EUROPEAN ACADEMIC RESEARCH
Vol. XII, Issue 9/ December 2024

Impact Factor: 3.4546 (UIF)
DRJI Value: 5.9 (B+)

ISSN 2286-4822
www.euacademic.org

The Impact of Data Grouping on the Standard Errors of the Regression Coefficients and the Variance of the Estimated Multiple Linear Regression Model

ADIL MOUSA YOUNIS WANISS[1]
Associate Professor, Department of Management Information Systems
College of Administration and Economics, Qassim University, Buraidah, P.O. Box 6633, KSA
A.waniss@qu.edu.sa
NIZAR ABDULLAH ISMAIL ALSUFI
Assistant Professor, Department of Management Information Systems
College of Administration and Economics, Qassim University, Buraidah
na.alsofi@qu.edu.sa

Abstract

This paper is completely focused on the impact of data grouping on the standard errors of the regression coefficients and the variance and standard error of the estimated multiple linear regression , we generate random numbers from Excel and built multiple linear regression model with two independent variables ,we compared ungrouped multiple linear regression model with different sex aggregated linear regression models (15 , 10, 8 , 6 , 5 and 4) we find that for all grouped models , except in case of grouping by less than five grouped the standard errors of the regression coefficients are more consistent and less dispersion than those of the ungrouped model, also the variances and the standard errors of the different grouped are less than the variance and standard error of the grouped model. The most important finding is that the estimated regression coefficients of grouped models are unbiased estimators of the estimated regression coefficients of grouped model. Groupings don't affect the significance of the different models. We find that grouping should not go a bove twenty groups or below four groups. The most important flinging is that all ANOVA tables for different groupings gives significant p- values, and grouping should be applied for all studies in order to save time and cost since it doesn't affect the statistics of the ungrouped model.

Keywords: data grouping, standard errors, regression coefficients, variance of the estimated multiple linear regression model

1-INTRODUCTION:

Aggregation refers to the process of combining or grouping multiple data points or pieces of information into a summary or a single representative form. This technique is often used in statistics, data analysis, and research to simplify complex data sets and highlight key trends or insights. In this paper we will use statistical aggregation, this includes methods like mean, median, mode, and standard deviation to summarize data. companies use aggregation to analyze sales data, customer feedback, and market trends. 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

[1] Corresponding author: A.waniss@qu.edu.sa or younisadil2002@gmail.com

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). 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 rojecteuclid.org). In this paper we will aggregate stimulated data of dependent variable with multiple independents variables in order to investigate it is impact on the standard errors of the multiple regression coefficients of the multiple linear regression. The use of random estimates in regression models has been gaining more attention in recent years (JUDITSKY, A. and NEMIROVSKI, A. (2000). Functional aggregation for nonparametric regression. Ann. Statist. 28 681–712. MR1792783)

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 multiple coefficient of determination could similarly be influenced.

2- STANDARD ERRORS OF THE MULTIPLE LINEAR REGRESSION COEFFICIENTS:

The multiple liner regression model is

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

Where y is the dependent variable , x_i are the independent variables, $\beta_o, \beta_1, \beta_2, , \cdots, \beta_{k-1}$ are the regression coefficients of the model and ε is the residuals, 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).

According to the least square method the coefficients are calculated as follows:

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

$$B = A^{-1} \cdot C \quad (2)$$

In multiple linear regression, the standard errors of the regression coefficients provide a measure of the variability or uncertainty associated with each estimated coefficient. These standard errors are essential for hypothesis testing and construction of confidence intervals T test for the regression coefficients as in the followings equations,

$$T = \frac{b_i - \beta_{io}}{S(b_i)} \quad (3)$$

$$b_i \pm T(\alpha / 2, n-2) S(b_i) \quad (4)$$

Here's how they are generally calculated and interpreted:

The Standard error of the multiple linear regression coefficients will be calculated from the variance covariance matrix as follows :

$$\sigma^2 = \frac{SSE}{n-k} \cdot A^{-1} \qquad (5)$$

Where $\frac{SSE}{n-k}$ is the estimated multiple linear regression model variance and A^{-1} is the invers matrix.
We get from the matrix diagonal
We get from the matrix diagonal

$$S^2(b_0)\ , S^2(b_1)\ , \ldots, \ S^2(b_{k-1}) \qquad (6)$$

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

$$S(b_0)\ , S(b_1)\ , \ldots, \ S\ \ (b_{k-1}) \qquad (7)$$

are the standard of errors of the regression coefficients (Frost, J., 2023).
Now on the followings paragraphs we will calculate the variances and standard errors of the regression coefficients and the estimated multiple linear regression model variance for ungrouped data and for different grouped data , then we compare the resulted values of different grouped with each other as well as with the value of grouped to see the impact of grouping data on the them .

3- GENERATION OF UNGROUPED DATA :

The data for this study generated randomly from Excel for 130 values according to the relation between quantity demanded (y) as dependent variable and income (x_1) and the price of the commodity (x_2) as independent variables, the data generated according to the roles and laws of the correlation between demanded quantity and it is price which is negative correlation and between quantity demanded and income of the consumer which is positive correlation We calculate the multiple regression model for ungrouped data from Excel (2010), using both of "Linest" and "Regression Statistics" (Greg Harvey, Microsoft Excel 2010).

3-1Summary of the calculated Statistics from Ungrouped Multiple Model:

Table (1): ANOVA for Ungrouped multiple regression :

S.O.V.	*S.S.*	*df*	*MSS.*	*F.*	*P value*
Regression	45231.01	2	22615.51	3258.203	7.4E-110
Residual	881.5193	127	6.941097		
Total	46112.53	129			

From table (1), the variance of the multiple linear regression model is 6.94 and it is standard error is 2.63, in the following paragraphs we will grouped the same data into different groups and calculate their models with variances and standard errors and compare them with above ungrouped statistics.

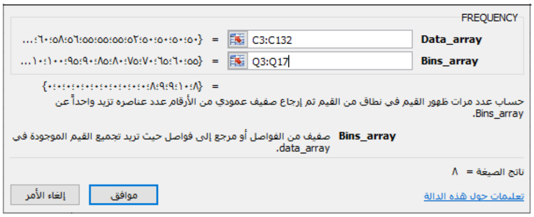
Table (2): Regression Coefficients and their Standard Errors for Ungrouped multiple regression Model :

	Coefficients	*Standard Error*	*t – test*	*P-value*
Intercept	90.95274	12.29484	7.397634	1.66E-11
Price	-4.8228	1.161283	-4.15299	5.98E-05
Income	2.131759	0.601952	3.541407	0.000557
Model		2.634596		

From table (2), the standard errors of the multiple linear regression coefficients model are 12.29 , 1.16 and 0.63 respectively and it is standard error of the model is 2.63, in the following paragraphs we will grouped the same data into different groups and calculate models coefficients and their standard errors along with the models standard errors and compare them with above ungrouped statistics.

4- GROUPING 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 :



FREQUENCY

Data_array C3:C132 = {٥٠؛٥٠؛٥٠؛٥٠؛٥٢؛٥٥؛٥٥؛٥٥؛٥٦؛٥٨؛٦٠؛...

Bins_array Q3:Q17 = {٥٥؛٦٠؛٦٥؛٧٠؛٧٥؛٨٠؛٨٥؛٩٠؛٩٥؛١٠٠؛١٠...

= {٨؛١٠؛٩؛٩؛٨؛٠؛٠؛٠؛٠؛٠؛٠؛٠؛٠؛٠؛٠؛٠}

حساب عدد مرات ظهور القيم في نطاق من القيم ثم إرجاع صفيف عمودي من الأرقام عدد عناصره تزيد واحداً عن Bins_array.

Bins_array صفيف من الفواصل أو مرجع إلى فواصل حيث تريد تجميع القيم الموجودة في data_array.

ناتج الصيغة = ٨

4-1 Summary of the calculated Statistics from grouped Multiple Model with 15 class intervals:

Table (3): ANOVA for grouped multiple regression with 15 class intervals :

S.O.V.	*S.S.*	*df*	*MSS.*	*F.*	*P value*
Regression	45081.39	2	22540.69	3942.538343	5.055E-115
Residual	726.0978	127	5.717305		
Total	45807.49	129			

From table (3), the variance of the grouped multiple linear regression model with 15 class intervals is 5.71 and it is standard error is 2.39, now if we compare them with the variance and standard error of the ungrouped model, we observed that the both of them were less than that of ungrouped multiple linear regression model. That means grouping is minimizing the measures of dispersion .

Table (4): Regression Coefficients and their Standard Errors for grouped multiple regression with 15 class intervals:

	Coefficients	*Standard Error*	*t – test*	*P-value*
Intercept	94.51695	12.05586	7.397634	1.66E-11
Price	-5.16822	1.138538	-4.15299	5.98E-05
Income	1.962769	0.590191	3.541407	0.000557
Model		2.391087		

From table (4), the standard errors of the estimated regression coefficients with 15 grouped class model are 12.06 , 1.14 and 0.59 respectively and it is standard error of the model is 2.39, if we compare them with the standard errors of the estimated standard errors of ungrouped model it is clear that the standard errors of regression coefficients of the grouped model are more consistent and less dispersion than those of the grouped model. But the regression coefficients of the ungrouped model are approximately unbiased estimator of the regression coefficients of the grouped model.

4-2 Summary of the calculated Statistics from grouped Multiple Model with 10 class intervals:

Table (5) :ANOVA for grouped multiple regression with 10 class intervals :

S.O.V.	*S.S.*	*df*	*MSS.*	*F.*	*P value*
Regression	44774.78	2	22387.39	4201.658	9.4E-117
Residual	676.6848	127	5.328227		
Total	45451.46	129			

From table (5), the variance of the grouped multiple linear regression model with 10 class intervals is 5.32 and it is standard error is 2.31, now if we compare them with the variance and standard error of the ungrouped model and of 15 class intervals, we observed that both of them were less than that of ungrouped multiple linear regression model and of 15 class intervals as well . That means grouping is minimizing the measures of dispersion more and more whenever we intense the grouping.

Table (6): Regression Coefficients and their Standard Errors for grouped multiple regression with 10 class intervals:

	Coefficients	*Standard Error*	*t − test*	*P-value*
Intercept	90.60594	11.61374	7.801613	1.95E-12
Price	-4.80812	1.097312	-4.38172	2.44E-05
Income	2.159327	0.568255	3.799928	0.000223
Model		2.30829		

From table (6), the standard errors of the 10 class grouped multiple linear regression model are 11.61 , 1.097 and 0.57 respectively and it is standard error is 2.31, if we compare them with the standard errors of the ungrouped and of 15 grouped class intervals, it is clear that the standard errors of the regression coefficients of the grouped model are more consistent and less dispersion than those of the grouped mode as well of the 15 grouped. But the regression coefficients of the 10 grouped model are approximately unbiased estimator of the regression coefficients of the ungrouped model.

4-3 Summary of the calculated Statistics from grouped Multiple Model with 8 class intervals:

Table (7): ANOVA for grouped multiple regression with 8 class intervals :

S.O.V.	*S.S.*	*df*	*MSS.*	*F.*	*P value*
Regression	44403.68	2	22201.84	4615.789385	2.6285E-119
Residual	610.867	127	4.809976		
Total	45014.54	129			

From table (7), the variance of the grouped multiple linear regression model with 8 class intervals is 4.81 and it is standard error is 2.19, now if we compare them with the variance and standard error of the ungrouped model and of 15 and 10 class intervals, we observed that both of them were less than that of ungrouped multiple linear regression model and of 15 and 10 class intervals as well . That means grouping is minimizing the measures of dispersion more and more whenever we intense the grouping.

Table (8): Regression Coefficients and their Standard Errors for grouped multiple regression with 8 class intervals:

	Coefficients	*Standard Error*	$t-test$	*P-value*
Intercept	91.91627	11.68455	7.801613	1.95E-12
Price	-4.93022	1.103312	-4.38172	2.44E-05
Income	2.09433	0.572022	3.799928	0.000223
Model	91.91627	2.193166		

From table (8), the standard errors of the 8 class grouped multiple linear regression model are 11.68, 1.0337 and 0.57 respectively and it is standard error is 2.19, if we compare them with the standard errors of the ungrouped and of 15 and 10 grouped class intervals, it is clear that the regression coefficients of the 8 grouped model are more consistent and less dispersion than those of the ungrouped mode as well of the 15 and 10 grouped. But the regression coefficients of the 8 grouped model are approximately unbiased estimator of the regression coefficients of the grouped model.

4-4 Summary of the calculated Statistics from grouped Multiple Model with 6 class intervals:

Table (9): ANOVA for grouped multiple regression with 6 class intervals :

$S.O.V.$	$S.S.$	df	$MSS.$	$F.$	P $value$
Regression	43748.09	2	21874.05	4994.400083	1.8794E-121
Residual	556.2238	127	4.379715		
Total	44304.32	129			

From table (9), the variance of the grouped multiple linear regression model with 8 class intervals is 4.38 and it is standard error is 2.09, now if we compare them with the variances and standard errors of the ungrouped model and of 15 and 10 class intervals, we observed that both of them were less than that of ungrouped multiple linear regression model and of 15 and 10 class intervals as well. That means grouping is minimizing the measures of dispersion more and more whenever we intense the grouping.

Table (10): Regression Coefficients and their Standard Errors for grouped multiple regression with 6 class intervals:

	Coefficients	*Standard Error*	$t-test$	*P-value*
Intercept	90.60594	13.75099	7.801613	1.95E-12
Price	-4.80812	1.296438	-4.38172	2.44E-05
Income	2.159327	0.673998	3.799928	0.000223
Model		2.092777		

From table (10), the standard errors of the 8 class grouped multiple linear regression model are 13.75, 1.30 and 0.67 respectively and it is standard error is 2.09, if we compare them with the standard errors of the ungrouped and of 15 , 10 and 8 grouped class intervals, it is clear that the regression coefficients of the 6 grouped model are more consistent and less dispersion than those of the ungrouped mode as well of the 15 , 10 and 8 grouped. But the regression coefficients of the 6 grouped model are approximately unbiased estimator of the regression coefficients of the ungrouped model.

4-5 Summary of the calculated Statistics from grouped Multiple Model with 5class intervals:

Table (11): ANOVA for grouped multiple regression with 5 class intervals:

$S.O.V.$	$S.S.$	df	$MSS.$	$F.$	$P\ value$
Regression	43233.25	2	21616.62	4994.400083	1.8794E-121
Residual	446.1202	127	3.512758		
Total	43679.37	129			

From table (11), the variance of the grouped multiple linear regression model with 8 class intervals is 3.5 and it is standard error is 1.87, now if we compare them with the variances and standard errors of the ungrouped model and of 15 and 10 class intervals, we observed that both of them were less than that of ungrouped multiple linear regression model and of 15 and 10 class intervals as well . That means grouping is minimizing the measures of dispersion more and more whenever we intense the grouping.

Table (12): Regression Coefficients and their Standard Errors for grouped multiple regression with 5 class intervals:

	Coefficients	*Standard Error*	$t-test$	*P-value*
Intercept	99.75919	11.77596	7.801613	1.95E-12
Price	-5.67863	1.113877	-4.38172	2.44E-05
Income	1.715602	0.575123	3.799928	0.000223
Model		1.874235		

From table (12), the standard errors of the 8 class grouped multiple linear regression model are 11.78 , 1.11 and 0.58 respectively and it is standard error is 1.87, if we compare them with the standard errors of the ungrouped and of 15 , 10 , 8 and 6 grouped class intervals, it is clear that the regression coefficients of the 5 grouped model are more consistent and less dispersion than those of the ungrouped mode as well of the 15 , 10 , 8 and 6 grouped. But the regression coefficients of the 5 grouped model are approximately unbiased estimator of the regression coefficients of the ungrouped model.

4-6 Summary of the calculated Statistics from grouped Multiple Model with 4 class intervals:

Table (12): ANOVA for grouped multiple regression with 4 class intervals :

$S.O.V.$	$S.S.$	df	$MSS.$	$F.$	$P\ value$
Regression	41579.84	2	20789.92	4773.791	3.2E-120
Residual	553.0866	127	4.355013		
Total	42132.93	129			

From table (12), the variance of the grouped multiple linear regression model with 4 class intervals is 4.4 and it is standard error is 2.09, now if we compare them with the variances and standard errors of the ungrouped model and of 15 and 10 class intervals, we observed that both of them were less than that of ungrouped multiple linear regression model and of 15 and 10 class intervals as well . That means grouping is minimizing the measures of dispersion more and more whenever we intense the grouping.

Table (13): Regression Coefficients and their Standard Errors for grouped multiple regression with 5 class intervals:

	Coefficients	*Standard Error*	*t − test*	*P-value*
Intercept	90.60594	13.70618	7.801613	1.95E-12
Price	-4.80812	1.293892	-4.38172	2.44E-05
Income	2.159327	0.670857	3.799928	0.000223
Model		2.086867		

From table (13), the standard errors of the 4 class grouped multiple linear regression model are 13.71 , 1.29 and 0.67respectively and it is standard error is 2.09, if we compare them with the standard errors of the ungrouped and of 15 , 10 , 8 ,6 and 5 grouped class intervals, it is clear that the regression coefficients of the 4 grouped model are more less consistent and more dispersion than those of the ungrouped mode as well of the 15 , 10 , 8 , 6 and 5 grouped. But the regression coefficients of the 4 grouped model are not unbiased estimator of the regression coefficients of the ungrouped model.

5- CONCLUSION:

On the light of the above results, if we compare the standard errors of the regression coefficients of the sex grouped models with that of ungrouped model we conclude the Grouping doesn't affect the statistics of the ungrouped model as shown in table (14).

Table (14): Standard errors of the multiple regression Coefficients for different grouped and ungrouped models:

	Ungrouped	15	10	8	6	5	4
Intercept	12.29484	12.0558	11.61374	11.68455	13.75099	11.77596	13.70618
Price	1.161283	1.13853	1.097312	1.103312	1.296438	1.113877	1.293892
Income	0.601952	0.59019	0.568255	0.572022	0.673998	0.575123	0.670857
Model	2.634596	2.39108	2.30829	2.193166	2.092777	1.874235	2.086867

Table (15): Regression Coefficients for ungrouped and different grouped multiple regression models:

	Ungrouped	15	10	8	6	5	4
Intercept	90.95274	94.516	90.60594	91.91627	90.60594	99.75919	90.60594
Price	-4.8228	-5.168	-4.80812	-4.93022	-4.80812	-5.67863	-4.80812
Income	2.131759	1.9627	2.159327	2.09433	2.159327	1.715602	2.159327

According to the above results, if we compare the coefficients of regression of the different grouped models with that of ungrouped model it is clear that all coefficients are unbiased estimators of the grouped coefficients. We conclude that grouping doesn't affect the biasedness of the ungrouped model as shown in table (15).

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 • projecteuclid.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

[6] Yang, Y. (2004). Aggregating regression procedures to improve performance. Bernoulli 10 25–47. MR2044592

[7] Tsybakov, A. B. (2003). Optimal rates of aggregation. In Learning Theory and Kernel Machines. Lecture Notes in Artificial Intelligence 2777 303–313. Springer, Heidelberg.

[8] Greg Harvey, Microsoft Excel 2010 All-in-One for Dummies, Published by Wiley Publishing, Inc. 111 River Street Hoboken, NJ 07030-5774.

[9] Kor, K. & Altun, G., 2020. Is Support Vector Regression method suitable for predicting rate. Journal of Petroleum Science and Engineering,194

[10] Frost, J., 2023. Statistics By Jim Making statistics intuitive. [Online] Available at: https://statisticsbyjim.com/regression/mean-squared-error-mse/

[11] Ali, P. A. & Younas, A. A., 2021. Understanding and Interpreting Regression Analysis. Evid Based Nurs, 24(4), pp. 116-118.

[12] Adil M. Younis and Abdulmajied Ali Balkash. International Journal of Science and Research (IJSR)