Working with heteroscedastic models

Final project: STA6446, Regression Analysis

Sotuyo, Carlos

Introduction:

Homoscedasticity in one of the conditions required by the Gauss-Markov theorem in order to assure that the ordinary least square is a valid procedure. The term homoscedastic[1] refer to the fact that $var(\epsilon_i) = var(y_i)$ is constant, σ^2 in short, constant variance of the residuals. The violation of this conditions leads to what is called heteroscedasticity, which force the analyst to reconsider the model in question (including the possibility of missing independent variables) or to try a theoretical procedure in order to overcome the heteroscedasticity of the model. Among the theoretical solutions to consider, we have variance stabilizing transformations and WLS, weighted least squares. In the present study, the author examines three data sets exploring different approaches working with R statistical software.

Heteroscedasticity does not result in biased parameter estimates; however, they are no longer consider BLUE[2]. That is, the standard errors are biased, thus inferences about the parameters (test statistics and confidence intervals) cannot be trusted. There are two chief ways to detect heteroscedasticity: one, is a plot of residuals versus fitted values, (whenever the residuals increase as independent variables increase, the scatter plot takes the shape of a megaphone) or by a statistical test whose null hypothesis is homoscedasticity. One of the most widely test is the Breusch Pagan test (in R, bptest part of the lmtest package). However, browsing the topic of heteroscedastic, this student found out that examining the residual plot is for some analysts enough criteria to judge heteroscedasticity. A case study developed by The Pennsylvania State University statistics pages[3] is a relevant example. The technical reasons for this choice of the plot of the residuals over the statistical test, has been studied for statisticians (Zaman,2000)[4]. In a intuitive way we may say that since the test depends on n, the sample size, for large sample size even when the variance is constant or approximately constant, the test yield significant p values. In technical terms, Zaman writes: *The test Breusch-Pagan Test is consistent if and only if there is some linear relationship between the error variances and the regressors.*

In dealing with heteroscedasticity in the case studies of the present work, the author has considered different methods: box cox transformations, WLS (weight least squares) and logarithmic transformations. Finally, robust estimators calculations are assess in order to correct the biased standard errors of the parameters that result from heteroscedastic models.

[1] A. Sen; M. Srivastava, Regression Analysis, pag 111.
[2] Williams, R. Heteroskedasticity, University of Notre Dame: https:/www3.nd.edu/ rwilliam.
[3] https://onlinecourses.science.psu.edu/stat501/node/3971.
[4] Zaman, A. The Inconsistency of the Breusch-Pagan Test, Journal of Economic and Social Science research 2(1), 2000, 1-11.

Statistical methods and case studies:

In the cases studies considered in the present work, the Box Cox transformations are the first to be considered.

The Box Cox transformation allows to modify non-normal dependent variables into a normal distributed random variables. But, since the theoretical distribution of the depended variable is a cause of heteroscedasticity [5], by correcting the lack of normality of the depended variables (and therefore the residuals), *it also reduces heterocedasticity*[6].
The Box-Cox transformation is a particularly useful family of transformations. It is defined as:

$$T(Y) = \frac{(Y^{\lambda} - 1)}{\lambda} \quad \text{when } \lambda \neq 0$$

Where Y is the response variable and λ is the transformation parameter. For $\lambda = 0$, the natural log of the data is taken instead of using the above formula[7]: $T(Y) = log(y_i)$, when $\lambda = 0$.

R script to find lambda:

> bc=boxcox(mymodel)
> lambda=bc$x[which.max(bc$y)]
> lambda

As variance increases with the independent variables, the it can be stated that $var(\epsilon_i) = \sigma_i^2 = c_i^2\sigma^2$, where c_i^2 are known constants. Weighted least squares consists of dividing both sides of the OLS model by c_i. This results in the weighted sum of squares[8]:

$$\sum w_i(y_i - \beta_0 - \beta_1 x_{i1} ... - \beta_k x_{ik})^2$$

Where w_i denotes the *weights*. A procedure to generate weights is described by Jeffrey M. Wooldridge[9]:
1. Run the regression of y on $x_1, x_2, , x_k$ and obtain the residuals e.
2. Create $\log(e^2)$ by first squaring the OLS residuals and then taking the natural log.
3. Run the regression of the log of the square of the residuals vs independent variables and obtain the fitted values.
4. Exponentiate the fitted values from the previous step: these are the weights, w.
5. Run the initial model using $1/w$ as weights.

[5] A. Sen; M. Srivastava, Regression Analysis, pag 122.
[6] A. Sen; M. Srivastava, Regression Analysis, pag 206.
[7] Engineering statistics handbook: http://www.itl.nist.gov/div898/handbook/eda/section3/eda336.htm
[8] A. Sen; M. Srivastava, Regression Analysis, pag 119.
[9] Wooldridge,J. M. Econometrics, a modern approach, Cengage Learning, 2012. page 287

Professor M. Crowson[10], University of Alabama, in a lecture, proceed to calculate weights in a slightly different way:

1. Run the OLS model.
2. Save the absolute value of the standardized residuals.
3. Run an auxiliary model using the st residuals as depending variable.
4. Save the standardized fitted fitted values of the aux model: these are the weights=w.
5. Run the WLS model: same structure as OLS, using $1/w^2$ as weights.

As weights we may also consider $z_i, 1/z_i,\ and\ \sqrt{y_i}$ for poisson distributed dependent variables. But these type of weights also leads to bias (Ashin Sen, p123) since the expected value of $E(y_i)$ is not known, although we estimate $E(y_i)$ with $w[y_i]$.

Montgomery et al in Introduction to Linear Regression Analysis[11], suggests the following *weights* based on residual analysis:

1. If variance of errors is a function of one of the regressors, use $w_i = \frac{1}{x_i}$;
2. If y_i is actually an average of n_i observations at x_i and if all original observations have constant varianceσ^2, use $w_i = n_i$.

Logarithmic transformation, when applied only to the depended variable is a case of Box Cox transformation; on the other hand, when the residual plot suggest the linearity of the model is compromised, we apply log to achieve linearity.

Robust estimators are resistant to outliers and when used in regression modeling, are robust to departures from the normality assumption.

M-estimators are a maximum likelihood type estimator [12]. M estimation minimizes: sum $\rho(e_i)$ from i to n where ρ is some function with the following properties:

$\rho(r) \geq 0$ for all r and has a minimum at 0.

$\rho(r) = \rho(-r)$ for all r.

$\rho(r)$ increases as r increases from 0, but does not get too large as r increases.

Huber proposed an M-Estimator that has the following $\rho(z_i)$ function:

$\rho(z_i) = \frac{1}{2}z_i^2,\ \text{if}\ |z_i| < c$

$\rho(z_i) = c|z_i| - \frac{1}{2}c^2,\ \text{if}\ |z_i| \geq c$

So, it is basically OLS for abs values of errors less than a certain c, and a more complex definition for values greater than c.

We implement robust regression in R in different ways (and packages). One is the function rlm (MASS package) or simply we obtain the standard errors (robust, unbiased) by running:

> coeftest(mymodel, vcov=hccm) # from package lmtest.

Case study I: Violent Crime[13]. Fifty small cities violent crime count (Reference: Life In America's Small Cities, By G.S. Thomas). The model dependent variable is annual violent crimes per city, the independent variables are police funding, and percentages people over 25 with high school diploma, percentage of 16-19 yrs old not in high school, percentage of 18-24 yrs old in college and percentage of over 25 with a bachelor degree.

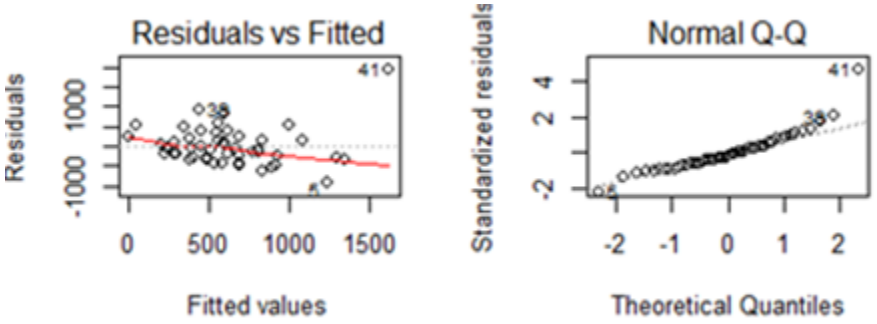
[10] $https://youtu.be/enPK_SXILnA$

[11] Montgomery, D. Introduction to Regression Analysis,Willey, New York, 2003 page 196

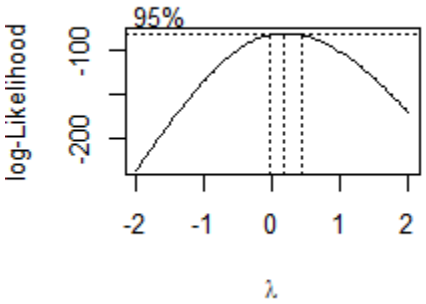
[12] Bellio, R. An introduction to robust estimators with R functions. Dept of Statistics, University of Udine. Retrieve from https://www.researchgate.net/publication/228906268

[13] Crime data: Reference: Life In America's Small Cities, By G.S. Thomas, $https://web.stanford.edu/\ hastie/StatLearnSparsity_files/DATA/crime$

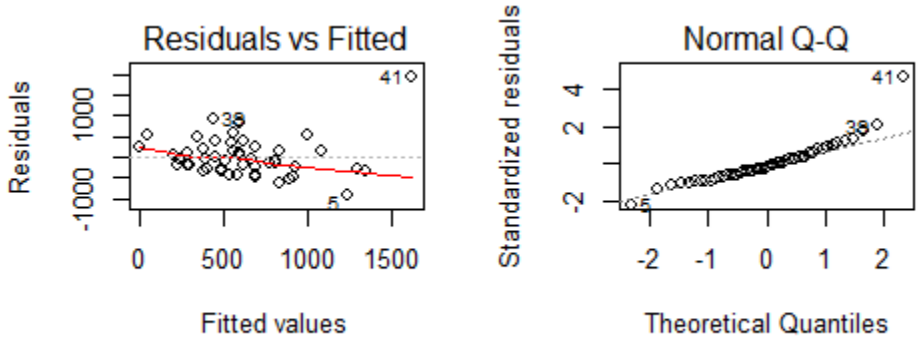
A Shapiro-Wilk normality test with a p-value = 0.0004912 rejects the assumption of normality; the studentized Breusch-Pagan test with a p-value = 0.01744 rejects the assumption of homoscedasticity. Multiple R sq =0.34, and the overall F statistics is significant: p-value 0.002303.



Both, the residual and the normality plot show an outlier observation (city 41). A box cox transformation in R finds lambda to be 0.181818:



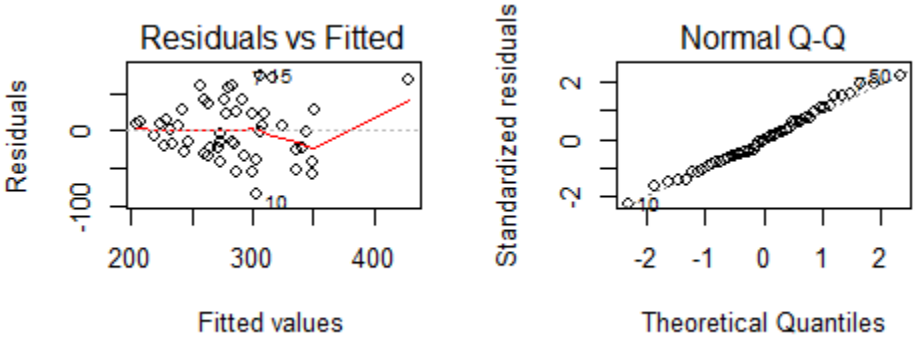
After running the model using lambda = 1/5,the Shapiro-Wilk normality test p value is 0.2537 and the studentized Breusch-Pagan test p-value = 0.4495.



Multiple R-squared: 0.3695; F-statistic: 5.158, p-value: 0.0008263

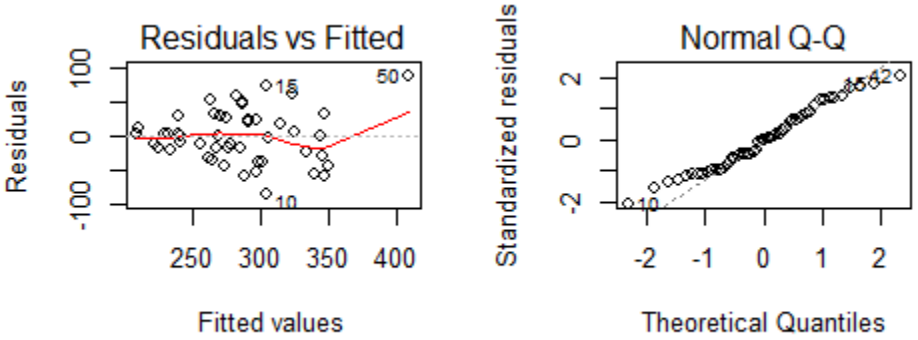
Case study II: Per capita expenditure on public education in a state, projected for 1975[14]. Considering two independent variables: per capita personal income, and population under 18.

The studentized Breusch-Pagan test OLS model has a p value of 0.01158; a Shapiro-Wilk normality test p-value = 0.6494. Therefore, the assumption of homoscedasticity is rejected while the normality assumption is met.



The megaphone shape of the scatter plot of the residuals shows an obvious pattern of heteroscedasticity.

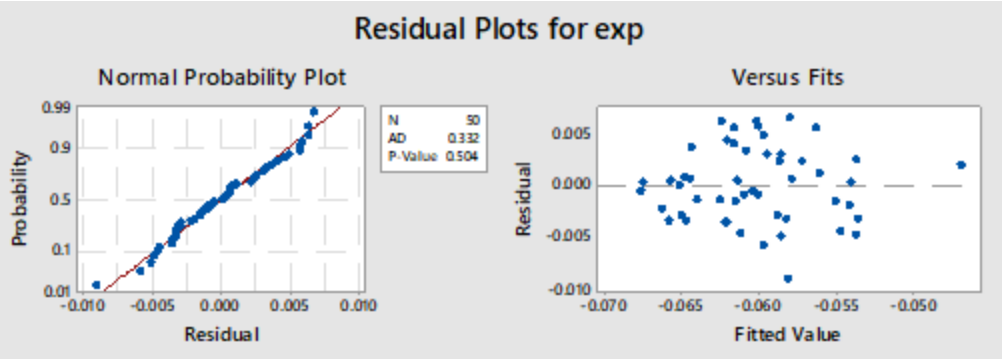
The procedure suggested by professor Crowson was follow in order to determine the weights. The WLS regression yields a studentized Breusch-Pagan test p value of 0.01158, which is the same p value for the OLS model. The plots looks also very similar:



Another variable in the data set is urban population. A new WLS model was run taking the reciprocal of the urban population as weights, having exactly the same results as the previous WLS model.

A box cox transformation yields a lambda of -0.5. The transformed model of the dependent variable shows a studentized Breusch-Pagan test p-value p-value = 0.2297; the transformed model plots in Minitab 18 is:

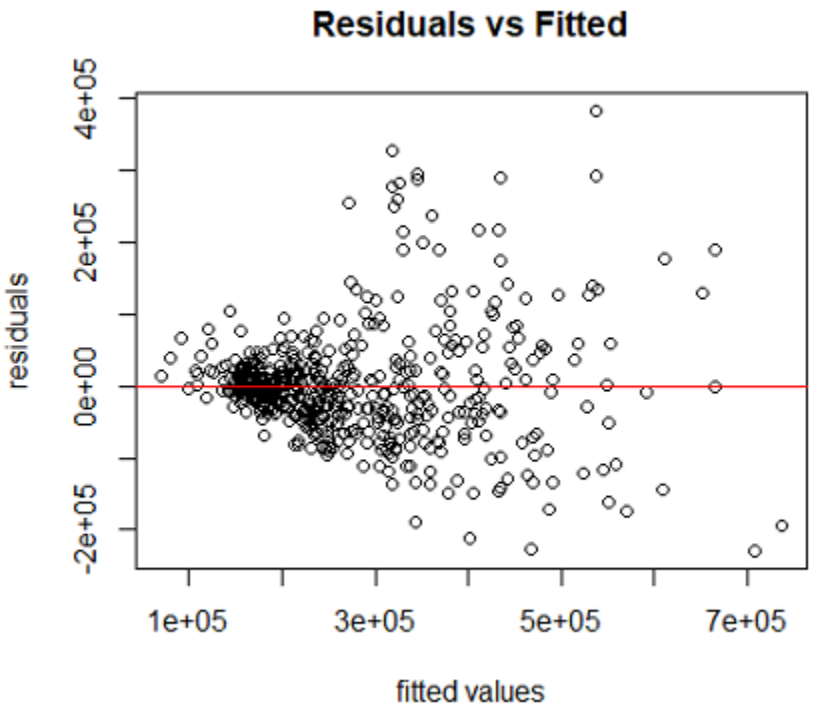
[14]Education expenditures 975, R data set



Showing a more homoscedastic model as anticipated by the Breusch-Pagan test.

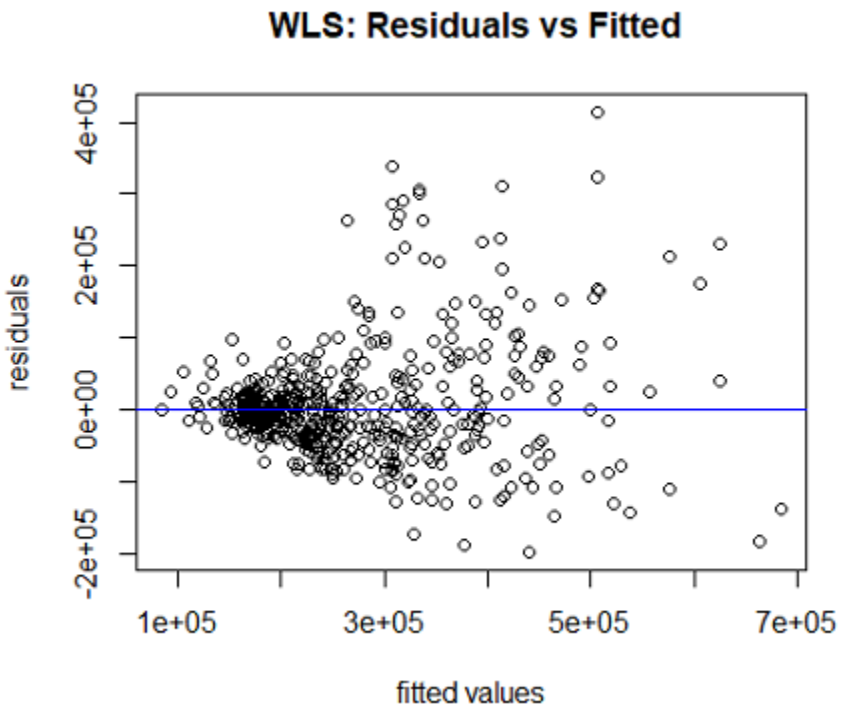
Case study III: Home price data set: University of Pennsylvania website: Home price as a function of area of the lot and area of the house, in square feet[15].

The studentized Breusch-Pagan test on the OLS model yields a p-value of zero (¡ 2.2e-16); the plot of the residuals shows a megaphone shape. note: a test for collinearity was conducted, the VIF for both variables are about 1. Thus, collinearity between independent variable does not exist.

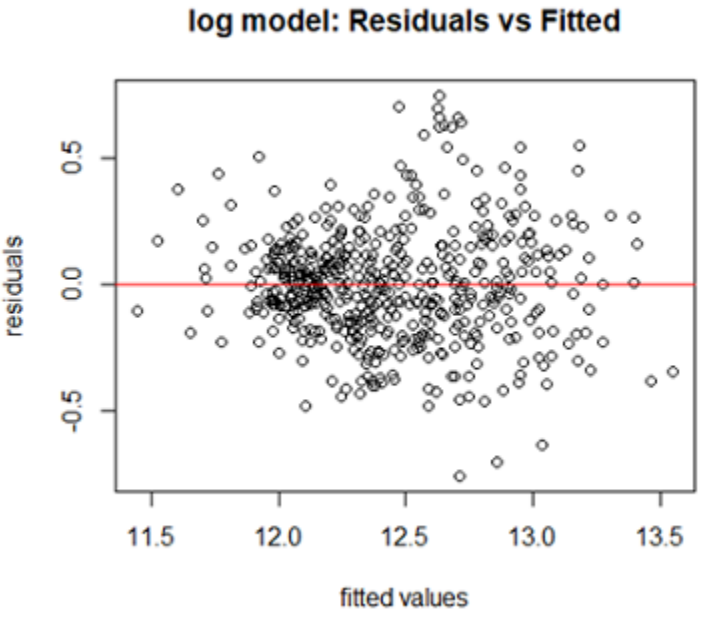


A WLS model was run. The weights calculated as suggested by Wooldridge,J. M. However, the Breusch-Pagan test shows the same very significant p value, and the scatter plot of the results looks exactly the same:

[15] *https* : *//onlinecourses.science.psu.edu/stat*501/*node*/397



Penn State website suggest to apply log transformation of both, dependent and independent variables. Once a log model is calculated, the plot of the residuals improves:



The log model eliminate the megaphone shape; however, the Breusch-Pagan test p-value = 3.575e-08 still very low.
At this point experience seems to confirm what theory anticipates: it is very difficult to find the weights that make the WLS model an alternative to OLS model when heteroscedasticity is present. To the rescue of the analyst come the theoretical fact that heteroscedasticity doesnt affect the parameters of the models, but the standard errors. In the original paper in Econometrica, 1980, Halbert White proved that his covariance matrix is consistent in the presence of heteroscedasticity. In R, the code to implement the covariance matrix (robust estimators covariance matrix) is:

> library(lmtest)

> coeftest(model, vcov=hccm)

In our case of the house prices, the log model is: log(Price) = log(Ah) + log(A lot)

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.96361    0.31286   6.276 7.33e-10 ***
log(Ah)      1.21976    0.03401  35.867  < 2e-16 ***
log(Alot)    0.11034    0.02412   4.575 5.97e-06 ***
```

The standard errors change after the robust estimator function is applied:

```
t test of coefficients:

            Estimate Std. Error t value  Pr(>|t|)
(Intercept) 1.963614   0.351710  5.5831 3.815e-08 ***
log(Ah)     1.219758   0.033665 36.2327 < 2.2e-16 ***
log(Alot)   0.110340   0.029938  3.6856 0.0002522 ***
```

It is shown that the standard errors of the parameters (except for the intercept in this case) is smaller with respect to the OLS model. That is something we expect from a WLS model, as proven by A. Shen, regression analysis, p121: $var(a'b_{wls}) \leq var(a'b_{ols})$ In short, the effect of robust estimators may well have the same implications that the WLS model: reduce the standard errors of the estimates that, otherwise would be inflated due to the non constant variance in the model.

Conclusions:

Finding appropriate weights for WLS models becomes very inefficient in practice. In some instances, when heteroscedasticity of the model is not severe, the box cox transformation not only solve the lack of normality of the residuals but often the non-constant variance issue. In cases where none of the transformations, neither the WLS model solve the non-constant variance in the model, a robust regression (robust estimators) may be the solution. Since Heteroscedasticity does not result in biased parameter estimates, finding robust estimators is an alternative: robust estimators are easy to implement in R and in other statistical software.

References

[1] A. Sen; M. Srivastava, Regression Analysis. Springer-Verlag, New York, 1990.

[2] Bellio, R. An introduction to robust estimators with R functions. Dept of Statistics, University of Udine. Retrieve from https://www.researchgate.net/publication/228906268

[3] Montgomery, D. Introduction to Regression Analysis, Willey, New York, 2003.

[4] Williams, R. Heteroskedasticity, University of Notre Dame. Retrieved from https:/www3.nd.edu/ rwilliam.

[5] Wooldridge,J. M. Econometrics, a modern approach, Cengage Learning, Mason, OH, 2012.

[6] Zaman, A. The Inconsistency of the Breusch-Pagan Test, Journal of Economic and Social Science research 2(1), 2000, 1-11.