Adjusted and Unadjusted "R^2" - Further Evidence from Irish Data

John L. Pratschke*

Abstract

The author considers the use of adjusted and unadjusted coefficients of determination in the evaluation of different L. S. regression models of the Engel function applied to Irish data. He finds the adjustment to be appreciable in some cases of logarithmic functions and also for functions where the dependent variable is the expenditure proportion. The selection of the regression equation of best fit is significantly different depending on whether or not the coefficient of determination is adjusted.

Resumé

L'auteur considère l'emploie des coéfficients de détermination ajustés et non-ajustés dans l'estimation des divers types des régressions moindres carres de la fonction Engel quand appliqués a la matière irlandaise. Il trouve que l'ajustement est appréciable dans certains eas de fonctions logarithmiques et aussi, dans les fonctions ou la variable dépendante est la proportion de dépense. Le choix de l'equation de régression de meilleur ajustement différe d'une manière significative, selon que le coéfficient de determination est ajusté ou non.

In a recent application of Engel curve analysis to Irish data [2] , the problem arose of selecting a basis on which to select the best fitting of a number of different algebraic formulations of the Engel function. While the problem

* The author is deeply indebted to his colleague T. O'Connell for his assistance in preparing the data for this study.

Table 1:

Alternative Algebraic Forms of the Engel Function Fitted to Data of Five Major Expenditure Groups

Function Type	Function No.	Form of Engel Function
Linear	1.1	$v_i = \alpha_i + \beta_i E + \Upsilon_i n + \epsilon_i$
	1.2	$v_i = \alpha_i + \beta_i E + \Upsilon_i \log n + \epsilon_i$
	1.3	$v_i/n = \alpha_i + \beta_i (E/n) + \epsilon_i$
Semi-log	2.1	$v_i = \alpha_i + \beta_i \log E + \Upsilon_i n + \epsilon_i$
	2.2	$v_i = \alpha_i + \beta_i \log E + \Upsilon_i \log n + \epsilon_i$
	2.3	$v_i = \alpha_i + \beta_i \log (E/n) + \epsilon_i$
Double-log	3.1	$\log v_i = \alpha_i + \beta_i \log E + \Upsilon_i n + \epsilon_i$
	3.2	$\log v_i = \alpha_i + \beta_i \log E + \Upsilon_i \log n + \epsilon_i$
	3.3	$\log (v_i/n) = \alpha_i + \beta_i \log (E/n) + \epsilon_i$
Log-inverse	4.1	$\log v_i = \alpha_i + \beta_i /E + \Upsilon_i n + \epsilon_i$
	4.2	$\log v_i = \alpha_i + \beta_i /E + \Upsilon_i \log n + \epsilon_i$
	4.3	$\log (v_i/n) = \alpha_i + \beta_i (n/E) + \epsilon_i$
Linear in w_i	5.1	$w_i = \alpha_i + \beta_i E + \Upsilon_i n + \epsilon_i$
	5.2	$w_i = \alpha_i + \beta_i E + \Upsilon_i \log n + \epsilon_i$
Semi-log in w_i	6.1	$w_i = \alpha_i + \beta_i \log E + \Upsilon_i n + \epsilon_i$
	6.2	$w_i = \alpha_i + \beta_i \log E + \Upsilon_i \log n + \epsilon_i$
Leser	7.1	$w_i = \alpha_i + \beta_i \log E + \Upsilon_i /E + \delta_i n + \epsilon_i$
	7.2	$w_i = \alpha_i + \beta_i \log E + \Upsilon_i /E + \delta_i \log n + \epsilon_i$

Notes

v_i is the household expenditure in good i,

E is household total expenditure $(\Sigma_i v_i)$

n is the number of persons per household

w_i is the expenditure proportion v_i/E

has been recognised for some time, it is only with the advent of large scale computer-oriented econometric research that analysts have been able to use large numbers of function types and that the problem has become acute. Apart from questions regarding the application of classical probability theory to the evaluation of correlation coefficients from large numbers of alternative least-square regressions on one body of data, the question of developing criteria for goodness of fit is now coming to the forefront.

In his analysis, the writer fitted eighteen formulations of the Engel function to data for five commodity groups reported in an Irish household budget survey [1]. The function types are set out in Table 1 following.

(Table 1)

The regression estimates were based on a two-way classification of household average weekly expenditures, in which four classifications of household disposable weekly income and four classifications of household size were used. There were, therefore, sixteen observations for each regression.

The question of selecting the function of best fit then arose. Clearly, the coefficient of determination $\underline{R^2}$ is not strictly comparable as between functions: in (1.1) it is the percentage variance of v_i that is explained, while in (3.1) it is the percentage variance of $\underline{\log v_i}$ and in (6.1) it is the percentage variance of $\underline{w_i}$ that is explained by the regression, where $\underline{w_i}$ is the expenditure proportion $\underline{v_i /E}$. The same point is made both by Pratschke (op. cit) and, more recently, by Mahajan [3] .

Table 2: Comparison of Goodness of Fit of Alternative Forms of the Engel Function using R', R and R*

Function Type	Criterion	Linear			Semi-log			Double-log			Log-inverse			Linear in w_i		Semi-log in w_i		Leser	
Function No.		1.1	1.2	1.3	2.1	2.2	2.3	3.1	3.2	3.3	4.1	4.2	4.3	5.1	5.2	6.1	6.2	7.1	7.2
Food	R'	.985	.984	.750	.973	.965	.980	.978	.998	.982	.949	.959	.970	.973	.977	.989	.997	.979	.997
	R			.977			.969	.988	.998	.986	.979	.980	.930	.966	.970	.981	.996	.954	.975
	R*	.965	.942	.929	.895	.839	.777	.979	.987	.966	.972	.973	.878						
Clothing	R'	.983	.984	.983	.936	.935	.977	.977	.981	.975	.879	.875	.943	.990	.992	.988	.990	.991	.979
	R			.978			.952	.991	.991	.984	.958	.958	.951	.749	.779	.745	.769	.720	.750
	R*	.611	.609	.312	.466	.470	.365	.754	.781	.116	.188	.181	.466						
Fuel and Light	R'	.848	.847	.813	.837	.838	.728	.718	.718	.781	.648	.649	.633	.686	.597	.812	.812	.849	.849
	R			.872			.804	.887	.886	.882	.838	.838	.812	.865	.869	.936	.937	.957	.957
	R*	.956	.955	.464	.936	.940	.022	.955	.955	.493	.919	.921	.416						
Housing	R'	.965	.960	.949	.969	.972	.777	.981	.979	.951	.934	.931	.650	.984	.992	.990	.991	.987	.983
	R			.985			.904	.986	.990	.980	.944	.958	.895	.890	.939	.891	.939	.773	.848
	R*	.795	.799	.481	.609	.768	.109	.908	.944	.586	.621	.726	.191						
Sundries	R'	.998	.998	.991	.966	.969	.881	.994	.996	.986	.919	.909	.887	.996	.996	.998	.999	.999	.993
	R			.996			.922	.998	.999	.992	.963	.968	.937	.957	.955	.988	.995	.973	.996
	R*	.970	.956	.854	.698	.756	.016	.978	.984	.883	.686	.728	.234						

In an attempt to correct for this lack of comparability between $\underline{R}^2$'s, it was decided to calculate the values of $\underline{v_i}$ predicted by each regression form, corresponding to each of the sixteen pair of values of the independent variables $\underline{E}$, $\underline{n}$, and to correlate this predicted $\underline{v_i}$, which is styled $\underline{v_{i(c)}}$, with the observed values of $\underline{v_i}$. This adjusted coefficient of correlation is styled $\underline{R'}$. For forms (1.1), (1.2), (2.1), and (2.2), $\underline{R}$ = $\underline{R'}$.

An alternative adjusted correlation coefficient involves the correlation of $\underline{w_i}$ with the corresponding predicted values of the expenditure proportion, $\underline{w_{i(c)}}$, where $\underline{w_{i(c)}} = \underline{v_{i(c)}}/\underline{E}$. If this correlation coefficient of $\underline{w_i}$ on $\underline{w_{i(c)}}$ is styled $\underline{R^*}$, then $\underline{R^*}$ compares the different regression estimates of the Engel function in terms of the percentage variance of the expenditure proportion explained by the regression. Clearly, for forms (5.1), (5.2), (6.1), (6.2), (7.1) and (7.2), $\underline{R} = \underline{R^*}$.

(Table 2)

The actual values of $\underline{R}$, $\underline{R'}$, and $\underline{R^*}$ are given in Table 2. It is interesting to note the differences between $\underline{R}$ and the two adjusted $\underline{R'}$s, and, in particular, to see if one would have selected a different form of the Engel function using $\underline{R}$, $\underline{R'}$ or R* alone. Mahajan (op. cit) has reported on a similar experiment, using $\underline{R'}$s and $\underline{R'}$'s derived from Indian consumption data.

It is clear from the results that the adjusted correlation coefficient $\underline{R'}$ is substantially different from $\underline{R}$. The $\underline{R'}$ is less than $\underline{R}$ in the logarithmic forms (3.1),

Table 3: Comparison of Ranking of Functions of Best Fit Judged by R, R' and R*.

Function Type	Criterion	Linear			Semi-log			Double log			Log-inverse			Linear in w_i		Semi-log in w_i		Leser	
Function No.		1. 1	1. 2	1. 3	2. 1	2. 2	2. 3	3. 1	3. 2	3. 3	4. 1	4. 2	4. 3	5. 1	5. 2	6. 1	6. 2	7. 1	7. 2
Food	R	5	6	10	12	16	14	3	1	4	9	8	18	15	13	7	2	17	11
	R'	5	6	16	12	14	8	10	1	7	17	15	13	12	11	4	2+	9	2+
	R*	9	13	14	15	17	18	4	2	10+	7	6	16	10+	8	3	1	12	5
Clothing	R	5	3+	6	11	12	9	1+	1+	3+	7+	7+	10	16	13	17	14	18	15
	R'	7+	6	7+	14	15	16	11	9	12	17	18	13	3+	1	5	3+	2	10
	R*	9	10	15	12+	11	14	4	1	18	16	17	12+	6	2	7	3	8	5
Fuel and Light	R	11	12	8	16	13+	18	5	6	7	13+	13+	17	10	9	4	3	1+	1+
	R'	3	4	7	6	5	11	12+	12+	10	16	15	17	14	18	8+	8+	1+	1+
	R*	3	4+	16	9+	7	18	4+	4+	15	12	11	17	14	13	9+	8	1+	1+
Housing	R	7	8	3	6	5	13	2	1	4	10	9	14	16	11+	15	11+	18	17
	R'	11	12	14	10	9	17	7	8	13	15	16	18	5	1	3	2	4	6
	R*	9	8	16	14	11	18	4	3	15	13	12	17	6	1+	5	1+	10	7
Sundries	R	2+	2+	5	13	11	18	2+	1	7	14	12	17	15	16	8	6	10	9
	R'	3+	3+	11	14	13	18	9	6+	12	15	16	17	6+	6+	3+	1+	1+	10
	R*	7	9	12	15	13	18	5	4	11	16	14	17	8	10	2	1	6	3

Note:

+ Indicates that one or more functions have the same R or R' or R* and have accordingly been assigned the same rank number.

(3. 2), (3. 3), (4. 1), (4. 2), (4. 3), and notably greater than R for the forms where w_i is the dependent variable (namely (5. 1) through (7. 2)).

The coefficient R* is notably lower than either R or R' for clothing throughout all function forms, but is lower for all commodity groups in the homogeneous forms (1. 3), (2. 3), (3. 3) and (4. 3).

As regards the problem of selecting the form of best fit, the results are also interesting. In Table 3 following, the different function forms have been ranked for each commodity group, using each of R, R' , and R* in turn.

(Table 3)

Rank 1 is assigned to the function form with the highest R or R' or R*. Tied ranks are marked. The three function forms having the lowest rank (i. e. 1, 2, or 3) are italicised for easier reading. For food, function (3. 2) would have been selected using R or R' , but (6. 2) is the best fitting form judged by R*. For clothing the picture is more confused: any of (3. 1), (3. 2) or (5. 2) might have been selected as best fitting. On the other hand, for fuel and light, it is clear that forms (7. 1) and (7. 2) are the best fitting regardless of the criterion of selection used. The results are again different for housing and sundries, depending on which criterion is preferred. Clearly however, the selection made as to the best fitting function differs depending on whether R, R', or R* is used as the criterion.

This further reinforces Mahajan's conclusions (op. cit) in showing substantial differences between adjusted and unadjusted correlation coefficients, and, more important at the practical level, the fact that the selection of the best fitting function depends to a large extent on whether or not one adjusts the coefficient of determination, and if so, how.

The Economic and Social Research Institute
Dublin.

25 May 1970.

References

[1.] Central Statistics Office (1969). Household Budget Inquiry, 1965-1966, Dublin: Stationery Office.

[2.] Pratschke, J. L. (1969). Income-Expenditure Relation for Ireland, 1965-1966, Dublin: The Economic and Social Research Institute, Paper No. 50.

[3.] Mahajan, B. M. (1969), On Some Aspects of Unadjusted and Adjusted "R^2" Estimated from Indian Consumption Data. Review of the International Statistical Institute, 37:3, 294-302, The Hague.