THE DISTRIBUTION OF $R^2/(1- R^2)$ IN THE AGGREGATED LINEAR REGRESSION MODEL

ADEL MOUSA YOUNES[1] AND BASSAM YOUNES IBRAHIM [1]

ABSTRACT

This study deals with simple linear regression model in the grouped data. The coefficient of determination R^2 based on grouped data will always be higher than that based on ungrouped data. Since the increments of the R^2 will be smaller for smaller number of data group points, therefore the statistic $R^2/(1- R^2)$ was taken which shows the significance of these increments. The data concerned with monthly income and monthly total expenditure (10000 S.D) are collected from 1000 families in different areas in Khartoum. The results show the good fitting of the exponential model to the aggregated data depending on some statistical criteria.

الملخص

تناولت هذه الدراسة نموذج الانحدار الخطي البسيط في حالة البيانات المتجمعة، وتبين ان معامل التحديد لنموذج الانحدار للبيانات المتجمعة اكبر من معامل التحديد R^2 لنموذج الانحدار للبيانات المجتمعة اكبر من معامل التحديد لنموذج الانحدار للبيانات غير المتجمعة. ونسبة إلى أن الزيادات في R^2 معامل التحديد " تكون صغيرة للتجميع الصغير عليه أخذت الإحصائية $R^2/(1- R^2)$ والتي توضح معنوية الزيادات. بيانات الدراسات متعلقة بالدخول الشهرية والإنفاق الكلي الشهري (عشرة الف دينار سوداني) حيث جمعت من ١٠٠٠ أسرة سودانية مختلفة في مدينة الخرطوم. بينت النتائج جودة توفيق البيانات على النموذج الأس للبيانات المتجمعة وذلك اعتمادا على بعض المعايير الإحصائية.

INTRODUCTION

In the fitting linear regression models to the relation between household income and expenditure on distinct commodities it has for some time been standard practice to avoid large-scale computations by grouping the individual households into a small number of income classes, and then to fit regression lines to the class means, weighted by the number of households in each class. As far as estimation is concerned, this method has been justified by Prais and Aitchison[1], who have shown that: (i) the regression estimates are unbiased, (ii) they have a larger variance than estimates based on individual observations, and (iii) this loss of efficiency

[1] Dept. of Applied Statistics, Faculty of Science (SUST)

depends on the manner of grouping and is least if the observations are grouped together according to the value of the explanatory variable income. Prais and Aitchison did not deal with the effect of grouping on the correlation coefficient, although they noted that the correlation coefficient based on grouped data is quite unsatisfactory estimate of the correlation in the population and in consequence of little statistical interest. This view now is keeping with the current trend of emphasizing regression rather than correlation. It is true that in the regression model R^2, the coefficient of determination, is a mere sample statistic that admits of no probabilistic interpretation. Still, as long as it is continued to compute R^2, one should admit that a certain importance to it is attached. In fact high values are regarded as satisfactory and low values as warning evidence that much remains unexplained.

In this respect the results of grouped data are completely unreliable as they lead to much larger values of R^2 than individual observations. While this in itself is widely understood, it is believed that it is rarely appreciated to what extent grouping may increase R^2. Gollnick fitted Engel curves for two broad food groups to individual observations as well as to the same data grouped by income classes of varying importance. The results of this study showed that the slopes of the regression lines are little affected, while the correlation coefficients are considerably increased as more and more households are grouped together[1].

For most commodities taken separately, income seldom accounts for more than 20 percent of the individual variation of expenditure[2]. The proportion is larger for broad commodity groups, and it may reach a quite respectable level if one considers, say, all food expenditure with total expenditure in lieu of income as the explanatory variable.

The primary purpose of this study is to examine the effect of grouping on the coefficient of determination, and so on the distribution of the statistic $R^2/(1- R^2)$ which gives the significance of little increments on the R^2 from aggregated model, considering the regression model with fixed values of explanatory variable and additive disturbances as the only random element. The discussion relies to a large extent on approximations which permit in the end to evaluate the increase in R^2 that under usual conditions of budget survey analysis may be expected to occur if (N) households are grouped into (g) income classes.

THE MODEL :

Consider the simple linear regression model

$$y_i = \alpha + \beta x_i + u_i \qquad \text{.....(1)}$$

Where α and β are unknown constants, u_i is the disturbance term which satisfies the following assumptions:

(i) $u_i \sim N(0, \sigma_u^2)$

(ii) $Eu_iu_j=0 \quad \forall i = j \qquad i, j = 1,2,...,N$

(iii) $Eu_ix_i=0 \quad \forall \qquad i = 1,2,...,N$

The estimated model for model (1) is:

$$\hat{y}_i = \hat{\alpha} + \hat{\beta}x_i \qquad \text{.....(2)}$$

where $\hat{\alpha}$ and $\hat{\beta}$ are the least squares estimators such that:

$$\hat{\beta} = \frac{S_{xy}}{S_{xx}} \qquad (3)$$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$

and

$$S_{xy} = \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}$$

$$S_{xx} = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2$$

It is easy to show that the slope $\hat{\beta}$ is unbiased estimator for β, i.e. $E\hat{\beta}=\beta$.

Now, consider the regression model for a sample of N observations (x_{ij}, y_{ij}) which is from the outset divided into g groups of n_i observations each, so that $i=1,2,..g$ and $j=1,2,...,n_i$ within each group, $\sum_{i=1}^{g} n_i = N$. At this stage no assumption was made whatever about the nature of the groups. x_{ij} are fixed and given, and y_{ij} are in repeated samples defined by:

$$y_{ij} = \alpha + \beta . x_{ij} + u_{ij} \qquad \text{.....(4)}$$

Where u_{ij} satisfies the above assumptions.

Overall sample means are denoted by a double bar and group means by a single bar, so that:

$$\bar{\bar{x}} = \frac{1}{N}\sum_{i=1}^{g}\sum_{j=1}^{n_i} x_{ij} \qquad \bar{\bar{y}} = \frac{1}{N}\sum_{i=1}^{g}\sum_{j=1}^{n_i} y_{ij}$$

$$\bar{x}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} x_{ij} \qquad \bar{y}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij} \qquad \text{.....(5)}$$

The sums of squares and cross-products of the variables, centered around the overall sample means are:

$$S_{xx} = \sum_{i=1}^{g}\sum_{j=1}^{n_i}(x_{ij} - \bar{\bar{x}})^2$$

$$S_{xy} = \sum_{i=1}^{g}\sum_{j=1}^{n_i}(x_{ij} - \bar{\bar{x}})(y_{ij} - \bar{\bar{y}}) \qquad \text{.....(6)}$$

The corresponding expressions calculated from the weighted group means are :

$$S_{\bar{x}\bar{x}} = \sum_{i=1}^{g} n_i(\bar{x}_i - \bar{\bar{x}})^2$$

$$S_{\bar{x}\bar{y}} = \sum_{i=1}^{g} n_i(\bar{x}_i - \bar{\bar{x}})(\bar{y}_i - \bar{\bar{y}}) \qquad \text{.....(7)}$$

Therefore the linear regression model for grouped data becomes:

$$\bar{y}_i = \tilde{\alpha} + \tilde{\beta}\bar{x}_i + \bar{u}_i \qquad \text{.....(8)}$$

here $\bar{u}_i$ satisfies the assumptions of the ordinary regression model, and the least squares estimators of this model are :

$$\hat{\tilde{\beta}} = \frac{S_{\bar{x}\bar{y}}}{S_{\bar{x}\bar{x}}} \qquad \text{.....(9)}$$

$$\hat{\tilde{\alpha}} = \bar{\bar{y}} - \hat{\tilde{\beta}}.\bar{\bar{x}}$$

Therefore, the estimated model for eq.(8) is :

$$\hat{\bar{y}}_i = \hat{\bar{\alpha}} + \hat{\bar{\beta}}\,\bar{x}_i \qquad \ldots\ldots\ldots(10)$$

It is easy to show that[3] $E\,\hat{\bar{\beta}} = \bar{\beta}$

The coefficient of determination in the simple linear regression model is :

$$R^2 = \frac{SS_{Reg.}}{SS_{Tot.}} \qquad \ldots\ldots\ldots(11)$$

In the ordinary regression model (2), the sum of squares of regression is:

$$SS_{Reg.} = \hat{\beta}\,S_{xy} = \frac{\left[\sum_{i=1}^{N} x_i y_i - N\bar{x}\bar{y}\right]^2}{\sum_{i=1}^{N} x_i^2 - N\bar{x}^2} \qquad (12)$$

And the sum of squares of totals is :

$$SS_{Tot.} = S_{yy} = \sum_{i=1}^{N} y_i^2 - N\bar{y}^2 \qquad (13)$$

Eq.(11) becomes :

$$R^2 = \frac{\left[\sum_{i=1}^{N} x_i y_i - N\bar{x}\bar{y}\right]^2}{\left[\sum_{i=1}^{N} x_i^2 - N\bar{x}^2\right]\left[\sum_{i=1}^{N} y_i^2 - N\bar{y}^2\right]} \qquad (14)$$

While, in the aggregated regression model (10), it is noted that :

$$SS_{Reg.} = \hat{\beta} S_{\bar{x}\bar{y}} = \frac{[\sum_{i=1}^{g} n_i \bar{x}_i \bar{y}_i - N \bar{\bar{x}} \bar{\bar{y}}]^2}{\sum_{i=1}^{g} n_i \bar{x}_i^2 - N \bar{\bar{x}}^2} \quad (15)$$

$$SS_{Tot.} = S_{\bar{y}\bar{y}} = \sum_{i=1}^{g} n_i \bar{y}_i^2 - N \bar{\bar{y}}^2 \quad (16)$$

Therefore, the R^2 in the aggregated model is:

$$R_k^2 = \frac{[\sum_{i=1}^{g} n_i \bar{x}_i \bar{y}_i - N \bar{\bar{x}} \bar{\bar{y}}]^2}{[\sum_{i=1}^{g} n_i \bar{x}_i^2 - N \bar{\bar{x}}^2][\sum_{i=1}^{g} n_i \bar{y}_i^2 - N \bar{\bar{y}}^2]} \quad (17)$$

But the means in equation (12) are equal to the overall means in equation (17), i.e.

$$\bar{\bar{x}} = \bar{x} \quad , \quad \bar{\bar{y}} = \bar{y} \quad (18)$$

And it is easy to prove that the following inequalities hold

$$\begin{aligned} \sum_{i=1}^{g} n_i \bar{x}_i \bar{y}_i &\geq \sum_{i=1}^{N} x_i y_i \\ \sum_{i=1}^{g} n_i \bar{x}_i^2 &\leq \sum_{i=1}^{N} x_i^2 \\ \sum_{i=1}^{g} n_i \bar{y}_i^2 &\leq \sum_{i=1}^{N} y_i^2 \end{aligned} \quad (19)$$

Therefore, from equations (18) and (19), it is concluded that:

$$R_k^2 \geq R^2 \qquad (20)$$

Thus the coefficient of determination based on grouped data will always be higher than that is based on ungrouped data.

EFFICIENCY OF THE MODEL

Any simple increment in the coefficient of determination may not be significant especially for the economic variables. Therefore, one may take the statistic $R^2/(1-R^2)$ which show the real significance of these increments.

Now, it could be shown that the statistic $I= R^2/(1-R^2)$ decreases exponentially with increase in size of group according to the model:

$$\hat{I}_k = a + e^{b+cG_k} \qquad k = 1,2,\ldots,g \qquad (21)$$

depending on :

1. F-test from ANOVA .
2. R_k^2 Coefficient of determination for each aggregated model.
3. χ^2-test for observed and estimated values.
4. Theil's U-statistic test for the forecasting power of the model.

The vaule of F from ANOVA table using ungrouped data is computed from :

$$F = \frac{SS_{Reg.}}{MS_{error}} = \frac{SS_{Reg.}}{[SS_{Tot.} - SS_{Reg.}]/N-2} \qquad \ldots\ldots\ldots(22)$$

while in grouped data, $SS_{Reg.}$ and $SS_{Tot.}$ are computed from equations (15) and (16), and the degree of freedom of the error becomes N-g.

The χ^2 statistic is computed from :

$$\chi^2 = \sum_{k=1}^{g} \frac{(I_k - \hat{I}_k)^2}{\hat{I}_k} \qquad \ldots\ldots\ldots(23)$$

where I_k is the observed value of $R_k^2/(1-R_k^2)$, and $\hat{I}_k$ is the estimated value of I_k computed from eq (21).

The Theil's U-statistic which is used to test the forecasting power of the model(21) is defined by[4]:

$$U = \frac{SS_{Res.}}{\sqrt{\sum_{k=1}^{g} I_k^2} + \sqrt{\sum_{k=1}^{g} \hat{I}_k^2}} \quad \ldots\ldots (24)$$

$0 \le U \le 1$, when $U \to 0$ that means a very high forecasting power of the model, while $U \to 1$ shows the weak forecasting power of the model.

THE DATA

Data were collected from 1000 Sudanese families from different areas in Khartoum city. This data contains two variables, the monthly income(100 SDD) represented by x and the monthly total expenditure(100 SDD) represented by y, we have aggregated the data according to ten classes of income as in the following table.

Table(1): Income classes for the aggregated data

Classes	No.of Groups	Income classes(100 SDD)
1	2	<300 , 300-600
2	3	<200 , 200-400 , 400-600
3	4	<150 , 150-300 , 300-450 , 450-600
4	5	<120 , 120-240 , 240-360 ,360-480 , 480-600
5	6	<100 , 100-200 , 200-300 , 300-400 , 400-500, 500-600
6	7	<86 , 86-172 , 172-258 , 258-344 , 344-430, 430-5160, 516-600
7	8	<75 , 75- , 150-225 , 225-300 , 300-3750,375-450 , 450-525 , 525-600
8	9	<67 , 67-134 , 134-201 , 201-268 , 268-33, 335-402 , 402-4690 , 469-536 , 536-600
9	10	<60 , 60-120 , 120-180 , 180-240 , 240-300, 300-360 , 360-420 , 420-480 , 480-540 , 540-600

RESULTS[2] AND DISCUSSION:

For ungrouped data with 1000 observations, one gets the following results:

$\hat{y}_i = 303.597 + 0.537 x_i$ i=1,2,...,1000

R^2=0.70 F=2332.846

After aggregating the data according to Table (1), one gets the following results:

Group 2 :

Table (2): Aggregated data for group 2

n_i	$\bar{x}_i$	$\bar{y}_i$
958	759.84	717.57
42	4118.52	2387.84

$\hat{\bar{y}}_i = 339.7965 + 0.4972 \bar{x}_i$ i=1,2 R_1^2=0.99998 F=49899002

Group 3 :

Table (3): Aggregated data for group 3

n_i	$\bar{x}_i$	$\bar{y}_i$
929	714.12	695.59
51	2619.12	1617.51
20	5195.40	2951.60

$\hat{\bar{y}}_i = 339.5263 + 0.4975 \bar{x}_i$ i=1,2,3 R_2^2=0.9996 F=2491503

Group 4 :

Table (4): Aggregated data for group 4

n_i	$\bar{x}_i$	$\bar{y}_i$
875	661.02	660.61
83	1801.60	1318.09
26	3280.69	1880.77
16	5480.00	3211.83

$\hat{\bar{y}}_i = 320.247 + 0.5189 \bar{x}_i$ i=1,2,3,4 R_3^2=0.994 F=165004

[2] SPSS and STATISTICA packages are used for all required computations

Group 5

Table (5): Aggregated data for group 5

n_i	$\bar{x}_i$	$\bar{y}_i$
848	641.12	644.57
99	1579.03	1265.91
32	2909.53	1739.64
7	4186.86	1917.58
14	5607.14	3336.33

$\hat{\bar{y}}_i = 322.8596 + 0.5160\,\bar{x}_i$ i=1,2,3,4,5 R_4^2=0.9740 F=37274.23

Group 6

Table (6): Aggregated data for group 6

n_i	$\bar{x}_i$	$\bar{y}_i$
776	605.80	617.42
153	1263.54	1092.06
29	2224.31	1421.92
22	3139.55	1875.33
12	4659.00	2434.40
8	6000.00	3727.40

$\hat{\bar{y}}_i = 315.9227 + 0.5237\,\bar{x}_i$ i=1,2,3,4,5,6 R_5^2=0.9757 F=319911.35

Group 7

Table (7): Aggregated data for group 7

n_i	$\bar{x}_i$	$\bar{y}_i$
712	577.17	592.59
208	1134.74	1018.22
36	2092.44	1422.02
19	2991.84	1837.94
9	3786.44	1833.75
7	4882.86	2588.59
9	5944.44	3696.57

$\hat{\bar{y}}_i = 313.1300 + 0.5268\,\bar{x}_i$ i=1,2,3,4,5,6,7 R_6^2=0.9666 F=28737.539

Group 8 :

Table (8): Aggregated data for group 8

n_i	$\bar{x}_i$	$\bar{y}_i$
712	577.17	592.59
208	1134.74	1018.22
36	2092.44	1422.02
19	2991.84	1837.94
9	3786.44	1833.75
7	4882.86	2588.59
9	5944.44	369.57

$\hat{\bar{y}}_i = 310.3371 + 0.5299\,\bar{x}_i$ i=1,2,3,4,5,6,7,8 $R^2_7 = 0.9613$ F=24641.07

Group 9 :

Table (9): Aggregated data for group 9

n_i	$\bar{x}_i$	$\bar{y}_i$
471	483.84	913.11
397	857.88	825.32
61	1556.62	1260.25
27	2180.74	1433.74
19	2991.84	1837.94
5	3570.00	1772.23
5	4145.60	2256.20
6	4946.67	2413.63
9	5944.44	2696.57

$\hat{\bar{y}}_i = 311.1479 + 0.5290\,\bar{x}_i$ i=1,2,3,4,5,6,7,8,9 $R^2_8 = 0.9620$ F=25087.95

Group 10 :

Table (10): Aggregated data for group 10

n_i	$\bar{x}_i$	$\bar{y}_i$
349	435.91	470.47
499	784.64	766.34
74	1443.76	1210.66
25	1979.44	1429.44
11	2539.55	1410.08
21	3103.33	1912.26
4	3975.00	1128.54
3	4469.33	2969.64
5	5000.00	2687.90
9	5944.44	3696.57

$\hat{\bar{y}}_i = 306.8236 + 0.5338\,\bar{x}_i$ i=1,2,3,4,5,6,7,8,9,10 $R^2_9 = 0.9371$ F=14749.27

The results of Tables (2)-(10) are summarized in Table (11)

Table(11): The results of all aggregated regression models

No. of groups	Groups	$\hat{\beta}$	R_k^2	F
1	2	0.4972	0.99998	49899002
2	3	0.4975	0.9996	2491503
3	4	0.5189	0.9940	165004
4	5	0.5160	0.9740	37274.23
5	6	0.5237	0.9757	319911.35
6	7	0.5268	0.9667	28737.54
7	8	0.5299	0.9613	24641.07
8	9	0.5290	0.9620	25087.95
9	10	0.5338	0.9371	14749.27
	ungrouped	0.5370	0.7000	2332.85

From Table (11), one notes that there is no significant difference between the regression coefficients. This is true also for the values of R_k^2 except R^2 for ungrouped data. Therefore, one computes $I_k = R_k^2/(1- R_k^2)$ for each group, as in Table (12).

Table(12): The values of I_k for all aggregated regression models

No. of groups	Groups(G_k)	I_k
1	2	45000.000
2	3	2499.000
3	4	165.667
4	5	37.462
5	6	40.184
6	7	29.030
7	8	24.826
8	9	25.337
9	10	14.900

To fit the exponential model as in equation (21) to the data in Table(12), QUASI-NEWTON which is one of the non-linear estimation methods is used [5], to get :

$\hat{I}_k = 27.6301 + \exp(16.561 - 2.9012\, G_k)$ $k=1,2,\ldots,9$ .....(25)

$R^2=0.99998$ $F=35459.25$

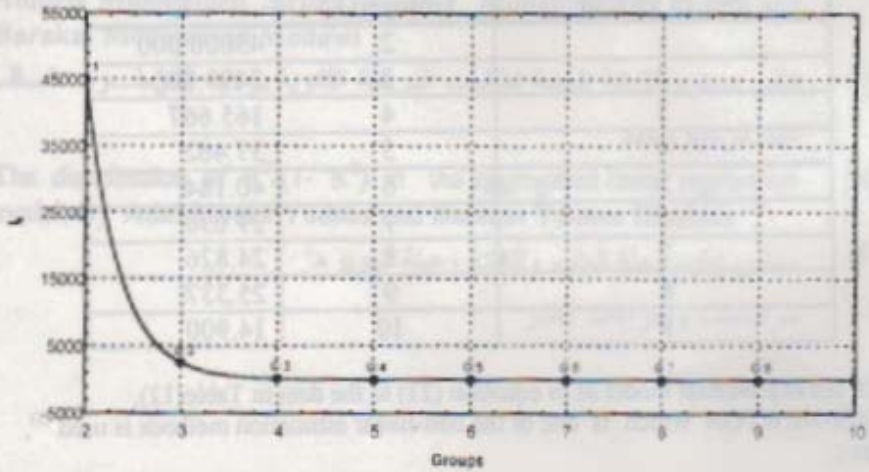
Very high values of R^2 and F reflects the good fitting of the above exponential model to the data. And by using χ^2 test as in equation (23), Table(13) is constructed as follows:

Table(13): Chi-square test for the observed and estimated values of I_k

k	I_k	$\hat{I}_k$	$(I_k - \hat{I}_k)^2 / \hat{I}_k$
1	45000.000	44999.991	0.000
2	2499.000	2499.273	0.000
3	165.667	163.475	0.029
4	37.462	35.103	0.159
5	40.184	28.049	5.259
6	29.030	27.653	0.068
7	24.826	27.638	0.285
8	25.337	27.635	0.190
9	14.900	27.637	5.865
			$\chi^2=11.857$

The tabulated χ^2 is $\chi^2_{8,0.05}=15.51$, so there is no significant difference between the observed and estimated values of I_k. This means that the exponential model reflects a good representation for the data.

Figure (1) shows the good fitting of the exponential model to the data.



Figure(1): Fitting the exponential model for the I_k

The forecasting power model (25) by using equation (24) is U=0.005, This value supports obtained results.

REFERENCES:

1. **Prais, S.J and Aitchison, J.** The grouping of observations in regression analysis. The meeting of the econometric society, Innsbruck, August/September 1953, 1-15.
2. **Cramer, J.S.** Efficient grouping, regression and correlation in engel curve analysis, American Statistical Association Journal (JASA), March 1964, 233-249.
3. **Koutsoyiannis, A.** Theory of econometrics, 2nd edition, Chapter 12,258-293, Harper and Row Publishers, Inc., 1977.
4. **Makridakis, S.; Wheelwright, S.C. and McGee, V.E.,** Forecasting: methods and applications, 2nd, Chapter 2,16-62, John Wiley and Sons, New York, U.S.A.
5. **Myers, R.H.** Classical and modern regression with applications, Chapter 9, 300-324, Duxbury Press, Boston, U.S.A.