Analysis of messy data with heteroscedastic in mean models

Cite as: AIP Conference Proceedings **1716**, 020009 (2016); https://doi.org/10.1063/1.4942992
Published Online: 29 February 2016

Nurvita Trianasari, and Cucu Sumarni

ARTICLES YOU MAY BE INTERESTED IN

Sensitivity analysis of Welch's t-test
AIP Conference Proceedings **1605**, 888 (2014); https://doi.org/10.1063/1.4887707

Analysis Of Messy Data With Heteroscedastic In Mean Models

Nurvita Trianasari[1,2,a,b] and Cucu Sumarni[3,4,c]

[1]Departement of Statistics, Student of Doctoral Program at Sekolah Pasca Sarjana-IPB
[2]Lecturer of Telkom Economics and Business School, Telkom University, Bandung
[3]Departement of Statistics, Student of Doctoral Program at Sekolah Pasca Sarjana-IPB
[4]Staf Badan Pusat Statistik Provinsi Bali.

[a)]vita.statistik@gmail.com
[b)]nurvitatrianasari@telkomuniversity.ac.id
[c)]cucu_s@bps.go.id

Abstract. In the analysis of the data, we often faced with the problem of data where the data did not meet some assumptions. In conditions of such data is often called data messy. This problem is a consequence of the data that generates outliers that bias or error estimation. To analyze the data messy, there are three approaches, namely standard analysis, transform data and data analysis methods rather than a standard. Simulations conducted to determine the performance of a third comparative test procedure on average often the model variance is not homogeneous. Data simulation of each scenario is raised as much as 500 times. Next, we do the analysis of the average comparison test using three methods, Welch test, mixed models and Welch-r test. Data generation is done through software R version 3.1.2. Based on simulation results, these three methods can be used for both normal and abnormal case (homoscedastic). The third method works very well on data balanced or unbalanced when there is no violation in the homogenity's assumptions variance.

For balanced data, the three methods still showed an excellent performance despite the violation of the assumption of homogeneity of variance, with the requisite degree of heterogeneity is high. It can be shown from the level of power test above 90 percent, and the best to Welch method (98.4%) and the Welch-r method (97.8%). For unbalanced data, Welch method will be very good moderate at in case of heterogeneity positive pair with a 98.2% power. Mixed models method will be very good at case of highly heterogeneity was negative negative pairs with power. Welch-r method works very well in both cases.

However, if the level of heterogeneity of variance is very high, the power of all method will decrease especially for mixed models methods. The method which still works well enough (power more than 50%) is Welch-r method (62.6%), and the method of Welch (58.6%) in the case of balanced data. If the data are unbalanced, Welch-r method works well enough in the case of highly heterogeneous positive positive or negative negative pairs, there power are 68.8% and 51% consequencly. Welch method perform well enough only in the case of highly heterogeneous variety of positive positive pairs with it is power of 64.8% . While mixed models method is good in the case of a very heterogeneous variety of negative partner with 54.6% power. So in general, when there is a variance is not homogeneous case, Welch method is applied to the data rank (Welch-r) has a better performance than the other methods..

Keywords: Messy Data, Welch test, fixed effect and random effects, mixed models, ANOVA and Welch-R test.

1. WHAT IS MESSY DATA?

In analyzing data, we often find some trouble in data, which the data cannot meet some assumptions, so the standarized analytic methods is not fitted. This condition is called messy data. Messy data are data which have not been cleaned or in other words, dirty data. This data are caused by incomplete data or incorrect data. (Hand, 2008).

A set of data is incomplete when there is some missing data. Some data can be missed randomly because of some causes which are not related with the research, for example, the non response case in data handling, error in sampling and outlier.

Application of Mathematics in Industry and Life
AIP Conf. Proc. 1716, 020009-1–020009-6; doi: 10.1063/1.4942992

According to (Harley and Lewis, 1963) data can be called messy when some assumptions of standardized model are not met. We can see it from error pattern from a model which deny some assumption, such as, the data don't have a normal distribution or the error variance is not homogen. The data which contains some error will disturb the assumption of the standard model.

Generally, to analyze messy data, there are three approaches. First, accept the data just like it is and do standarized analysis. However, this method could cause bias estimation, possibly creating a wrong conclusion. Second one, transform the data so we can do standarized analysis. But this method require an expert in finding the right transformation, also it is difficult to interpret the results. Third, accept the data just like it is, but find the analysis method which are more valid than standarized analysis. Unstandarize analysis is often needed because they are some violation to assumptions of standarized analysis, such as the sample sizes are not the same for each treatment, the distribution of the data is different from the one of a standarize analysis, There are an equal variances, there is outliers and mixed distribution, etc (Johnson and Milliken in the Encyclopedia of Statistical Sciences, 2006).

And it will be found out immediately after the examination. However messy data like any other is cause by non homogenous error or violation of assumption of the standard model certainly are not easy to be analyzed. It takes a certain method to detect it. Therefore, we will discuss about messy data cases which is caused by a violation of homogeneity variances.

On mean models or like in the experimental design, the assumption of homogeneity of error is one of the main requirements to do next analysis, such as the average comparative treatment test. If the experiment has been done and then we find out there is violation of assumption of homogeneity of variance, so there are two approaches methods which can be used. First, the transformation to stabilize the variance, so analysis of variance (ANOVA) can be used. However, the disadvantage of the first approach is more difficult to do, especially if the researcher does not know the theoretical distribution of experimental data (Montgomery, 2001) and the interpretation is not easy (Milliken and Johnson, 2009). The second approach is to do a variety of non-standard analysis of existing data, or without transformation (unequal variance models). The advantage of this second approach, it is easy to do (can use SAS software and R).

Milliken and Johnson (2009), proposed Welch test procedure and procedure mixed model for mean comparative test between treatments in a variety of models are not the same (unequal variance models). While David and Vikki, (2008) proposed nonparametric test procedure known as modern procedures which is robust to the assumptions of normality and homogeneity of variance. Nonparametric testing procedures are easy to implement, because it is robust one of the example to test on the data rank and Welch test data rank (Welch-r) (Cribbie, et al, 2007).

Thus, this paper will discuss the messy data because of the assumption of homogeneity of variance error in experimental designs, how to explore and analyze it. Beside, this paper wil discuss how to know the performance of all three mean testing methods when there is messy data due to a variety of not homogeneous errors variance.

2. HOW TO EXPLORE HETEROSCEDASTIC?

Irregularities in variety of error are also known as Heteroscedasticity. Detection towards heteroskedatisitas can be done through methods of informal and formal methods (Gujarati, 1995). Informal method is performed by plotting the residual or error model $\hat{\varepsilon}_{ij}$ to the fitted value $(\hat{y}_{ij})$. Data with homogeneous variance will generate residual plots, spreads around the value of $\hat{\varepsilon}_{ij} = 0$ with the same width of the distribution. While the data is heteroskedasticity, distribution of residu is will have the width of distribution which is not homogeneous. Patterns forming, funnel-shaped, linear or non linear pattern, cluster or non cluster around, i$\hat{\varepsilon}_{ij} = 0$ indicate the presence of heteroscedasticity.

To analyse the violation of assumption we can do informaly (residual plotting) and formally (error model). This model can be tested by the homogeneity of variance of treatment. For example, (in one-way treatment structure in a completely randomized design) with treatment as *t* and the number of observations $N = \sum_{i=1}^{t} n_i$, then the model can be written as follows:

$$y_{ij} = \mu_i + \varepsilon_{ij} \; for \; i = 1,2,\ldots,t, \quad j = 1,2,\ldots,n_i \tag{1}$$

In the standard model, a error model is assumed to have normal and independent distribution with zero mean and the variance is σ^2, $\varepsilon_{ij} \sim independent \; N(0,\sigma^2)$. If the analysis is not standard, then the error models are assumed $\varepsilon_{ij} \sim independent\ N(0, \sigma_i^2)$ meaning that error ε_{ij} all independent, normally distributed with variance varied for each treatment (population). The best estimation value based on the model are:

$$\hat{\mu}_i = \sum_{j=1}^{n_i} \frac{y_{ij}}{n_i} = \bar{y}_{i.}\ , \quad i = 1,2,\dots,t$$

and

$$\hat{\sigma}_i^2 = \frac{\sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i.})^2}{n_i - 1}$$

Testing homogeneity of variance is performed to test the hypothesis of homogeneity of variance as follow:

$$H_0 : \sigma_1^2 = \sigma_2^2 = \dots = \sigma_t^2 \text{ vs } H_a : (not\ H_0)$$

The recommended significance level to reject the null hypothesis (H_0) is $\alpha \le 0,01$. If there is not sufficient evidence to conclude that the variance among treatments is not the same, then the standard model (by assuming the error variance is the same), can be used for futher analizer. If there is sufficient evidence to considered that the variance is not the same, then we can use analytical technique to analyze further.

There are homogeneity of variance test procedures, such as: F-max Hartley test, Bartlett test, Levene test, Brown and Forsythe test and O'Brien test. Research shows that there is not any strongest procedures and most robust test for all situations. However Milliken and Johnson (2002) recommends many diagnostic tests for a different conditions, namely:

1. If the distribution have a heavy tail then used the Brown-Forsythe test.
2. If the distribution is skewed, then used O'Brien test.
3. If the data are approximately normal distributed then all of the test can be used.

Levene's test and O'Brien can be easily customized to be used in experiments which are designed to involve more than one factor, including analysis of covariance. Levene's and O'Brien test and Brown-Forsythe test have been proved as good as Bartlett and Hartley test for normally distributed data. However those test are better for not normally distributed data.

3. HOW TO ANALYSIS THE MEAN MODEL WITH THE HETEROSCEDASTIC?

Comparison of mean test among treatments is one of the goals in experimental designs. In unequal variance models, Welch test procedure and mixed models procedure (Milliken and Johnson, 2009) can be used for nonparametric procedure test, we can use Welch test or Welch-r test (Cribbie , et al, 2007).

1). Welch Test

Average ratio test introduced by Welch in 1951 make the researchers do the test on ratio average test with assumption that the homogeneity of variance error is violated and the data is unbalanced, a condition in which the sample size of each treatment is not the same , The null hypothesis is: $H_0 := \mu_1 = \mu_2 = \dots = \mu_t$ vs $H_a : (not\ H_0 :)$

Welch test is as follow:

$$F_c = \frac{\sum_{i=1}^{t} W_i \frac{(\bar{y}_{i\bullet} - \bar{y}^*)}{(t-1)}}{1 + 2(t-1)\Lambda / (t^2 - 1)} \tag{2}$$

Where $W_i = n_i / \hat{\sigma}_i^2$ is the weight specified,,

$\overline{y}^* = \sum_{i=1}^{t} W_i \overline{y}_i / \sum_{i=1}^{t} W_i$ is the weighted average of the sample

$\Lambda = \sum_{i=1}^{t} \frac{(1 - W_i / W_\bullet)^2}{n_i - 1}$ with $W_\bullet = \sum_{i=1}^{t} W_i$.

Welch test (2) has F-distribution with any given degrees $v_1 = t - 1$and $v_2 = (t^2 - 1)/3\Lambda$. Thus, the null hypothesis $H_0 := \mu_1 = \mu_2 = \ldots = \mu_t$ is rejected if $F_c > F_{\alpha, v_1, v_2}$. In the SAS program, Welch test can be obtained, through the PROC GLM procedure by specifying Welch as an option on the means statement (see Milliken and Johnson, 2009 page 35).

2). Mixed Model Procedure

Comparison of mean test based on mixed model procedures is similar with ANOVA, but it is performed on a mixed model. Mixed model is a model consisting of a fixed effect and random effect. First model (1) only contains a fixed effect mean.

The common forms of the mixture of linear models can be written as follow (Jiang, 2007):

$$y = X\beta + Z\alpha + \epsilon \qquad (3)$$

Where y is a vector of observation
X is the matrix of known covariate
β is regression coefficients vector or fixed effect
Z is the known matrix
α is a vector of random effect
ϵ is a vector error

If we assume that the experimental units in each treatment are chosen randomly, so the mixed models for model (1) can be written as:

$$y_{ij} = \mu_i + \alpha + \varepsilon_{ij} \, for \; i = 1,2, \ldots, t, \quad j = 1,2, \ldots, n_i \qquad (4)$$

Whre α is random intercept which is assumed to be independent with error model.

SAS procedure, (in the case PROC MIXED) can be used to fit the model with unequal variance among treatment groups. Syntax details can be found in Milliken and Johnson, 2009 page 37.

3). Welch-r Test

This test is similar with Welch test, but the data that is used, are rank transformed data.

4. SIMULATION STUDY

Simulations study are carried out to determine the performance of the three procedures on comparison of means test on non homogeneous variance model, with the aim of:

- To determine whenever the three procedures worked well in the case of unbalanced data (the sample sizes is not the same).
- To find out how large the variety of irregularities that can be tolerated so that the third test is still working well.

Simulation data generated by the model (1), with 4 treatment and the average treatment is defined as follows:
$\mu_1 = 1$, $\mu_2 = 3$, $\mu_3 = 5$, $\mu_4 = 7$.

The scenarios created in this simulation can be seen in Table 1.

Table 1. Simulation Scenario

Sample Size	The Condition of Variance Among Treatment	
The Same ($n_1 = n_2 = n_3 = n_4 = n$)	Homogenous ($s_1 = s_2 = s_3 = s_4 = s$)	$n = 10, s = 1$
	Moderate Heterogenous (range standard deviation: 1-4)	$n = 10, s_1 = 1, s_2 = 2, s_3 = 3, s_4 = 4$
	Highly Heterogenous (range standard deviation: 1-8)	$n = 10, s_1 = 1, s_2 = 3, s_3 = 5, s_4 = 8$
The Different	Homogenous	$n_1 = 3, n_2 = 5, n_3 = 8, n_4 = 12, s = 1$
	Moderate Heterogenous, positive positive pairs	$n_1 = 3, n_2 = 5, n_3 = 8, n_4 = 12, s_1 = 1, s_2 = 2, s_3 = 3, s_4 = 4$
	Moderate Heterogenous, negative negative pairs	$n_1 = 3, n_2 = 5, n_3 = 8, n_4 = 12, s_1 = 4, s_2 = 3, s_3 = 2, s_4 = 1$
	Highly Heterogenous, positive positive pairs	$n_1 = 3, n_2 = 5, n_3 = 8, n_4 = 12, s_1 = 1, s_2 = 3, s_3 = 5, s_4 = 8$
	Highly Heterogenous, negative negative pairs	$n_1 = 3, n_2 = 5, n_3 = 8, n_4 = 12, s_1 = 8, s_2 = 5, s_3 = 3, s_4 = 1$

Simulation data of each scenario is generated for 500 times, then the analysis of the average comparison test is performed by using three methods, namely Welch, mixed models and Welch-r test. The data generation is done using software R version 3.1.2. The simulation results can be seen in Table 2.

Table 2. The Percentage of power test of Welch test, mixed models procedures F test and Welch-r test based on 500 time generated data

Sample Size	The Condition of Variance Among Treatment	Welch	Mixed	Welch-r
The Same	Homogenous	100	100	100
	Moderate Heterogenous	**98,4**	92,8	97,8
	Highly Heterogenous	58.6	33	**62.6**
The Different	Homogenous	100	100	100
	Moderate Heterogenous, positive positive pairs	**98,2**	75	**98,2**
	Moderate Heterogenous negative negative pairs	88,8	**94,8**	91,8
	Highly Heterogenous, positive positive pairs	64,8	12	**68,8**
	Highly Heterogenous, negative negative pairs	37,4	**54,6**	51

Based on simulation results in Table 2, the third method can be done for normal and abnormal case (there is violation of homogeneity of variance assumption or homoscedastic). The third method works very well on balanced data (the sample sizes among the treatments is the same) or unbalanced data.

For balanced data, the three methods still showed an excellent performance despite the violation of the assumption of homogeneity of variance, with the requisite degree of heterogeneity is high. It can be shown from the level of power test above 90 percent, and the best to Welch method (98.4%) and the Welch-r method (97.8%). For unbalanced data, Welch method will be very good moderate at in case of heterogeneity positive pair with a 98.2% power. Mixed models method will be very good at case of highly heterogeneity was negative negative pairs with power. Welch-r method works very well in both cases.

However, if the level of heterogeneity of variance is very high, the power of all method will decrease especially for mixed models methods. The method which still works well enough (power more than 50%) is Welch-r method (62.6%), and the method of Welch (58.6%) in the case of balanced data. if the data are unbalanced, Welch-r method works well enough in the case of highly heterogeneous positive positive or negative negative pairs, there power are 68.8% and 51% consequencly. Welch method perform well enough only in the case of highly heterogeneous variety of positive positive pairs with it is power of 64.8%. While mixed models method is good in the case of a very heterogeneous variety of negative partner with 54.6% power. So in general, when there is a variance is not homogeneous case, Welch method is applied to the data rank (Welch-r) has a better performance than the other methods..

5. CONCLUSION

Every statistical models which is fail to calculate the existence of messy will cause a group of data tends to lead to a wrong conclusion. That's why unstandarized methods or transformation is needed so we can perform a standardized method. An unstandardized method are not hard to apllied in messy data. Mean comparison testing procedure in mean model (experimental design model's) which can accommodate messy data because violation homogeneity assumption of variance. One of them is Welch test and Mixed models procedure and Welch test ini rank data (Welch-r). Based on those three simulation test, they work very well when heterogeneity of variance is on moderate rank. Welch test in rank data is more robust then violation assumption homogeneity of variance then a normal Welch test dan F mixed model test.

BIBLIOGRAPHY

Cribbie, R.A., Wilcox, R.R, Bewell, C., dan Keselman, H.J., 2007. *Test of Treatment Group Equality When Data are Nonnormal and Heteroscedastic*. Journal of Modern Applied Statistical Methods May 2007, Vol 6 No. I,117-132.

David, M. dan Vikki, M., 2008. *Modern robust statistical methods an easy way to maximize the accuracy and power of your research*. American Psychologist Vol 63, No. 7, 591-601.

Gujarati, D., 1995. *Basic Econometrics Third Edition*. Mcgraw-Hill, Inc. New York.

Hafley, W.L. dan Lewis, J.S., 1963. *Analizing messy data*. Industrial and Engineering Chemistryl Vol 55(4) pp 37-39.

Hand, D.J., 2008. *Statisics, a very short introduction*. Oxford University Press Inc. New York.

Jiang, J. (2007) *Linear and Generalized Linear Mixed Models and Their Applications.* Springer. New York.

Johnson, D.E. dan Milliken, G.A., 2006. *Messy data in Encyclopedia of Statistical Sciences published online: 15 Aug 2006.* John Willey and Son, Inc.

Milliken, G.A.dan Johnson, D.E., 2009. *Analysis of messy data volume 1: designed experiments second edition*. Chapman & Hall/CRC Taylor & Francis Group. Bocca Raton.

Montgomery, D.C., 2001. *Design and Analysis of Experiments 5th Edition.* John Willey & Son Inc. New York.