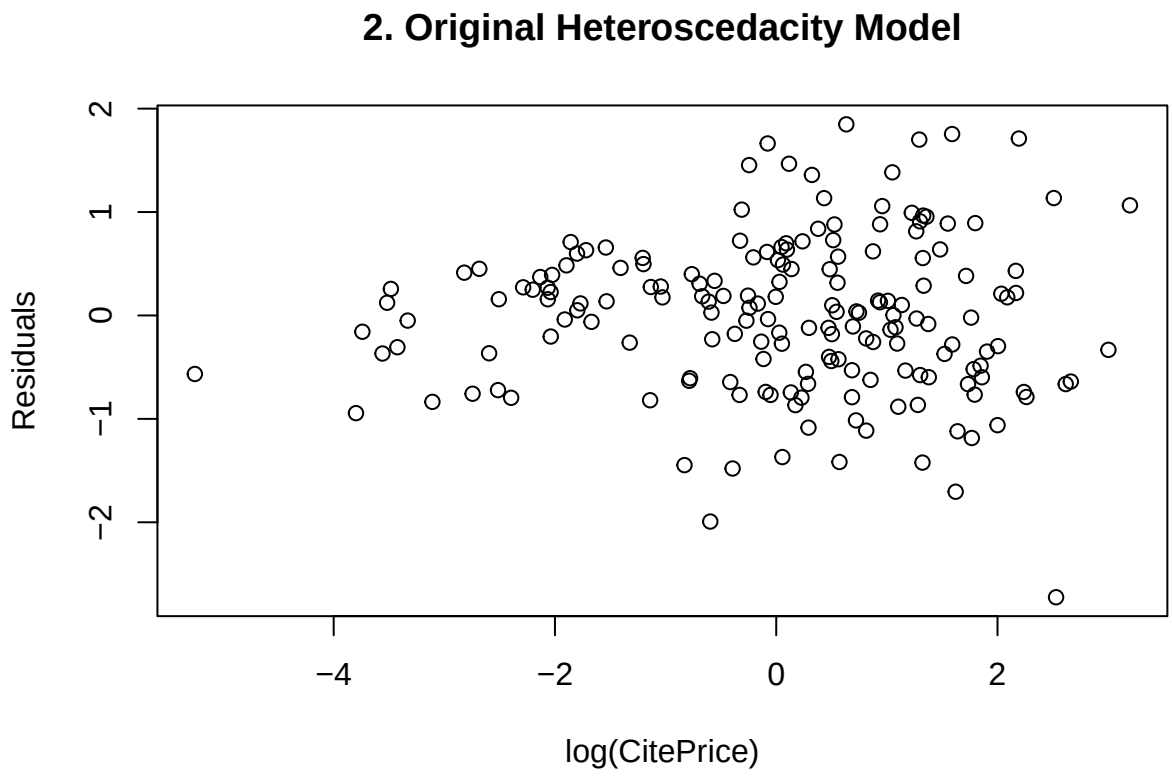
```
##
## Call:
## lm(formula = log(subs) ~ log(citeprice), data = journals, weights = jlgs_w)
##
## Weighted Residuals:
##     Min      1Q  Median      3Q     Max
## -5.9192 -1.2302  0.0388  1.3220  4.5771
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     4.78461    0.05422   88.24   <2e-16 ***
## log(citeprice) -0.51763    0.03370  -15.36   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.848 on 178 degrees of freedom
## Multiple R-squared:  0.5699, Adjusted R-squared:  0.5675
## F-statistic: 235.9 on 1 and 178 DF,  p-value: < 2.2e-16
```

7) Graphically Checking the Results

```r
plot(sqrt(jlgs_w)*log(journals$citeprice),sqrt(jlgs_w)*residuals(j_lmw), xlab = "log(CitePrice)", ylab =
```




```
plot(log(journals$citeprice), jres, main = "2. Original Heteroscedacity Model", xlab = "log(CitePrice)"
```



From the two graphs, you can notice that graph1, is less heteroscedastic compared to graph 2