International Journal of Science and Research (IJSR)
ISSN: 2319-7064
SJIF (2022): 7.942

The Effect of Data Grouping on the Significance of the Coefficients of the Multiple Linear Regression Model

Adil Mousa Younis Waniss[1], Abdulmajied Ali Balkash[2]

[1]Associate Professor, Department of Management Information Systems, College of Administration and Economics, Qassim University, *Buraidah*, *P.O. Box 6633*, KSA
Email: *younisadil2002[at]gmail.com*

[2]Professor, Department of Management Information Systems, College of Administration and Economics, Qassim University, *Buraidah*, *P.O. Box 6633*, KSA
Email: *d.balkash[at]gmail.com*

Abstract: ***This study is devoted for investigating the effect of grouping on the multiple linear regression coefficient's particularly on their significance, estimated confidence intervals and their standard error of the multiple linear regression coefficients, we generate random numbers from Excel and built multiple linear regression model with two independent variables, we compared regression coefficient's for an ungrouped multiple linear regression model with different five aggregated regression coefficient's, we find that, multiple linear regression coefficient's and their confidence intervals as well as the standard error of the multiple linear regression coefficients are unbiased for the different levels of grouping and except in case of grouping by less than five grouped a Sharpe decreases happened to their values. Grouping gives the same results for the regression coefficient for different groups, and for the different estimated confidence intervals as well as for standard error of regression coefficients. We find that grouping should not goes above twenty groups nor below five groups. The most important flinging is that, all regression coefficients for different groups are unbiased estimate and gives same p-values for different groups and different groups gives all most the same estimated confidence intervals and standard errors for the deferent multiple regression coefficients.***

Keywords: data aggregation; regression coefficient's; confidence intervals; standard error of regression coefficients

1. Introduction

Now a days the flow of data or information is a common phenomenon throw out the world, everyone is familiar with data mining and big data analysis which led to the problem of how to deal with such huge information, one of the technique that is used by econometricians to reduce the effect huge data is the technique of data aggregation or data grouping, but as we known the aggregation of data generate some problem such as the problem of loss of information or over estimation of some statistical parameters.

The aggregation problem is a common problem in data analysis in nearly all fields of study, including social science and some fields of physics. In its broadest definition, the aggregation problem is the loss of information that occurs when aggregate, or large-scale, data is replaced by individual, or small-scale, data. Data aggregation is the process of collecting data to present it in summary form. This information is then used to conduct statistical analysis and can also help company executives make more informed decisions about marketing strategies, price settings, and structuring operations, among other things (Lukas Racickas 2023). The use of random estimates in regression models has been gaining more attention in recent years (4, 5, 6).

New ideas on combining different procedures for estimation, coding, forecasting and learning have recently been considered in statistics and several related fields, leading to a number of very interesting results (Y Yang Bernoulli, 2004 projecteuclid.org).

It is well known that using aggregate data might result in correlation coefficients that are significantly biased above their individual values [Adil M. Youniss2002], has demonstrated that the regression coefficients could similarly be influenced. It is commonly known that one should never assume that relationships that exist at one level of analysis would necessarily be equally strong at another level.

The purpose of this paper is to demonstrate the effect of aggregation on the significance of the regression coefficients of the multiple linear regression model when the data is ungrouped against grouped data. We will show on the following paragraphs different types of grouped data with (15, 10, 8, 6, 5, 4) class intervals, along with their test of hypothesis and confidence intervals for each class interval and compare them with their cross ponding ungrouped measures with discussion of results. Throw out these paragraphs a detailed explanation of how aggregation is done by Excel and what types of functions we used.

2. Multiple Linear Regression for Un-grouped data

The multiple linear regression model is

$$y = \beta_o + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_{k-1} x_{k-1} + \varepsilon \qquad (1)$$

Where y is the dependent variable, are the independent x_i variables, $\beta_o, \beta_1, \beta_2, , \cdots, \beta_{k-1}$ to minimize the sum of the squares of the residuals we use the least square method to

estimate the regression coefficients of the model Where, K - 1 is the number of the independent variables, and K is the number of the regression coefficients (kor & altun, 2020), (Ali, P. A. & Younis,. A. A., 2021).

2.1 The parameters of the model. The parameters of the model are estimated by the least square method as follows.

$$B = A^{-1} \cdot C \tag{2}$$

The matrixes are defined as follows:

$$A = \begin{bmatrix} n & \sum x_{1i} & \sum x_{2i} \\ \sum x_{1i} & \sum x_{1i}^2 & \sum x_{1i} \cdot x_{2i} \\ \sum x_{2i} & \sum x_{1i} \cdot x_{2i} & \sum x_{2i}^2 \end{bmatrix}, \quad B = \begin{bmatrix} b_0 \\ b_1 \\ b_2 \end{bmatrix}$$

$$C = \begin{bmatrix} \sum y_i \\ \sum x_{1i} \cdot y_i \\ \sum x_{2i} \cdot y_i \end{bmatrix} \tag{3}$$

2.2 Standard Error of the Multiple Linear Regression Coefficients. The Standard error of the multiple linear regression coefficients will be calculated from the variance covariance matrix as follows:

$$\sigma^2 = \frac{SSE}{n-k} \tag{4}$$

Where, $\frac{SSE}{n-k}$ is the estimated multiple linear regression model variance. We get from the matrix diagonal

$$S^2(b_0), S^2(b_1), S^2(b_2), \cdots, S^2(b_{k-1}) \tag{5}$$

Which are the variances of errors of the regression coefficients and.

$$S(b_0), S(b_1), S(b_2), \cdots, S(b_{k-1}) \tag{6}$$

Which are the standard errors of the regression coefficients (Frost, J., 2023).

2.3 The Test of Significance of the Multiple Linear Regression Coefficients. To test the significance of the multiple linear regression model (Ali, P. A. & Younis. A. A, 2021), we follow the usual test of hypothesis steps as shown in the following table:

Table 1: Test steps multiple linear regression model table

H_o	H_1	Test statistic	Reject H_o If
$\beta_i \le \beta_{io}$	$\beta_i > \beta_{io}$	$T = \frac{b_i - \beta_{io}}{S(b_i)}$	$T > T(\alpha, n-k)$
$\beta_i \ge \beta_{io}$	$\beta_i < \beta_{io}$		$T < -T(\alpha, n-k)$
$\beta_i = \beta_{io}$	$\beta_i \ne \beta_{io}$		$T < -T(\alpha/2, n-k)$ OR $T > T(\alpha/2, n-k)$

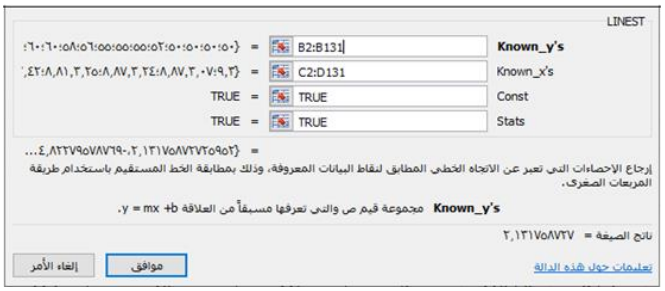
2.4 The Confidence Intervals of the Multiple Regression Coefficients. The confidence intervals of the multiple regression coefficients are calculated as follows

$$b_i \pm T(\alpha/2, n-k) S(b_i) \tag{7}$$

3. The Data, Application and Results

We use Excel (2016) to generate 130 random numbers and we defined them as is the demand for a certain y commodity x_1, x_2 are the price of the commodity and monthly income rate respectively. We calculate the multiple regression model for ungrouped data from Excel (2016), using both of "Linest" and "Regression Statistics" (Greg Harvey, Microsoft Excel 2016).

3.1 Test of hypotheses and confidence interval for regression coefficients for ungrouped multiple linear regression model. We calculated the measurements of the ungrouped multiple linear regression, by using "Linest" and the function "Regression Statistics" in Excel as follows:



For the ungrouped multiple linear regression model, the estimated regression model and the test of hypotheses of the estimated coefficients and their confidence intervals are as the followings:

$$\hat{y} = 90.95 - 4.82x_1 + 2.13x_2$$

Table 2: Measurements of Ungrouped Multiple Linear Regression Model

	β_i	St. Err	-Stat.	P-value	Low 95%	Up 95%
Intercept	90.95	12.29	7.39	1.66E-11	66.62	115.
Price	-4.82	1.16	-4.15	5.98E-05	-7.12	-2.52
Income	2.13	0.60	3.54	0.0006	0.94	3.32

It is very clear from table (2), that all the coefficients are significant and the confidence intervals does not contain zero number.

3.2 Test of hypotheses and confidence interval for regression coefficients for grouped multiple linear regression model. To calculate the above 4 equations and test the hypothesis of the regression coefficients for grouped data we aggregate the data according to the following steps:

3.2.1 Data is aggregated for different class intervals, 15, 10, 8, 6, 5 and 4 class intervals with upper and lower boundary and frequency for each class interval.

3.2.2 Each class interval frequencies will be represented by mean value with their cross-ponding frequencies, instead of mid-point as usual.

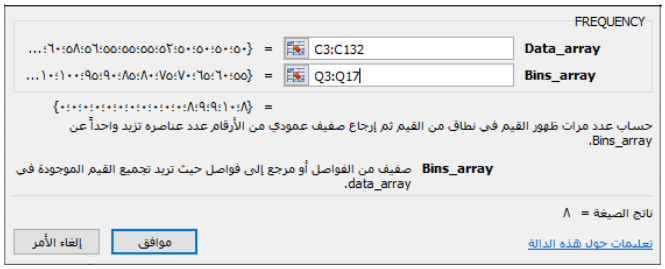
3.2.3 For estimating the regression coefficients, testing their significance and constructing confidence interval for them, we use the same procedures as for ungrouped data but we take in account the means and frequencies together.

3.2.4 The grouping of the data is don only for the dependent variable y while the grouping of the independent variables

x_1, x_2 is done automatically according to the rule of aggregation.

Calculation of Grouped data from Excel:

3.3 Calculation of Grouped data from Excel. We calculated the measurements of the grouped multiple linear regression, by using "FREQUENCY" in Excel, the calculation for 15, 10, 8, 6, 5 and 4 class intervals groups is done as follows:



3.4 The parameters of the grouped model are estimated by the least square method for 15 class interval with their means and frequencies as follows:

$$B=\begin{bmatrix} b_0 \\ b_1 \\ b_2 \end{bmatrix} \quad , \quad C=\begin{bmatrix} \sum_{i=1}^{15} \bar{y}_i \cdot f_i \\ \sum_{i=1}^{15} \bar{x}_{1i}\bar{y}_i \cdot f_i \\ \sum_{i=1}^{15} \bar{x}_{2i} \cdot \bar{y}_i \cdot f_i \end{bmatrix} = \begin{bmatrix} 10949 \\ 56615 \\ 114619 \end{bmatrix}$$

$$A=\begin{bmatrix} \sum_{i=1}^{15} f_i & \sum_{i=1}^{15} \bar{x}_{1i} \cdot f_i & \sum_{i=1}^{15} \bar{x}_{2i} \cdot f_i \\ \sum_{i=1}^{15} \bar{x}_{1i} \cdot f_i & \sum_{i=1}^{15} \bar{x}_{1i}^2 \cdot f_i & \sum_{i=1}^{15} \bar{x}_{1i} \cdot \bar{x}_{2i} \cdot f_i \\ \sum_{i=1}^{15} \bar{x}_{2i} \cdot f_i & \sum_{i=1}^{15} \bar{x}_{1i} \cdot \bar{x}_{2i} \cdot f_i & \sum_{i=1}^{15} \bar{x}_{2i}^2 \cdot f_i \end{bmatrix} =$$

$$=\begin{bmatrix} 130 & 732 & 1246 \\ 732 & 4685 & 5931 \\ 1246 & 5931 & 14032 \end{bmatrix}$$

$$A^{-1}=\begin{bmatrix} 25.42 & -2.4 & -1.24 \\ -2.4 & 0.23 & 0.12 \\ -1.24 & 0.12 & 0.06 \end{bmatrix}$$

$$B=A^{-1}\times C=\begin{bmatrix} 25.42 & -2.4 & -1.24 \\ -2.4 & 0.23 & 0.12 \\ -1.24 & 0.12 & 0.06 \end{bmatrix} \times \begin{bmatrix} 10949 \\ 56615 \\ 114619 \end{bmatrix}$$

$$=\begin{bmatrix} 94.517 \\ -5.158 \\ 1.963 \end{bmatrix} = \begin{bmatrix} b_0 \\ b_1 \\ b_2 \end{bmatrix}$$

The estimated regression model and the test of hypotheses of the estimated coefficients and their confidence intervals are as the followings:

$$\hat{y}=94.525-5.16x_1+1.96x_2$$

Table 3: Measurements of grouped Multiple Linear Regression Model with 15 class intervals

	β_i	St. Err	t-Stat.	P-value	Lo 95%	Up 95%
Intercept	94.52	12.06	7.84	1.59E-12	70.66	118.37
MEANX1	-5.17	1.14	-4.54	1.29E-05	-7.42	-2.92
MEANX2	1.96	0.59	3.33	0.00115	0.79	3.13

It is very clear from table (3), that all the coefficients are significant and the confidence intervals does not contain zero number.

3.5 The parameters of the grouped model are estimated by the least square method for 4 class interval with their means and frequencies as follows. The calculation for 4 class intervals grouping is done by Excel as follows:

$$A=\begin{bmatrix} \sum_{i=1}^{4} f_i & \sum_{i=1}^{4} \bar{x}_{1i} \cdot f_i & \sum_{i=1}^{4} \bar{x}_{2i} \cdot f_i \\ \sum_{i=1}^{4} \bar{x}_{1i} \cdot f_i & \sum_{i=1}^{4} \bar{x}_{1i}^2 \cdot f_i & \sum_{i=1}^{4} \bar{x}_{1i} \cdot \bar{x}_{2i} \cdot f_i \\ \sum_{i=1}^{4} \bar{x}_{2i} \cdot f_i & \sum_{i=1}^{4} \bar{x}_{1i} \cdot \bar{x}_{2i} \cdot f_i & \sum_{i=1}^{4} \bar{x}_{2i}^2 \cdot f_i \end{bmatrix} =$$

$$=\begin{bmatrix} 130 & 731.99 & 1245.63 \\ 731.99 & 4642.231 & 6012.14 \\ 1245.63 & 6012.14 & 13872.02 \end{bmatrix} \Rightarrow$$

$$A^{-1}=\begin{bmatrix} 43.14 & -4.07 & -2.11 \\ -7.07 & 0.38 & 0.19 \\ -2.11 & 0.19 & 0.10 \end{bmatrix}$$

$$B=\begin{bmatrix} b_0 \\ b_1 \\ b_2 \end{bmatrix}, \quad C=\begin{bmatrix} \sum_{i=1}^{4} y_i \cdot f_i \\ \sum_{i=1}^{4} \bar{x}_{1i} y_i \cdot f_i \\ \sum_{i=1}^{4} \bar{x}_{2i} \cdot y_i \cdot f_i \end{bmatrix} = \begin{bmatrix} 10949 \\ 57001.35 \\ 113879.99 \end{bmatrix}$$

$$B=A^{-1}\times C=\begin{bmatrix} 43.14 & -4.07 & -2.11 \\ -7.07 & 0.38 & 0.19 \\ -2.11 & 0.19 & 0.10 \end{bmatrix} \times$$

$$\begin{bmatrix} 10949 \\ 57001.35 \\ 113879.99 \end{bmatrix} = \begin{bmatrix} 82.08 \\ -3.99 \\ 2.57 \end{bmatrix} = \begin{bmatrix} b_0 \\ b_1 \\ b_2 \end{bmatrix}$$

The estimated regression model and the test of hypotheses of the estimated coefficients and their confidence intervals are as the followings:

$$\hat{y}=82.08-3.996x_1+2.57x_2$$

Table 4: Measurements of grouped Multiple Linear Regression Model with 4 class intervals

	β_i	St. Er	t-Stat	P-value	Lo 95%	Up 95%
Intercept	82.08	13.71	5.99	2.03E-08	54.96	109.21
MEANX1	-3.99	1.29	-3.089	0.00251	-6.55	-1.43
MEANX2	2.57	0.67	3.83	0.000202	1.24	3.89

It is very clear from table (4), that all the coefficients are significant and the confidence intervals does not contain zero number. Similarly, we follow the same procedures for the class, 10, 8 and 6.

3.6 The parameters of the grouped model are estimated by the least square method for 10 class interval with their means and frequencies as follows. The estimated regression model and the test of hypotheses of the estimated coefficients and their confidence intervals are as the followings:

$$\hat{y} = 90.61 - 4.81x_1 + 2.16x_2$$

Table 5: Measurements of grouped Multiple Linear Regression Model with 10 class intervals

	β_i	St. Er	t-Stat	P-value	Lo 95%	Up 95%
Intercept	90.61	11.61	7.81	1.95E-12	67.62	113.59
MEANX1	-4.81	1.09	-4.38	2.44E-05	-6.98	-2.64
MEANX2	2.16	0.57	3.79	0.000223	1.03	3.28

It is very clear from table (5), that all the coefficients are significant and the confidence intervals does not contain zero number.

3.7 The parameters of the grouped model are estimated by the least square method for 8 class interval with their means and frequencies as follows. The estimated regression model and the test of hypotheses of the estimated coefficients and their confidence intervals are as the followings:

$$\hat{y} = 91.92 - 4.931x_1 + 2.09x_2$$

Table 6: Measurements of grouped Multiple Linear Regression Model with 8 class intervals

	β_i	St. Error	t-Stat	P-value	Lo 95%	Up 95%
Intercept	91.92	11.68	7.87	1.38E-12	68.79	115.04
MEANX1	-4.93	1.10	-4.47	1.72E-05	-7.11	-2.75
MEANX2	2.09	0.57	3.66	0.00037	0.96	3.23

It is very clear from table (6), that all the coefficients are significant and the confidence intervals does not contain zero number.

3.8 The parameters of the grouped model are estimated by the least square method for 6 class interval with their means and frequencies as follows. The estimated regression model and the test of hypotheses of the estimated coefficients and their confidence intervals are as the followings parameters of the grouped model are estimated by the least square method for 4 class interval with their means and frequencies as follows. The calculation for 4 class intervals grouping is done by Excel as follows:

$$\therefore \hat{y} = 96.69 - 5.3831x_1 + 1.86x_2$$

Table 7: Measurements of grouped Multiple Linear Regression Model with 6 class intervals

	β_i	St. Err	t-Stat.	P-valu	Lo 95%	Up 95%
Intercept	96.69	13.75	7.03	1.12E-1	69.48	123.9
MEANX1	-5.38	1.29	-4.15	6.13E-0	-7.94	-2.81
MEANX2	1.86	0.67	2.76	0.00671	0.52	3.19

It is very clear from table (7), that all the coefficients are significant and the confidence intervals does not contain zero number.

3.9 The parameters of the grouped model are estimated by the least square method for 5 class interval with their means and frequencies as follows. The estimated regression model and the test of hypotheses of the estimated coefficients and their confidence intervals are as the followings:

$$\hat{y} = 99.76 - 5.68x_1 + 1.72x_2$$

Table 8: Measurements of grouped Multiple Linear Regression Model with 5 class intervals

	β_i	St. Er	t-Stat.	P-value	Lo 95%	Up 95%
Intercept	99.76	11.78	8.47	5.15E-14	76.46	123.06
MEANX1	-5.68	1.11	-5.09	1.21E-06	-7.88	-3.47
MEANX2	1.72	0.58	2.98	0.003423	0.58	2.85

It is very clear from table (8), that all the coefficients are significant and the confidence intervals does not contain zero number.

4. The effect of Grouping on the Regression coefficients of the Model

To discuss the effect of grouping on the regression coefficients of the multiple linear regression, we summarize the values of the multiple linear regression coefficients from the above tables in the following three tables:

Table 9: Estimated β_0 for ungrouped and different grouped models:

Grouping	Ungrouped	15	10	8	6	5	4
β_0	90.95	94.52	90.61	91.92	96.69	99.76	82.08
St. Error	13.70618	12.06	11.61	11.68	13.75	11.78	13.71
P-value	1.66E-11	1.59E-12	1.95E-12	1.95E-12	1.12E-10	5.15E-14	2.03E-08

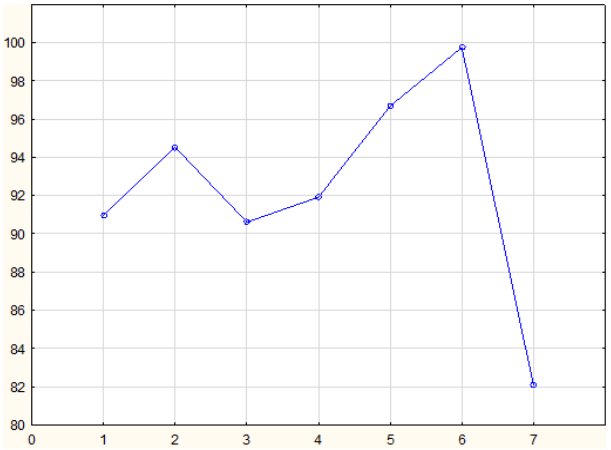


Figure 1: Estimated β_0 for ungrouped and different grouped models:

4.1. Discussion of the effect of Grouping on the Estimated β_0. From table (9) and figure (1) , we noticed that, all the values of the intercepts (estimated β_0) of the different grouped lies within the range of the confidence intervals of the ungrouped model (66.62, 115.28) which shows Significant confidence intervals, that means without any doubt, aggregation or grouping does not affect the actual ungrouped value of the intercept

Table 10: Estimated β_1 for ungrouped and different cropped models:

Grouping	Ungrouped	15	10	8	6	5	4
β_1	-4.8228	-5.17	-4.81	-4.93	-5.38	-5.68	-3.99
St. Error	1.161283	1.14	1.09	1.10	1.29	1.11	1.29
P-value	5.9E-05	1.2E-05	2.4E-05	1.7E-05	6.1E-05	1.2E-06	0.003

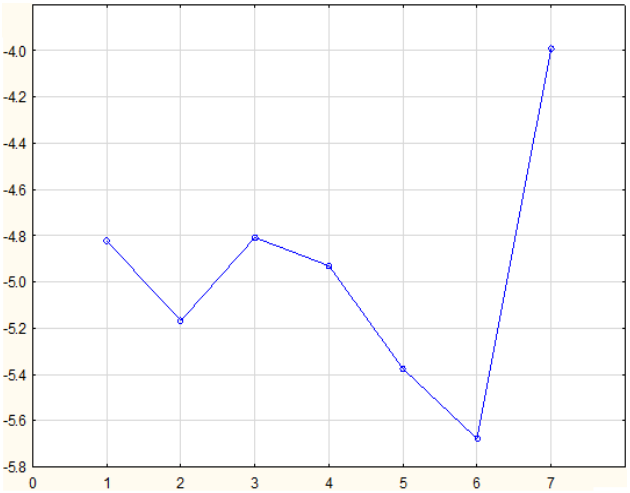


Figure 2: Estimated β_1 for ungrouped and different grouped models:

4.2. From table (10) and figure (2), we noticed that, all the values of the (estimated β_1) of the different grouped lies within the range of the confidence intervals of the ungrouped model (-7.12, -2.52) which shows Significant confidence intervals, that means without any doubt, aggregation or grouping does not affect the actual ungrouped value of the estimated β_1.

Table 11: Estimated β_2 for ungrouped and different grouped models

Grouping	Ungrouped	15	10	8	6	5	4
β_2	2.131759	1.96	2.16	2.09	1.86	1.72	2.57
St. Error	0.601952	0.59	0.57	0.57	0.67	0.58	0.67
P-value	0.000557	0.001	0.0002	0.0004	0.007	0.003	0.0002

:

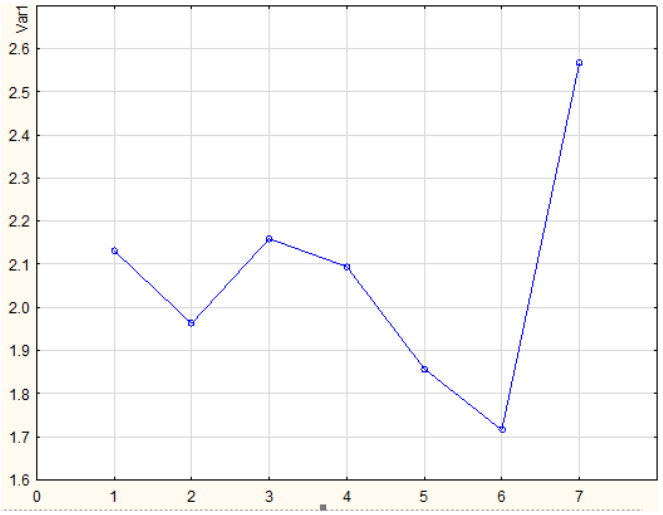


Figure 3: Estimated β_2 for ungrouped and different grouped models

4.3. Discussion of the effect of Grouping on the Estimated β_2. From table (11) and figure (3), we noticed that, all the values of the (estimated β_2) of the different grouped lies within the range of the confidence intervals of the ungrouped model (0.94, 3.324) which shows Significant confidence intervals, that means without any doubt, aggregation or grouping does not affect the actual ungrouped value of the estimated β_2.

From the above discussions, we can sum up the results of this study as followings:

4.3.1 Grouping gives the same results for the regression coefficient for different groups.

4.3.2 All regression coefficients for different groups are unbiased estimate.

4.3.3 All estimated confidence intervals for different groups are unbiased estimate.

4.3.4 All estimated standard errors for the different multiple regression coefficients different groups are unbiased estimate.

4.3.5 We should use only 15 to 20 class intervals for grouping at most.

4.3.6 We should use only 5 class intervals for grouping at least.

4.3.7 Grouping does not give misleading results regarding estimation.

References

[1] *Lukas Racickas March 21,2023 aggregation: Definition, Benefits, and Examples*. https://coresignal.com/blog/data- aggregation/
[2] Aggregating regression procedures to improve performance Y Yang Bernoulli, 2004 project Euclid .org.

[3] [Adil M. Youniss, a PhD dissertation 2002].
[4] JUDITSKY, A. and NEMIROVSKI, A. (2000) Functional aggregation for nonparametric regression. Ann. Statist. 28 681–712. MR1792783
[5] YANG, Y. (2004). Aggregating regression procedures to improve performance. Bernoulli 10 25–47. MR2044592
[6] TSYBAKOV, A. B. (2003). Optimal rates of aggregation. In Learning Theory and Kernel Machines. Lecture Notes in Artificial Intelligence 2777 303–313. Springer, Heidelberg.
[7] Greg Harvey, Microsoft Excel 2010 All-in-One for Dummies, Published by Wiley Publishing, Inc. 111 River Street Hoboken, NJ 07030-5774.
[8] Kor. K. & Altun, G., 2020. Is Support Vector Regression method suitable for predicting rate. Journal of Petroleum Science and Engineering,194.
[9] Frost, J., 2023. Statistics By Jim Making statistics intuitive. [Online] Available at: https://statisticsbyjim.com/regression/mean-squared-error-mse/.
[10] Ali, P. A. & Younas, A. A., 2021. Understanding and Interpreting Regression Analysis. Evid Based Nurs, 24(4), pp. 116-118.