Teaching Quantitative Methods Vignettes

Tutorial on Heteroskedasticity
using *HeteroskedasticityV3 SPSS macro*

Ahmad Daryanto [a] ✉ iD

[a]Lancaster University

Abstract ■ In this paper, I demonstrate how to assess the heteroskedasticity problems in cross-sectional studies that use linear regression models using my *HeteroskedasticityV3* SPSS macro. I present two illustrative examples inspired from real research. This paper also provides the annotations of the macro outputs. In my classroom demonstrations, students were asked to analyse data sets used in this paper and discuss their regression results with and without implementing robust standard errors. The merits of checking for the presence of heteroskedasticity prior to adjusting robust standard errors were also discussed in class.

Keywords ■ Heteroskedasticity, cross-sectional studies. **Tools** ■ SPSS, R.

Concept to be presented

Homoskedasticity is one of the basic assumptions in OLS (ordinary least squares) regression, which states that the regression error terms should have the same spreads across any values of independent variables. For simplicity, let us consider a simple linear regression model of customers' satisfaction scores on types of customers described by the following equation:

$$\text{Satisfaction}_i = \beta_0 + \beta_1 Loyal_i + \mu_i \quad (1)$$

for $i = 1, ..., N$, where N equals the number of customers in a random sample. $Loyal_i$ is a dummy variable that equals 1 for loyal customers and equals 0 for non-loyal customers. The error term μ_i is normally distributed with mean equals zero and variance equals $Var(\mu_i|Loyal_i) = \sigma_i^2$. The error terms are assumed to be independent, which means that they are not correlated, i.e., $Cov(\mu_i, \mu_j) = 0, for\ i \neq j$. [1]. For the above model, the homoskedasticity assumption requires the variance of the error terms to be constant across loyal vs. non-loyal customers. That is, $Var(\mu_i|Loyal_i = 1) = Var(\mu_i|Loyal_i = 0) = \sigma^2$, or simply put, $\sigma_{loyal}^2 = \sigma_{non-loyal}^2$. This implies that the variance of the distribution of satisfaction scores is the same for

loyal and non-loyal customers in a population. The distributions of the error terms within each group are assumed to be normal with mean zero.
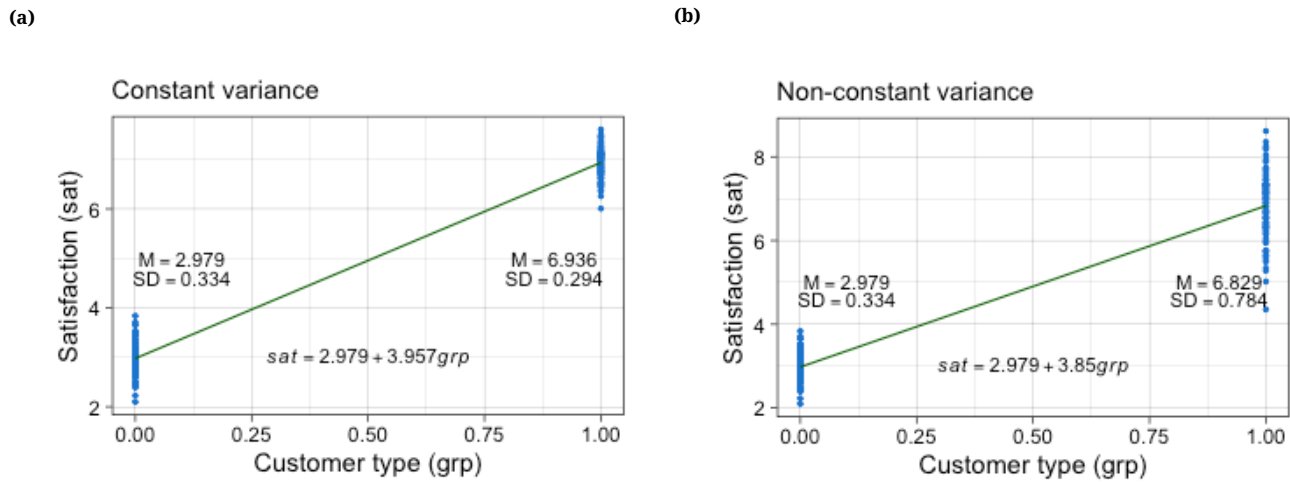
Homoskedasticity is a special case of heteroskedasticity. In this particular example, error terms μ_i are heteroskedastic if the variance of the error terms for the loyal customers is not the same as that of the non-loyal customers, i.e., $Var(\mu_i|Loyal_i = 1) \neq Var(\mu_i|Loyal_i = 0) = \sigma_i^2$, or simply put, $\sigma_{loyal}^2 \neq \sigma_{non-loyal}^2$. In general, heteroskedasticity exists if the variance of the error is not constant at any values of independent variables. In OLS regression that assumes homoskedastic errors, the variance of the regression parameters can be calculated according to the following formula:

$$Var(b) = \sigma^2(X^T X)^{-1} \quad (2)$$

where b is a vector of the coefficient estimates. For two parameters as in Equation 1, $b = (b_0\ b_1)^T$ where $E(b) = \beta$. X is an $N \times K$ matrix of independent variables where K is regression parameters including a constant, which equals 2 (i.e., the intercept and the slope) in Equation (1). The symbol σ^2 denotes the variance of the error terms, i.e., $Var(\mu) = \sigma^2$. However, when error terms are het-

---

[1]Hereafter, I use the term residuals instead of errors if the parameters of the above model are estimated using a random sample

Figure 1 ■ Scatter plot of customer type and satisfaction scores.

(a)

(b)

eroskedastic, the above formula to compute the standard errors of the regression coefficients are not correct anymore as σ^2 values are not constant across values of X. The general form of the variance of the regression coefficients can be seen below (see e.g., Stock & Watson, 2015; Davidson & MacKinnon, 2004; Fox, 2015):

$$Var(b) = (X^TX)^{-1}X^T\Omega X(X^TX)^{-1} \qquad (3)$$

where Ω is the covariance matrix of the error terms. If error terms are homoskedastic ($\Omega = \sigma^2\mathbf{I}$, where $\mathbf{I}$ is the $N \times N$ identity matrix), Equation (3) simplifies to Equation (2). My *heteroskedasticityV3* macro presented in this paper implements this general formula as a basis in calculating various options to adjust standard errors, which will be explained later. In literature, Equation (3) is also referred as the heteroskedasticity-consistent covariance matrix (HCCM) [2] and is also known as the sandwich covariance matrix (Davidson & MacKinnon, 2004).[3]

For a simple linear regression model, as we are interested in the variance of the coefficient estimate of b_1, the variance is given by the following formula:

$$Var(b_1) = \frac{\sum_{i=1}^{N} a_i^2\sigma_i^2}{\sum_{i=1}^{N} s_i^2} \qquad (4)$$

For homoskedastic errors where σ_i^2 is the same for all i, the variance of b_1 is given by

$$Var(b_1) = \frac{\sigma^2}{S_{xx}^2} \qquad (5)$$

where $a_i = x_i - \bar{x}$ (i.e., subtracting the mean of x_i's from x_i), and $S_{xx} = \sum_{i=1}^{N} a_i^2$ is the sum of squares of the a_i's. These results are covered in the textbooks mentioned above and presented here to make readers aware of the subtle differences in the formula for computing standard errors for b_1 when heteroskedasticity exists (Equation (4)) and when it does not (Equation (5)). I present the details of these derivations in the appendix.

The homoskedastic vs. heteroskedastic regression situations are typically explained using a graphical method. Figure 1 provides an illustration. In Panel A of Figure 1, the variance of satisfaction scores of loyal customers is roughly equal to that of the loyal customers. This is indicated by the non-significant difference in the standard deviations of the satisfaction scores between the two groups of customers. As shown in the figure, the vertical spreads of satisfaction across the two groups are relatively equal. Equal spreads in the satisfaction scores implies that the regression residuals will also have the same spreads. Thus, Panel A exhibits homoskedastic errors situations.[4]

In Panel B of Figure 1, the variance of satisfaction scores across the two groups of customers are unequal. As can be seen in the figure, the vertical spreads of the satisfaction scores across the two groups are unequal—satisfaction scores are more spread out for loyal customers than those for non-loyal customers. Unequal spreads in the satisfaction scores implies that the regression residuals will also have unequal spreads (Stock & Watson, 2015). Thus, Panel B exhibits heteroskedastic error situations. In both panels, simple regression lines and coeffi-

---

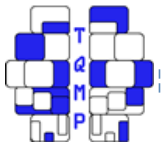[2]The word 'corrected' is also popular instead of 'consistent'.

[3]Due to the fact that $X^T\Omega X$ is placed between two slices of $(X^TX)^{-1}$.

[4]Instead of plotting the actual values, boxplots are recommended to use as it contains more information about data e.g., skewness, outliers, median, min and max values.

cient estimates were also plotted. Readers can easily verify the relationship between the coefficient estimates and the mean satisfaction scores of loyal vs. non-loyal customers.

When heteroskedasticity problem exists, the standard errors of the OLS regression coefficients will be biased. However, the regression coefficients will still be unbiased (i.e., over repeated samples, on average the estimates will be equal to population parameters) and OLS is still consistent (i.e., will get closer to population parameters as sample gets larger) (Stock & Watson, 2015). If regression errors are heteroskedastic but standard OLS is applied, the t and F test used to test hypotheses and confidence intervals associated with regression coefficients are wrong because the standard errors are wrong. Subsequently, OLS results could lead to an erroneous conclusion in regard to hypothesis testing. For example, non-significant relationship in a population can be significant in a sample with heteroskedastic errors. A simple remedy that has been recommended in the literature is to adjust the standard errors using the heteroskedasticity-adjusted or robust standard errors (Stock & Watson, 2015).[5]

This tutorial aims to demonstrate the use of my *HeteroskedasticityV3* SPSS macro to assess the impact of the heteroskedasticity problems on hypothesis testing. The macro allow users to eyeball the OLS results with or without the heteroskedasticity adjusted-standard errors simultaneously for a direct comparison. In addition, the macro produces outputs that includes results of heteroskedasticity tests (i.e., Breusch - Pagan and Koenker tests) and a graphical output for the assessment of the potential presence of heteroskedasticity. This tutorial provides two illustrative examples with hypothetical data sets that can be used in a classroom demonstration. In the next section, I explain the features of the macro and present two illustrative examples.

The HeteroskedasticityV3 *macro*

My *HeteroskedasticityV3* macro is written in both spd and spe format. As mentioned previously, it produces results of OLS with and without adjusted-standard errors, contains heteroskedasticity tests and automatically produces a scatter plot of fitted values vs. residuals. The *HeteroskedasticityV3* macro will appear in SPSS drop-down menu after being installed and outputs standard OLS results and those with heteroskedasticity-adjusted standard errors.

There are four options available on the macro for adjusting the standard errors due to the heteroskedasticity of unknown form: HC0, HC1, HC2, HC3, and HC4. HC0 is the well-known White's estimator (White, 1980). In HC0, the error term Ω in the HCCM of Equation (3) was estimated by the variance of the OLS residuals (e_i^2). HC1 and HC2 were developed by MacKinnon and White (1985). HC1 improves HC0 by adjusting for the degrees of freedom, i.e., $\frac{N}{N-K}$. HC2 modified HC1 by adjusting for the leverage value of observations (h_i). These $h_i's$ lie in the main diagonal of the projection matrix $P = X(X^T X)^{-1} X^T$. HC3 and HC4 was developed by Davidson and MacKinnon (1993) and Cribari-Neto (2004), respectively. HC3 modified HC2 slightly and HC4 modified HC3 by adjusting residuals with a leverage factor. I present the modifications of the matrix Ω to estimate $X^T \Omega X$ to calculate each of these options below:

$$\Omega_{HC0} = diag(e_i{}^2) \tag{6}$$

$$\Omega_{HC1} = \frac{N}{N-K} \; diag(e_i{}^2) \tag{7}$$

$$\Omega_{HC2} = diag\left(\frac{e_i{}^2}{1-h_i}\right) \tag{8}$$

$$\Omega_{HC3} = diag\left(\left(\frac{e_i{}^2}{1-h_i}\right)^2\right) \tag{9}$$

$$\Omega_{HC4} = diag\left(\frac{e_i{}^2}{(1-h_i)^{\delta_i}}\right), \delta_i = min\left\{4, \frac{Nh_i}{K}\right\} \tag{10}$$

Based on simulation study of Long and Ervin (2000), they found that HC3 performs well in a small sample ($N \leq 250$) - the statistical power associated with testing a hypothesis is higher than that of other adjustments. Later Cribari-Neto (2004) demonstrated that HC4 performs well with small samples and recommended researchers to use it when observations have a high leverage values and errors are not normally distributed. For these reasons HC4 is set as a default of the macro. (More detail explanations of each of these options and their differences can be read in the following articles: Hayes and Cai; Curto, Pinto, Morais, and Lourenço; Rosopa, Schaffer, and Schroeder; MacKinnon, 2007, 2011, 2013, 2013).

The *HeteroskedasticityV3* macro reports the results of two statistical tests commonly used to test the homoskedasticity assumption, namely the Breusch-Pagan (BP) and Koenker tests. The results of the two outputs could differ because BP is a large sample test and is not suitable for small samples. If data is non-normal, which is more realistic in real settings (Koenker, 1981), users should rely on the results of the Koenker test. For this particular data, both tests are significant, showing that heteroskedasticity is present. As the heteroskedasticity is present in the data, as mentioned in the introduction, the standard errors, p-values and confidence intervals associated with the regression coefficients are not correct anymore. It is important to note that, the conclusions are based on p-values and false conclusions can occur due to a lack of statistical power. Past studies shows that Koenker test outperforms BP test in terms of power (Lyon & Tsai, 1996).

---

[5]I used the word 'adjusted' and 'robust' interchangebly throughout the text as both terms are equally popular in literature.
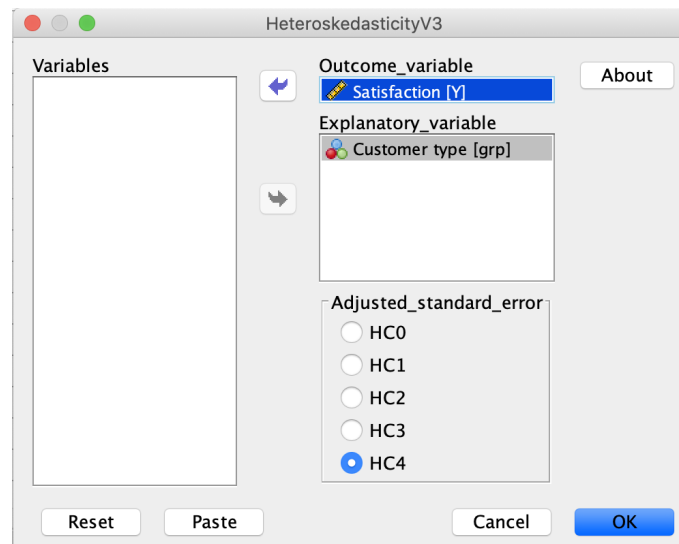
Figure 2 ∎ The HeteroskedasticityV3 macro

Illustration 1: The effects of types of customers on satisfaction

To illustrate the heteroskedasticity problem, it is best to start with a simple linear regression model with a dummy (0,1) independent variable and a continuous dependent variable. This is because the variance or the standard deviation of the dependent variable and residuals can be easily calculated at the two discrete values of the dummy variable allowing users to inspect the differences in standard deviations of the dependent variable. Furthermore, as two independent samples t-test was usually taught prior to learning regression, users might recall and relate the use of the Levene's test of the equality of variances to the tests commonly used in the context of linear regression models. However, as stated by Rosopa et al. (2013) "Levene's (1960) test may not have adequate statistical power to detect violations of the homogeneity of variance assumption when testing for the equality of two independent means." (p. 347).
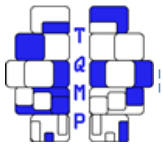
To this end, I use satisfaction scores as a dependent variable and types of customer as a dummy (0,1) independent variable with two groups: loyal vs. non-loyal customers. For example, loyal customers are those who holds loyalty memberships and non-loyal customers are regular customers that do not possess loyalty cards. In essence, I am using the simple linear regression model shown earlier in Equation (1). For this illustration, the macro input is shown in Figure 2.

Data for this illustration (satisfaction.sav) is generated using SPSS syntax in an sps format (see Listing 1) or R codes (see Listings 2). The SPSS syntax format is suitable for those that has no or limited background in R. In both formats, users can modify parameters in the codes to suit their modelling purposes. The satisfaction scores for loyal vs. non-loyal of customers were generated according to a normal distribution with different means and standard deviations. Heteroskedasticity is infused to the data by setting the standard deviation of the satisfaction scores of loyal customers to be higher than that of non-loyal customers. Hence, the form of the heteroskedasticity is known beforehand. In reality, the form of the heteroskedasticy in mostly unknown. The mean satisfaction scores of the loyal customers are also set to be higher than that of the non-loyal customers reflecting a more realistic situation (see e.g., Meyer-Waarden, 2008). In the generated sample, the means and standard deviations of satisfaction scores of the two groups are: $M_{loyal} = 2.979$, $SD_{loyal} = 0.334$; $M_{non-loyal} = 6.829$, $SD_{non-loyal} = 0.784$. Conducting the two–independent samples t-test on the data, the Levene's test of the equality of variances will be rejected while mean difference is significant.

Interpretations of the HeteroskedasticityV3 outputs

The outputs can be seen in Listings 3 (see appendix) and will be annotated as follows. First, the macro produces information on the name of the dependent variable (line 8) and the R^2 value (line 11). The OLS outputs under homoskedasticity and heteroskedasticity assumption are presented in lines 13-16 and lines 19 - 21, respectively. The

outputs report the coefficient estimates (b), standard errors (se), t-values (t), p - value (sig) and the lower (95%LB) and upper bound (95%UB) of the 95% confidence interval. The default option for heteroskedasticity-adjusted or robust standard errors is HC4 and is noted in line 23. For this example, as expected, the coefficient estimates for customer type denoted by variable grp in the outputs are unchanged after their standard errors are adjusted. [6] The standard error for the intercept after the adjustment is slightly lower than that of without adjustment, while the standard errors before and after the adjustment is relatively the same. Thus, for hypothesis testing, the conclusion regarding the significance of customer type (p = 0.000) is unchanged after adjustment. The ANOVA table from OLS with robust standard error is presented in lines 26 - 29.

Lines 33 - 59 produce outputs for the Breusch - Pagan and Koenker test. Important results are presented in lines 53 - 55. For this particular example, both tests are significant (p = 0.000) indicating the potential presence of heteroskedasticity. To augment this analysis, the *HeteroskedasticityV3* also produces a scatter plot of fitted values and residuals placed at the end of the output page. I have modified the plot and presented it earlier in the Panel B of Figure 1. As been previously explained, the scatter plot suggests that the regression error terms are heteroskedastic as the residuals are more spread out for loyal customers than those for non-loyal customers.

It is important to note that some researchers (see e.g., Ng & Wilcox, 2011; Long & Ervin, 2000; Stock & Watson, 2015) do not recommend to perform heteroskedasticity tests as it can lead to poor control over Type I errors. Instead, they recommend to directly apply the heteroskedasticity-adjusted standard errors without prior testing for the homoskedasticity assumption as according to them applying the standard error adjustment "..offer reasonable control over Type I errors under both homoscedasticity and heteroscedasticity" (Ng & Wilcox, 2011, p. 256). However, as pointed out by Wooldridge (2013), regardless of sample size, if the homoskedasticity assumption is not rejected, then the t-statistics from OLS will have an exact t distributions (Wooldridge, 2013, p. 273) and t statistics calculated using adjusted-standard errors in small samples will not be exactly distributed according to t distributions. Thus, there is an advantage of applying heteroskedasticity tests. For large cross-sectional data, Wooldridge (2013) recommends to always report the heteroskedasticity-adjusted standard errors together with the usual OLS standard errors. My *HeteroskedasticityV3* macro conveniently facilitates this task for users as both results are presented simultaneously in the outputs.
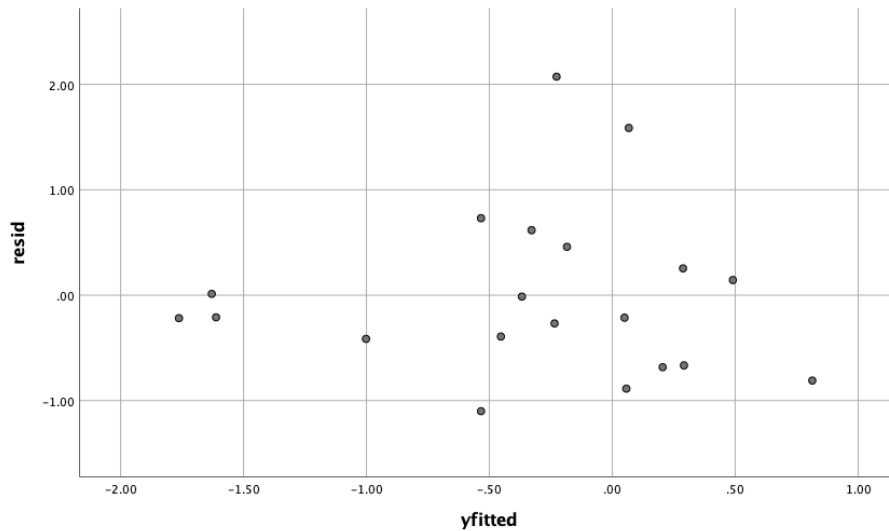
Illustration 2: The effects of team characteristics on team stability

For the second illustration (slotegraaf.sav), I chose group dynamics as a research context. When I used this example in my class, I pointed out to my students that this is a realistic setting that might occur to them when working in a group on their coursework. This illustration is inspired by a research paper by Slotegraaf and Atuahene-Gima (2011) that examined the decision-making processes of cross-functional teams defined as a group of workers of different management functions responsible for executing a project. Data for this illustration was generated from a correlation matrix reported in Slotegraaf and Atuahene-Gima (2011). Among several constructs reported in the paper, I chose three constructs that would be easy to grasp without reading the article prior to the class. These constructs are project team stability, project team size and project team-level debate. In the data set, these constructs are represented with variables tstability, teamsize and tdebate, respectively. My aim is to examine the influence of two team characteristics (i.e., project team size and project team level debate) on project team stability. By using the *HeteroskedasticityV3* macro, I want to show how controlling for heteroskedasticity could alter initial research conclusions derived from standard OLS that violates the homoskedasticity assumption. There are two hypotheses being tested. In the first hypothesis, team size is predicted to be negatively related to team stability – a large team could be unstable (e.g., members come and go during a project). In the second hypothesis, team-level debate is predicted to be negatively related to team stability – heated debates can make a team unstable (e.g., members resign).

Listing 4 shows the macro outputs. As can be seen in the standard OLS outputs. Both predictor variables (i.e., teamsize and tdebate) are significant ($b_{teamsize} = 0.514, p = 0.042; b_{tdebate} = -0.489, p = 0.036$). Note that despite being significant, the sign of $b_{teamsize}$ is positive, which is counterintuitive and contradicts our hypothesis. After applying the heteroskedasticity-robust standard errors, results show that team size and team-level debate are no longer significant although the p-value of the later is very close to being significant ($b_{teamsize} = 0.514, p = 0.128; b_{tdebate} = -0.489, p = 0.054$). The scatter plot of the fitted values vs. residuals is presented in Figure 3. The plot suggests that heteroskedasticity may be present in the data as the vertical spread of the residuals gets larger as the fitted value increases. In fact, the residuals on the left of the plot, i.e., below fitted values = -1, appear to be very cluttered compared to those on the right. This graphical inspection is confirmed by the results of the Breusch - Pa-

---

[6]Note that *grp* is a 0, 1 dummy variable – dummy variable should be coded as 0 or 1.

Figure 3 ■ Scatter plot of fitted values of tstability vs. residuals

gan and Koenker test where both tests were significant (p < 0.05).

Note that for this illustration, data was generated using information reported in a published paper (correlations, means and standard deviations) [7]. The heteroskedasticity that was detected was inherent in the generated data and its form is unknown. For the sake of learning, users can generate data by specifying how heteroskedasticity arises, for example by setting residuals variance depends on an independent variable. This task is more convenient to do in R than SPSS as the whole analysis as reported above can be coded in one script. I provided the R codes shown in Listing 5 to show one way of generating data with heteroskedasticity of known forms.[8]

In-classroom activities

The materials in this paper have been largely used in my classroom demonstrations at postgraduate-level module on applied quantitative research methods - without presenting mathematical formulas and derivations. In my class, first, I explained the concept of heteroskedasticity using the satisfaction. sav data. Next, students were asked to analyse the slotegraaf.sav presented in this paper and discuss their regression results with and without implementing robust standard errors. Next, the merits of checking for the presence of heteroskedasticity prior to adjusting robust standard errors were discussed. This strategy of teaching heteroskedasticity were well-received by my students who did not have strong background in statistics. One informal feedback I received was that "This is a difficult concept which is in need of the knowledge of statistics, but the instructor uses detailed and simple explanations to help us understand". Other students including two PhD students and one of my colleagues praise the simplicity of using the macro and the lay-out of the outputs that present the OLS results with and without adjusted standard errors next to each other.
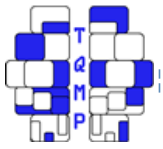
Conclusion

Heteroskedasticity problems remain a challenging topic in regression as well as challenging task in explaining the concept to non-statistical oriented audiences. In fact, after several decades since the seminal paper by White (1980)[9], the topic of finding new statistical methods to test and control for heteroskedasticity of an unknown form in cross-sectional data remains an active area of research (see e.g., MacKinnon, 2013; Lu & Wooldridge, 2020). For those who want to deepen their knowledge further, I recommend them to read these particular articles: MacKinnon (2013), Rosopa et al. (2013), Hayes and Cai (2007), which have been mentioned in previous sections.

Lastly, In this paper I have introduced my *HeteroskedasticityV3* SPSS macro to deal with heteroskedasticity problems in linear regression models that use cross-

[7]The sample size of the generated data was much smaller than that of reported in the published paper.
[8]Thanks to an anonymous reviewer who shared his/her R codes to replicate results of illustration 2 lines (18 - 41) modified in Listing 5.
[9]or Eicker-Huber-White's estimator to honour White (1980)'s predecessors: Eicker (1967) and Huber (1967).

sectional data. As aforementioned in the previous paragraph, the macro includes two statistical tests commonly used in practice (i.e., Breusch - Pagan and Koenker test), four heteroskedasticity adjusted-standard error options to control for heteroskedasticity and a scatter plot of fitted values vs. residuals. My aim is that the macro would be useful for applied researchers, practitioners and graduate students to assess heteroskedasticity problems in their works using SPSS.
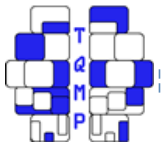
References

Cribari-Neto, F. (2004). Asymptotic inference under heteroskedasticity of unknown form. *Computational Statistics & Data Analysis*, *45*(2), 215–233.

Curto, J. D., Pinto, J. C., Morais, A. I., & Lourenço, I. M. (2011). The heteroskedasticity-consistent covariance estimator in accounting. *Review of Quantitative Finance and Accounting*, *37*(4), 427–449.

Davidson, R., & MacKinnon, J. G. (1993). Estimation and inference in econometrics. *OUP Catalogue*.

Davidson, R., & MacKinnon, J. G. (2004). *Econometric theory and methods*. Oxford University Press New York.

Eicker, F. (1967). Limit theorems for regressions with unequal and dependent errors. In *Proceedings of the fifth berkeley symposium on mathematical statistics and probability, volume 1: Statistics* (pp. 59–82). Berkeley, Calif.: University of California Press. Retrieved from https://projecteuclid.org/euclid.bsmsp/1200512981

Fox, J. (2015). *Applied regression analysis and generalized linear models*. Sage Publications.

Hayes, A. F., & Cai, L. (2007). Using heteroskedasticity-consistent standard error estimators in ols regression: An introduction and software implementation. *Behavior Research Methods*, *39*(4), 709–722.

Huber, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. In *Proceedings of the fifth berkeley symposium on mathematical statistics and probability, volume 1: Statistics* (pp. 221–233). Berkeley, Calif.: University of California Press. Retrieved from https : / / projecteuclid . org / euclid . bsmsp/1200512988

Koenker, R. (1981). A note on studentizing a test for heteroscedasticity. *Journal of Econometrics*, *17*(1), 107–112.

Long, J. S., & Ervin, L. H. (2000). Using heteroscedasticity consistent standard errors in the linear regression model. *The American Statistician*, *54*(3), 217–224.

Lu, C., & Wooldridge, J. M. (2020). A gmm estimator asymptotically more efficient than ols and wls in the presence of heteroscedasticity of unknown form. *Applied Economics Letters*, *27*(12), 997–1001.

Lyon, J. D., & Tsai, C.-L. (1996). A comparison of tests for heteroscedasticity. *Journal of the Royal Statistical Society: Series D (The Statistician)*, *45*(3), 337–349.

MacKinnon, J. G. (2013). Thirty years of heteroskedasticity-robust inference. In *Recent advances and future directions in causality, prediction, and specification analysis* (pp. 437–461). Springer.

MacKinnon, J. G., & White, H. (1985). Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties. *Journal of Econometrics*, *29*(3), 305–325.

Meyer-Waarden, L. (2008). The influence of loyalty programme membership on customer purchase behaviour. *European Journal of Marketing*.

Ng, M., & Wilcox, R. R. (2011). A comparison of two-stage procedures for testing least-squares coefficients under heteroscedasticity. *British Journal of Mathematical and Statistical Psychology*, *64*(2), 244–258.

Rosopa, P. J., Schaffer, M. M., & Schroeder, A. N. (2013). Managing heteroscedasticity in general linear models. *Psychological Methods*, *18*(3), 335.

Slotegraaf, R. J., & Atuahene-Gima, K. (2011). Product development team stability and new product advantage: The role of decision-making processes. *Journal of Marketing*, *75*(1), 96–108.

Stock, J. H., & Watson, M. W. (2015). *Introduction to econometrics*.

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica: Journal of the Econometric Society*, 817–838.

Wooldridge, J. M. (2013). *Introductory econometrics: A modern approach* (5th). Nelson Education.

Appendix

Listing 1: SPSS syntax to generate data for illustration 1

```
1  SET RNG=MC SEED=2000000.
2  MATRIX.
3
4  /*** You can modify this part.
5  COMPUTE nobs = 100.  /* set the number of observations/cases.
```
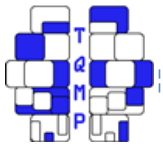
```
 6 COMPUTE one= make(nobs,1, 1).  /*create a dummy variable '1'.
 7 COMPUTE two= make(nobs,1, 0).  /*create a dummy variable '0'.
 8 COMPUTE mu1 = 7.  /*set the mean of group 1.
 9 COMPUTE sd1 = 0.8.  /*set the standard deviation of group 1.
10 COMPUTE mu2 = 3.  /*set the mean of group 2.
11 COMPUTE sd2 = 0.3.  /*set the standard deviation of group 1.
12
13 /*** Don't modify this part.
14 COMPUTE x = UNIFORM(nobs,2).  /* generate uniform random variates.
15 COMPUTE z = -ln((1/x)-1)/1.702.  /* transform to standard normal.
16 COMPUTE x1 = mu1 + z(:, 1)*sd1.  /* convert std normal to normal.
17 COMPUTE x2 = mu2 + z(:, 2)*sd2.  /* convert std normal to normal.
18 COMPUTE x1 = t({one, x1}).
19 COMPUTE x2 = t({two, x2}).
20 COMPUTE mydata =t({x1, x2}).  /* tidy up the data.
21 SAVE mydata/OUTFILE =* / VARIABLES = grp, Y.
22 END MATRIX.
```

Listing 2: R codes to generate data for illustration 1

```
 1 library(tidyverse)
 2 library(haven)
 3 library(fastDummies)
 4 set.seed(1233)
 5
 6 # Generate data.
 7 nobs <- 100
 8 m.loy <- 7
 9 sd.loy <- 0.8
10 m.nonloy <- 3
11 sd.nonloy <- 0.3
12 loyal <- rnorm(nobs, m.loy, sd.loy)
13 non_loyal <- rnorm(nobs, m.nonloy, sd.nonloy)
14
15 # Combine loyal and non_loyal and make it tidy.
16 dat <- data.frame(loyal) % > %
17   mutate(non_loyal) % > %
18   gather(key = "grp", value = "satisfaction") % > %
19   mutate(grp = as.factor(grp))
20
21 # Create dummy variable for grp.
22 dat <- dummy_cols(dat, select_columns = c('grp'),
23                     remove_selected_columns = FALSE)
24
25 # Store data into SPSS .sav format.
26 write_sav(dat,"/Users//Documents/Data/satisfaction.sav")
27 view(dat)
```

Listing 3: Outputs from the heteroskedasticity macro for illustration 1

```
 1 Run MATRIX procedure:
 2
```

3   written by Ahmad Daryanto

4

5 Original Regression model:

6

7 Dependent variable

8   Y

9

10 R-square

11   .912

12

13 OLS outputs

```
14              b          se         t         sig       95%LB      95%UB
15 constant 2.9790      .0603    49.4180      .0000    2.8602     3.0979
16 grp      3.8503      .0853    45.1633      .0000    3.6822     4.0184
```

17

18 OLS outputs with heteroskedasticity-robust standard errors:

```
19              b          se         t         sig       95%LB      95%UB
20 constant 2.9790      .0334    89.1141      .0000    2.9131     3.0450
21 grp      3.8503      .0853    45.1633      .0000    3.6822     4.0184
```

22

23 * Note: standard error is HC4 variant

24

25

26 ------- ANOVA TABLE --------

```
27               SS          df         MS           F          Sig
28 Model      741.2307     1.0000    741.2307   2039.7248       .0000
29 Residual    71.9527   198.0000       .3634   -999.0000   -999.0000
```

30

31 ==

32

33 Breusch-Pagan and Koenker test

34

35 ==

36

37 The tests use the scaled residuals from the original OLS above with no adjustment
    to standard errors.

38

39 OLS outputs

```
40              b          se         t         sig       95%LB      95%UB
41 constant  .3075      .1800     1.7087      .0891     -.0474       .6624
42 grp      1.3850      .2545     5.4415      .0000      .8830     1.8869
```
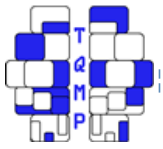
43

44 R-square

45   .130

46

47 ------- ANOVA TABLE --------

```
48               SS          df         MS           F          Sig
49 Model       95.9048     1.0000     95.9048     29.6096       .0000
50 Residual   641.3171   198.0000      3.2390   -999.0000   -999.0000
```

51

52 ---- Breusch-Pagan and Koenker test statistics and sig-values

```
53              LM          Sig
```
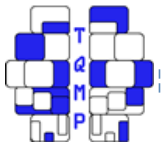
```
54 BP           47.9524       .0000
55 Koenker      26.0179       .0000
56
57 Null hypothesis:heteroskedasticity not present (homoskedasticity).
58 If sig-value less than 0.05, reject the null hypothesis.
59 Note: Breusch-Pagan test is a large sample test and assumes the residuals to be
      normally distributed.
```

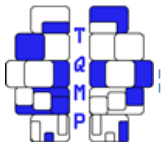### Listing 4: Outputs from the heteroskedasticity macro for illustration 2

```
 1 Run MATRIX procedure:
 2
 3  written by Ahmad Daryanto
 4
 5 Original Regression model:
 6
 7 Dependent variable
 8  tstabili
 9
10 R-square
11      .438
12
13 OLS outputs
14             b        se        t       sig      95%LB      95%UB
15 constant-1.9779    .8161   -2.4237    .0268   -3.6996    -.2561
16 teamsize  .5143    .2335    2.2028    .0417     .0217    1.0068
17 tdebate  -.4892    .2153   -2.2724    .0363    -.9435    -.0350
18
19 OLS outputs with heteroskedasticity-robust standard errors:
20             b        se        t       sig      95%LB      95%UB
21 constant-1.9779   1.0462   -1.8906    .0759   -4.1851     .2294
22 teamsize  .5143    .3210    1.6020    .1276    -.1630    1.1915
23 tdebate  -.4892    .2366   -2.0682    .0542    -.9883     .0099
24
25 * Note: standard error is HC4 variant
26
27 ------- ANOVA TABLE  --------
28              SS         df        MS         F         Sig
29 Model      9.4623    2.0000    4.7312    6.6345       .0074
30 Residual  12.1230   17.0000     .7131 -999.0000  -999.0000
31 ============================================
32
33 Breusch-Pagan and Koenker test
34
35 ============================================
36 The tests use the scaled residuals from the original OLS above with no adjustment
      to standard errors.
37 OLS outputs
38             b        se        t       sig      95%LB      95%UB
39 constant 3.6622   1.4044    2.6077    .0184     .6992    6.6251
40 teamsize 1.1915    .4018    2.9657    .0087     .3439    2.0392
41 tdebate   .6581    .3705    1.7763    .0936    -.1236    1.4398
```

```
42
43 R-square
44       .370
45
46 ------- ANOVA TABLE  --------
47                 SS          df          MS           F          Sig
48 Model       21.0477      2.0000     10.5238       4.9830       .0074
49 Residual    35.9029     17.0000      2.1119    -999.0000    -999.0000
50
51 --- Breusch-Pagan and Koenker test statistics and sig-values
52                 LM          Sig
53 BP          10.5238        .0052
54 Koenker      7.3916        .0248
55
56 Null hypothesis:heteroskedasticity not present (homoskedasticity).
57 If sig-value less than 0.05, reject the null hypothesis.
58 Note: Breusch-Pagan test is a large sample test and assumes the residuals to be
      normally distributed.
```

Listing 5. R codes to generate data with heteroskedasticity of known forms

```r
1 # R codes  related  to   Illustration   2
2 # Residual  variance  depends on  the  values  of  iv.
3
4 rm(list=ls())
5 library(lmtest)
6 library(sandwich)
7
8 obs <- 200 # number of observations
9 sigma.sq <- 0.2 # set variance
10
11 # generate  variables
12 teamsize <- rnorm(obs,10,2)
13 tdebate <- rnorm(obs,7,3)
14 ve <- sigma.sq*teamsize # set error variance
15 error <- rnorm(obs, 0, sqrt(ve))
16 tstability <- 0.2 - 0.3*teamsize - 0.4*tdebate + error
17
18 #OLS regression
19
20 reg.model <- lm(tstability ~ teamsize + tdebate)
21 summary(reg.model)
22 par(mfrow = c(2, 2))
23 plot(reg.model, ask=FALSE)
24
25 # Breusch — Pagan test
26 bp <- bptest(tstability ~ teamsize + tdebate, varformula = NULL, studentize = FALSE
      , data = list())
27 bp
28
29 # Koenker  test
```

```
30  kt <- bptest(tstability ~ teamsize + tdebate, varformula = NULL, studentize = TRUE,
        data = list())
31  kt
32
33  # calculate robust standard errors
34  coeftest(reg.model,vcov = vcovHC(reg.model,type="HC4"))
35
36  #compute 95%CI for robust se
37  covariance <- vcovHC(reg.model, type = "HC4")
38  t.critical <-qt(c(0.025,0.975),summary(reg.model)$df[2])
39  se <- sqrt(diag(covariance))
40  robust.ci <-coef(reg.model) + se %o% t.critical
41  robust.ci
```

Deriving the standard error of b_1 of a simple linear regression

Let a_i be $x_i - \bar{x}$ and $S_{xx} = \sum_{i=1}^{N} a_i^2$. The estimate of b_1 is given by (see e.g., Wooldridge, 2013):

$$b_1 = \frac{\sum_{i=1}^{N} a_i y_i}{S_{xx}} = \beta_1 + \frac{\sum_{i=1}^{N} a_i \mu_i}{S_{xx}} \tag{11}$$

where $y_i = \beta_0 + \beta_1 x_i + \mu_i$, $\mu_i \sim N(0, \sigma_i^2)$, $Cov(\mu_i, \mu_j) = 0$ for $i \neq j, i = 1, 2, ..., N$. β_0 and β_1 are the population parameters.

Using the last term of Equation 11 and taking the conditional variance:

$$Var(b_1|X) = Var((\beta_1 + \frac{\sum_{i=1}^{N} a_i \mu_i}{S_{xx}})|X)$$
$$= \frac{\sum_{i=1}^{N} a_i^2 \sigma_i^2}{S_{xx}^2}. \tag{12}$$

This result is obtained using the fact that we are taking the conditional variance on data assumed to be fixed despite being drawn randomly from a population, hence S_{xx} is also not a random variable. Furthermore, β_1 is a population regression parameter, which is a constant and mostly unknown – otherwise, we do not need to estimate it. We use the following properties of variance to get the final form: $Var(a) = 0, Var(aX) = a^2 Var(X)$, where a is a constant; $Var(\sum_{i=1}^{N} a_i \mu_i) = \sum_{i=1}^{N} a_i^2 Var(\mu_i)$ (i.e., the sum of the variance is the variance of the sum), and a_i is a constant.

Back to Equation 12, if errors are homoskedastic, $\sigma_i^2 = \sigma^2$ for all i, then the variance of b_1 is given by:

$$Var(b_1|X) = \frac{\sum_{i=1}^{N} a_i^2 \sigma^2}{S_{xx}^2}$$
$$= \frac{\sigma^2}{S_{xx}} \tag{13}$$

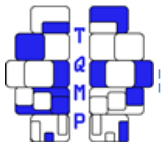***Macro Installation***

The heteroskedasticity SPSS macro is available both in .spd and .spe formats, which can be downloaded on my google website [10] or the journal's website. The .spd file can be installed on SPSS under version 24, 25, 26 and 27. The .spd file can be installed by simply double clicking the file. Alternatively, use the pull-down menu by clicking: Extensions → Utilities → Install Custom Dialog. The .spe file can be installed by clicking Extensions → Install Local Extension Bundle.

Open practices

⬤ The *Open Data* badge was earned because the data of the experiment(s) are available on the journal's web site.
⬤ The *Open Material* badge was earned because supplementary material(s) are available on the journal's web site.

Citation

Daryanto, A. (2020). Tutorial on heteroskedasticity using heteroskedasticityV3 SPSS macro. *The Quantitative Methods for Psychology*, *16*(5), v8–v20. doi:10.20982/tqmp.16.5.v008

Extended activity metadata

Concept illustrated	Heteroskedasticity	*Type of activity*	In-class
Prerequisite	*t*-test, ANOVA, simple linear regression	*Types of data*	Survey data
Co-requisite	Multiple linear regression	*Computation by*	R and SPSS
Suitable class size	75 postgraduate students	*Duration*	1 - 2 hours