Measurement of Action Result of Eliminating Bivariate based Cook's D: Cell Concentration and Turbidity

Ahmad Syauqi[1], Hari Santoso[2], Siti Nurul Hasana[3]
[1,2]Biological Department of Mathematics and Natural Sciences Faculty of Islamic University of Malang, Indonesia
[3]Mathematic Education Department of Teacher's Training and Education Faculty of Islamic University of Malang, Indonesia
email: * [1]syauqi.fmipa@unisma.ac.id, [2]harisantoso.m.biomed@gmail.com, [3]nurul.fkipunisma@gmail.com

***Abstract*- The method of turbidimetry interprets the effects of light scattering of the cell particles consisting of the mean of cell concentration. The relation between number of cell particles and turbidity has not been adequately expressed particularly to bivariate statistically. The objective of this paper is to discuss the action result of eliminating a bivariate-based Cook's D analysis, which can be measured with relative error of cell quantification in turbidimetry. Experiments were designed to research the regression design on two paired variables (bivariate), and according to the linear relationship characteristic between suspended solid and turbidity. Variables consisted of cell concentration and turbidity. Paired sample size was ten data as a dataset and replicated five times. The diagnosis of regression on linearity parameter against five sets of data was used as assumption of covariance analysis, *backward method*–residual analysis of Cook's D, and then measured by relative error value of cell concentration result by a turbidity value. The regression diagnostic of plotting fifty bivariate had the heterochedastisity character with five slopes being not homogenous. The diagnosis a linier curve was performed and elimination action toward to a bivariate at the furthest point using *Cook's Distance* value. One dataset gives the lowest relative error result of cell quantification and become consideration in turbidimetry as a standard curve.**

Keywords*— *turbidimetry; bivariate elimination; cook's D

I. INTRODUCTION

Current ethanol production relies on one cell fungus of *Saccharomyces cerevisiae* and requires highly efficient and effective process in order to be mass-produced. p The process of fermentation, having knowledge of the number of cells is one of the attempts at achieving it. The method of turbidimetry is appropriate for that purpose which technically interprets the effect of light scattering from cell particles. The relationship between scattered light and numbered particles has been found to be linear. Nonetheless, analysis on discovering the appropriate regression equation remains shallow. The scattered spectrum is attenuation (T) because it is blocked by suspended particles (TSS). Mishendo, Hovenier and Travis discovered the theoretical formulation which explains that T and TSS have a mathematically-defined relationship for homogenous particles [1].

$$TSS = \frac{2}{3}\left(\sum_{i=1}^{n}\frac{Q_{exti}p_{mi}}{\rho_{pi}d_{pi}}\right)^{-1} T = a \times T$$

Dimana: T = Light intensity reduction
a = A tribe that shows proportional

Light intensity reduction is technically a form of light scattering, measured by Nephelometric turbidity unit (NTU). The relationship between the number of cell particles and turbidity has been done statistically to data of bivariate. Mathematics is a form of science that deals with logic of shape, quantity and arrangement. The mathematical equation above shows a shape of linearity between TSS and T.

Although most research about regression diagnostic used the approach of deleting a single observation [2], Cook's D analysis managed to surface the discrepancy i.e. unusual Y value that influence the slope and intercept of a simple regression. The objective of this paper is to discuss the action result eliminating a bivariate based on Cook's D analysis, which can be measured with relative error of cell quantification in turbidimetry.

II. METHODS

Research was conducted through experiments applying the regression design on two paired variables (bivariate), and the linear relationship character [1]. The first variable of the experiment was suspended solid of cells, which is the known population of *S cerevisiae.* The other variable adopted was turbidity. Paired sample size was 10 data as a data set and replicated five times [3]. The diagnosis of regression on linearity parameter against five sets of data adopted the assumption of covariance analysis (A*NACOVA*), *Backward method* [2][4][5] – residual analysis of Cook's distance (Cook's D) and *Centered Leverage* [6].

The analysis assumed that the regression coefficient was not equal to zero, also termed as Covariance. Regression coefficients in respective dataset were tested with H_o hypotheses being homogenous or the five slopes of the line being parallel. An alternative hypothesis of H_1 was not homogenous and the slopes not parallel. Homogeneous is

rejected if Fs value > F table at minimum of confidence 95% (P=0.05). The result of test using covariance assumption leads forming five subsets of data respective ten dots is merged or not. The implementation of hypotheses test using the Fs value compared the critical value of F distribution and formulation was as follows;

$$Fs = \frac{\Sigma y^2 \text{ (Corrected by average regression in the group)} - \Sigma y^2 \text{ (Corrected by respective regression group)}}{[\Sigma y^2 \text{ (Corrected in respective group)}/N - 2k]} \quad (1)$$

where:

N : data population
k : group (replication)
df *Enumerator* = (k – 1)
df *Denumerator* = (N – 2k)

The experiment began with the independent variable of cell particles of *Saccharomyces cerevisiae* being suspended in water (cell/Cm^3). Following that, the Linearity relationship of particles and light were tested, with the turbidity of suspended cells as the dependent variable, and shown by determination value of regression of both variables. Normality of paired data was conducted to compare the results of before and after the test process of Cook's D and centered Leverage; with the possibility of eliminating the deviant paired data. Data elimination was performed by Cook's D value prior to be measured by relative error value.

Cell as particles were performed as follows: phosphate saline buffer was made with 125 Cm^3 of disodium hydrogen phosphate added with 125 Cm^3 Sodium di-hydrogen phosphate at 0,05M respectively, added HCl 10% to pH 4,9; Sodium Chloride at 0,095% and ethanol absolute added to be 10% (w/v). This formula was referred as BP-4.9ES [7][8]. Furthermore, granules of *Saccharomyces cerevisiae* were added with BP-4.9ES to be suspended and shaken six times with the hand, prior to being incubated for 30 minutes. Suspended solid was centrifuged at 8000g and contain two stages of 2 x 10 minutes.

Quantity of cells (X) calculated on haemocytometer at low concentration using grid of 1 mm^2 area on that in 4x4 boxes and thickness of 0.1 mm. High concentration of cell using 1 mm^2 were placed in the centre of 5x5 boxes. Calculating the quantity of cells was conducted at average volume of 1 Cm^3 or mL by conversion value. Turbidity value of suspension obtained from respective value of cell quantity in water. The bottle of turbidimeter was filled with 15 Cm^3 of water and then added cell suspension with reduced aggregate.

IV. Result and Discussions

The first analysis was based on the replication of research variable i.e. the five replications as five data subsets containing ten paired data and then merged. The second analysis was assumed that replication as dataset was reproducible [9][10] and result of Fs calculation using (1) was shown an application output of computing and gives heterogeneous slopes.

```
          Corrected Sum Square Each Group
-------------------------------------------------------------
   Group        df     SS        Slope  Correction  df
-------------------------------------------------------------
 Data Set 1      7  27166.578   2.086    971.184    6
 Data Set 2      8 132634.734   2.419   5347.422    7
 Data Set 3      7  35198.633   1.895   3694.992    6
 Data Set 4      7  30792.055   1.324    687.359    6
 Data Set 5      7  30948.875   1.465     39.145    6
-------------------------------------------------------------
   Total        36 256740.875           10740.102  31
-------------------------------------------------------------
              Fs Value of 1-3:
          Fs1 =  324.103 - (Table:    4.120)
          Fs2 =   10.307 - (Table:    2.520)
          Fs3 =    0.324 - (Table:    2.634)

Result Assumption Analysis of Covariance
     - Linier Regression
       F1: Fulfilling Requisite of Covariance Analysis (0.050)

     - Homogeneous slope
       F2: No Fulfilling Requisite of Covariance Analysis (0.050)

     - Variable of X Effect toward group
       F3: Fulfilling Requisite of Covariance Analysis (0.050)

App. Computation (update 2016) by:
          Ir. Ahmad Syauqi, M.Si.
          Ref.: Sudjana (1985)

Dos Box App.-Megabuild6-Printfil 5.22 Personal
Edition
```

Let knowing why the result is heterochedastisity. The regression equation is

$$Y_i = a + bX_i + e_i$$

with e referring to error and If $\Sigma e = \Sigma (Y_i - \bar{Y})$ than $[\Sigma (Y_i - \bar{Y})^2 / n\text{-}1]$ is variance. The heterochedastisity character in graph showed that at higher X value; the error of Y value becomes larger. In this case, at higher value of cell particles concentration; the scattered light is not consistent if it is compared with lower cell concentrations. These relationship characteristics require the selection of one among the five slopes (many slopes).

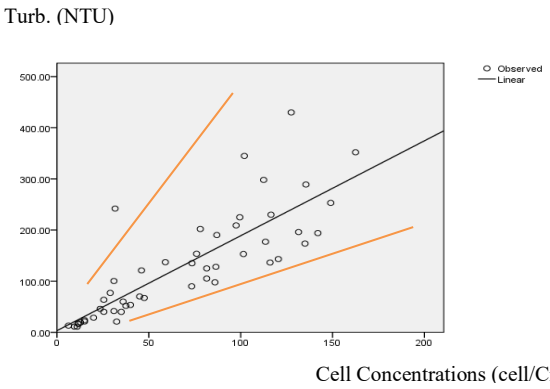


Fig. 1. Data Plot of 5 x 10 bivariate of Cell Concentration and Turbidity show the Heterochedastisity Character.

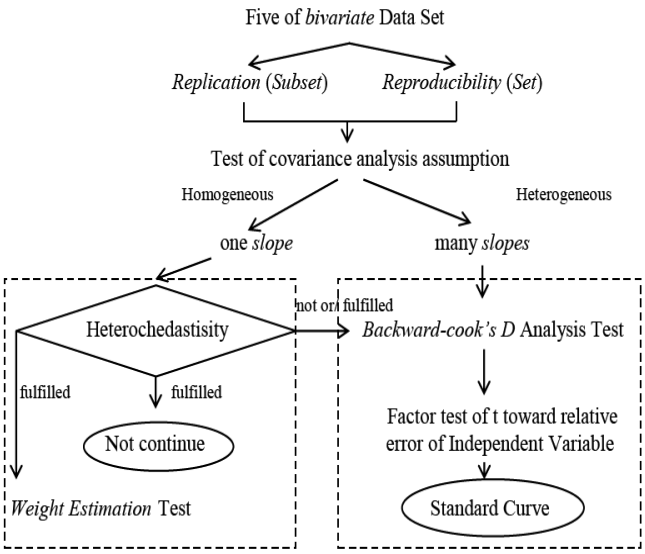


Fig. 2. Scheme of Analysis Process and Diagnose of Bivariate Dataset using Covariance Assumption Test and Cook's D [11][4][10]; without Test of *Weighted Least Square:* WLS from Reference [5].

Plotting bivariate illustrates the heterochedastisity character (Fig. 1) and processes of analysis in a regression or diagnostic as shown on scheme (Fig. 2). Why had the analysis not use test of weight estimation? The weakness of weight estimation with the test of *Weighted Least Square:* WLS [5] for this research was due to the predictor being unreal if it was estimated by turbidity value as data transforms. This technique was done by reference [12], also by the proportional reverse of the response variable [13]. Therefore the researcher considered it impractical to apply the WLS technique in respect of fermentation production. Besides that, the weakness of the Cook's D analysis makes it impractical to be applied in large scaled population paired data [2].

Calculation of error was done applying the formula in reference [5][11] and standard deviation of cell number estimate was translated as the form

$$s_X = s_Y/b.\ \sqrt{1 + 1/N + \frac{(NTU_{sample} - \overline{NTU}_{standard})^2}{b^2\,\Sigma\,(X^2) - [(\Sigma^2\,X)/N]}}$$

where

s_X = Standard deviation of Cell Concentration
s_Y = Standard deviation of Turbidity
NTU_{sample} = A value of turbidity
$\overline{NTU}_{standard}$ = Average of NTU dataset
X = Cell particles
N = Number of point (bivariate)

TABLE I
Parameters of Regression Equation by Value of *Cook's Distance* Diagnosis - Elimination, Simulation of Turbidity (Y) Value and Percentage Relative Error Value of Cell Concentration (X) used t distribution at α=0.95 one tail.

Repli-cation	*n* Data	Original Data a**)	Original Data *Slope*	Pasca elimination[1*)] *New Slope*	Expected Value NTU (Y)	Normality of Residual Standardized Data[2*)]	Relative Error (%) Predicted X Value[3*)]
1	10	-26,210	3,170				36,41
	9			2,727	200	Normal	36,61
	8			2,088			15.65④
2	10	12,220	2,213			Right skewed	69,08
	9			2,423	200	Normal	25,89④
	8			2,697		Left Skewed	24.5
3	10	-7,537	1,805			Normal	35,27
	9			2,114	200	Left Skewed	24,65
	8			1,902		Normal	28,10④
4	10	-17,940	1,547			Normal	21,39
	9			1,391	200	Right Skewed	17,8
	8			1,321		Normal	12,75④
5	10	-3,439	1,625		200	Normal	20,41
	9			1,448		Left Skewed	9,14
	8			1,470		Normal	3,14④

Information: 1 and 2*) Application of SPSS v.16 Windows 10 Operating system; **) Intersept
3*) *Under Dos application – DosBox Megabuild6 – Printfil 5.22 Personal Edition* on Windows 10 OS.

Statistical description at replication two (Table 1) shows the slope values indicated a change. The certainty of the regression relationship is shown in Table 2 and 3. Fig. 3 depicts the normality of residual standardized bivariate.

TABLE II
Analysis of Variance (ANOVA) of Bivariate Dataset in Replication Two

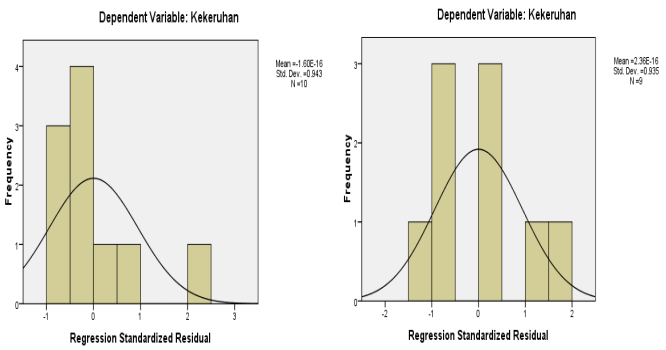
Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	109828.254	1	109828.254	25.390	.001[a]
	Residual	34604.798	8	4325.600		
	Total	144433.052	9			

a. Predictors: (Constant), Jumlah_Partikel_Sel
b. Dependent Variable: Kekeruhan

TABLE III
Regression Equation Coefficients of Bivariate Dataset in the Replication Two

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	12.220	32.637		.374	.718
	Cell particle	2.213	.439	.872	5.039	.001

a. Dependent Variable: Kekeruhan



A B

Fig. 3. Statistical Description on Indicating a Change of Residual Standardized Bivariate Normality: A before and B after taking the action of Eliminating Bivariate based Cook's D Analysis

Percentage (%) of relative error is:

$$\pm (t.\ s_X) / [(Trb.Val - a)/b].100$$

Where:

s_X = Standard deviation
Trb.Val = Value of Turbidity
a = Intersept of curve
b = Parameter of regression or *slope*
t = value at $t_{0.05(2),n-2}$

The example of action result eliminating one bivariate using an application and implementation value of 200 NTU used statistic of t distribution at α=0.95 is

Replication four with elimination of two bivariate

```
1. Linier equation  ---->   Y =  1.321 X + (-9.048)
         Correlation value  (r)  =  0.988
2. Value: TURBIDITY VARIANCE(Vy)  = 121.228515625
         CONC. VARIANCE    (Vx)  =  10.378
3. Error of mean Y at confidence of 95%:(plus minus)
   = 61.977
4. Result of Conc. X from Turbidity value = 200 NTU
   ---> 158.2202606201172   (plus minus)  25.396
5. Conclusion:

     RELATIVE ERROR OF CONC. > 10% YAITU  16.05%

Program oleh Syauqi
File : d:\doto\ul4rev2asli.RGS
- Dicetak dengan Aplikasi Dos BoxMegabuild6-Printfil
  5.2 Personal Edition
```

Replication five with elimination of one bivariate

```
1. Linier equation  ---->   Y =  1.448 X + ( 1.926)
         Correlation value  (r)  =  0.993
2. Value: TURBIDITY VARIANCE(Vy)  = 59.258
         CONC. VARIANCE    (Vx)  =  6.596
3. Error of mean Y at confidence of 95%:(plus minus)
   = 52.436
4. Result of Conc. X from Turbidity value = 200 NTU
   ---> 136.81396484375   (plus minus)  15.600
5. Conclusion:

     RELATIVE ERROR OF CONC. > 10% YAITU  11.40%

Program oleh Syauqi
File : d:\doto\ul5rev1asli.RGS
- Dicetak dengan Aplikasi Dos BoxMegabuild6-Printfil
  5.2 Personal Edition
```

Replication five with elimination of two bivariate

```
1. Linier equation  ---->   Y =  1.470 X + ( 2.955)
         Correlation value  (r)  =  0.999
2. Value: TURBIDITY VARIANCE(Vy)  =  6.668
         CONC. VARIANCE    (Vx)  =  2.189
3. Error of mean Y at confidence of 95%:(plus minus)
   = 62.135
4. Result of Conc. X from Turbidity value = 200 NTU
   ---> 134.0642395019531   (plus minus)   5.356
5. Conclusion:

     RELATIVE ERROR OF CONC. < 10% YAITU  3.99%

Program oleh Syauqi
File : d:\doto\ul5rev2asli.RGS
- Dicetak dengan Aplikasi Dos BoxMegabuild6-Printfil
  5.2 Personal Edition
```

Reference [14] using 5% - 8% of standard deviation for measuring precision of height estimate of tree in forest by biological factors from regression curve.

V. CONCLUSION

Eliminating bivariate could be done more than once. Linear regression parameters changes as a result of elimination and requires consideration of the determination standard curve on turbidimetry. The lowest value of relative error result could be used for such consideration. The dataset that uses Cook's D diagnosis with elimination of bivariate twice gives standard curve and at the respected value of 200 NTU, it has relative error value of ±3.99%.

ACKNOWLEDGMENT

This research was funded by the Ministry of Research, Technology and Higher Education in year 2018; focusing at renewable energy in general biotechnological science based letter No.: 0045/E3/LL/2018 date January 16th, 2018; in Islamic University of Malang (Unisma) No.: 029/F.05/U.I/LPPM/2018 date January 25th, 2018.

REFERENCES

[1] A. Hannouche, C Ghassan, G. Ruban, B. Tassin, B. Lemaire, "Relationship between turbidity and total suspended solids concentration within a combined sewer system." *Water Science and Technology* 64(12), pp. 2445-52 <10.2166/wst.2011.779>, 2011.

[2] K.E.Muller, M.C. Mok, "The distribution of Cook's D statistic." *Communication Stat Theory Methods* 26(3), pp. 525-546, Doi:10.1080/ 03610927708831932, 2011.

[3] Sugiyono, *Statistik Nonparametrik untuk Penelitian*. Alfabeta, Bandung: 2012, p. 13.

[4] Stang, *Aplikasi Statistik Multivariat dalam Penelitian Kesehatan*. Jakarta: Mitra Wacana Media, 2017.

[5] R.K. Sembiring, *Analisis Regresi,* Edisi kedua, Bandung: Penerbit ITB, 2003, p. 195.

[6] C. Trihendradi, *Kupas Tuntas Analisis Regresi Strategi Jitu Melakukan Analisis Hubungan Causal*. Yogyakarta:Andi, 2007, p. 25.

[7] A. Syauqi, H. Santoso dan S.N. Hasana. "Dispersi agregat sel *Saccharomyces cerevisiae* tersuspensi dalam air dengan pengaruh kimia-fisik." in *Prosiding Seminar Nasional Biologi 2018 Biodiversitas: Pembelajaran, Penelitian, dan Penerapannya dalam Pengelolaan Lingkungan.* Bandung : Pusat Penelitian dan Penerbitan UIN SGD. 2018, p 41-47.

[8] A. Syauqi, H. Santoso dan S.N. Hasana. "Dispersi dan Prosedur Dispersi Agregat Sel *Saccharomyces cerevisiae.*" unpublished.

[9] A. Syauqi. "Standard Curve of Biuret-Spectrofotometry Using Amonia for Protein Molecule." unpublished.

[10] A. Syauqi, "Comparative method of references and protein quantifications using Biuret-spectrophotometric." *Chimica et Natura Acta (Jcena),* inpress

[11] M.F. Qudratullah. *Analisis Regresi Terapan Teori, Contoh Kasus dan Aplikasi dengan SPSS.* Yogyakarta: Andi. 2013, p 3-4.

[12] Y.D. Setyningsih, dan Noeryanti. "Penggunaan Metode Weight Least Square untuk Mengatasi Masalah Heterokedastisitas dalam Analisis Regresi (Studi Kasus pada Balita Gizi Buruk Tahun 2014 di Provinsi Jawa Tengah)." *Jurnal Statistika Industri ddan Komputasi* 2(1), pp. 51-58, 2017.

[13] N. Hanifah, N. Heryanto dan F. Agustina. "Penerapan Model Weighted Lest Square untuk Mengatasi Heterokedastisitas pada Analisis Regesi Linier." *J. EurekaMatika* 3(1), pp. 105-114, 2015.

[14] A. van Laar and A. Akca. *Forest Mensuration.* Dordrecht: e-book of Springer. p. 124, 2007.