Using Dummy Variables in Improving the Simple Linear Regression Model for the Ratio of Consumption to the National Income in Iraq during the Period (1970-1994)

Fedaa N. Abdulahad
Department of Mathematics and Computer Applications, Collage of Science,
Al Nahreen University, Baghdad- Iraq.
E-mail: fn5382@yahoo.com.

Abstract

This paper discuss the concept of dummy variables and its importance usge in statistical analysis by transforming the qualitative variables to measurable quantitative variables and applying it in analyzing the linear regression in both simple and multiple forms. A comparison has been made between using dummy variables and power transformation methodology. This comparison aims to show which one of the two methods is better in improving the linear regression model by applying them on the data of ratio of consumption to the national income in Iraq for the period of (1970-1994). Depending on the data available of that period the results showed that the dummy variables are more efficient than power transformation in improving the regression model of the consumption to national income. The dummy variables helped explaining almost 80% from the consumption ratio in the given period in Iraq by making the data to be more intelligible and more homogeneous in the model.

Keywords: dummy variables; simple linear regression; power transformation.

Introduction

Dummy variable regression (DVR) and analysis of covariance (ANCOVA) are methods of statistical analysis that combine analysis of variance and multiple regression. In ANOVA, there are several groups of scores defined by the levels on one or more factors, while in regression; each value of the response variable Y is associated with scores on one or more X variables (called covariates rather than predictors). ANCOVA is really a special case of DVR, with a certain specific goal [7]. One approach to regression models is simply to include dummy variables for all individuals in the sample. For linear models Poisson regression models [4] and negative binomial regression models [2], this method works very well. For logistic regression, on the other hand, the inclusion of dummy variables for many individuals can lead to severe "incidental parameters" bais [6].

When there are exactly two observations for each individuals, logistic regression coefficients will be twice as large as they should be [1]. The solution to the incidental parameters problem for regression is to do conditional maximum likelihood, conditioning on the number of 1's and 0's for each individual this removes the dummy variable coefficients from the likelihood function [5].

This paper discuss the consumption ratio in Iraq for the period (1970-1994) with respect to national income in this period and it showed that Iraq has been throw abnormal situations between the normal conditions before the wars and into the Gulf war followed by economical sanctions which made the data incoherent and made it incapable of explaining the consumption ratio in a simple linear regression model. To solve this problem the dummy variables are used by dividing the studied period into three parts (70's, 80's and 90's) this helped improving the model. A comparison has been made between using the dummy variables method and using the power transformation method where both aiming to improve the model by using the transformation on the response variable (consumption ratio).

The paper also viewed briefly the theoretical aspect of the dummy variables and its use in a linear regression model also the theoretical aspect of power transformation method. After that we use these two methods on the data to show which method represents the studied period. Draper & Smith method [9] has been used to estimate the power parameter

λ for the response variable to have a maximum fitting in the model.

1. The Principles of Dummy Variables, [8]:

A dummy variable is an artificial variable created to represent an attribute with two or more distinct categories.

Regression analysis treats all independent (X) variables in the analysis as numerical. Numerical variables are interval or ratio scale variables whose values are directly comparable e.g. '10 is twice as much 5' or '3 minus 1 equal 2' often, however, you might want to include an attribute or nominal scale variable such as 'product brand' or 'type of defect' in your study. Say you have three types of defects, numbered '1', '2' and '3'. In this case, '3 minus 1' doesn't mean anything… you can't subtracting defect 1 from defect 3. The numbers here are used to indicate or identify the levels of 'Defect Types' and do not having intrinsic meaning of their own. Dummy variables are created in this situation to 'trick' the regression algorithm into correctly analyzing variables.

2. The General Linear Model for Dummy Variables Regression DVR:

Dummy variables usually indicate the presence or observe of "quality" or an attribute, such as male or female, black or white. The general linear model for DVR includes terms for both the categorical variables of ANOVA and numerical variables of regression, [7].

There is a way to circumvent the dummy variable trap (multicollinearity) by introducing as many dummy variables as the number of categories of that variable provided we do not introduce the intercept in such a model

$$Y_i = \beta_1 D_{i1} + \beta_2 D_{i2} + \beta_3 D_{i3} + \varepsilon_i \quad \text{..... (1)}$$

where i=1,2,…,n, n represents number of the observations of the variables, D_{i1}, D_{i2} and D_{i3} are dummy variables, β_1, β_2, and β_3 are unknown coefficients and ε_i represents the variable of random error.

We note that, running regression (1), there is no intercept option in regression package, with the intercept suppressed, and allowing a dummy variables for each category, we obtain directly the mean values of the various categories, [6].

Also, we note that equation with an intercept is more convenient.

Whenever intercept is present, the estimated coefficients β_1, β_2, and β_3 attached to the various dummies D_{i1}, D_{i2} and D_{i3} are now differential intercepts, showing by how much the average value of (1) differs from that of the benchmark quarter. Put differently, the coefficients on the seasonal dummies will give the seasonal increase or decrease in the average value of Y relative to the base season, [7].

3. Power Transforming of The Response Variables:

In statistics, the power transform is a family of transformations that map data from one space to another using power functions. This is a useful for data processing technique used to reduce data variation, make the data more normal distribution-like, improve the correlation between variables and for other data stabilization procedures.

Power transformations are ubiquitously used in various fields. For example, multi-resolution and wavelet analysis, statistical data analysis, medical research, modeling of physical processes … etc, [3].

The Box-Cox transformation of the response variables Y is used to make the linear model more appropriate to the data. It can be used to attempt or impose linearity, reduce the skewness or stabilize the residual variance.

The Box-Cox transformation is defined as [9]:

$$\tau(Y;\lambda) = \begin{cases} \frac{Y^{\lambda}-1}{\lambda} & if \quad \lambda \neq 0 \\ \ln(Y) & if \quad \lambda = 0 \end{cases} \quad \text{..... (2)}$$

It is clear that $\tau(1;\lambda) = 0$ for any λ, and the derivative with respect to Y exist for any λ.

The Box-Cox transformation is a power transformation, but done in such a way as to make it continues with the parameter λ at λ=0. This transformation has proved popular in regression analysis, including econometrics. Box and Cox also proposed a more general from of the transformation which incorporates a shift parameter α.

$$\tau(Y;\lambda,\alpha)=\begin{cases}\frac{[(Y+\alpha)^{\lambda}-1]}{\lambda} & if \quad \lambda\neq 0\\ \ln(Y+\alpha) & if \quad \lambda=0\end{cases} \quad \text{..... (3)}$$

4. Draper & Smith Method for Estimating the Parameter λ of Box-Cox Transformation, [9]:

This method is considered to be one of the oldest and simplest procedures to estimate the power transformation parameter in regression models whether it is applied on the predictor variable or the response variable.

This method is suggested a certain domain to the value of λ is to be chosen from and most of the cases the range of domain is (-2,2) and this domain can be modified to meet the requirement of the experiment. The values of λ selected from the domain are applied on the regression model to minimize or maximize a certain values (like MSE and R^2 …etc) to be a measure of λ. The following are the steps of the procedure:

1. Select the value of λ from the assumed domain.
2. Transforming the response variable $y \rightarrow \tau(y)$ where $\tau(y)$ is defined in equation (2).
3. Estimating the regression model parameters $\tau(y_i,\lambda)=\alpha+\beta x_i+\varepsilon_i$; i=1,2,…,n
   where α, β are the parameters of regression model.
4. Estimating the value of decision measure.
5. Making the decision.

5. The Applications Data:

In this section we used data from the consumption ratio and the national income in Iraq for the period (1970-1994). This data is not ready for statistical analysis because the period from (1970-1994) was abnormal, during that period Iraq was in very hard situations due to several wars and economical sanctions. The mentioned period is divided into three stages the first from (1970-1980) Iraq was relatively in good conditions, the second (1980-1988) was the Gulf war and the third (1988-1994) included the second Gulf war followed by sever economical sanctions. Table (1) shows the data used in the analysis. These data has been obtained from the annual statistical groups issued from the central institution for statistics and information technology Baghdad- Iraq.

Table (1)
Consumption ratio and the national income in Iraq for the period (1970-1994).

Year	*i*	*The national income (dinar)*	*Consumption ratio*
1970	1	5171.6	0.9271
1971	2	5425.8	0.8981
1972	3	5582.9	0.8168
1973	4	8062.3	0.6022
1974	5	7585.4	0.7464
1975	6	9094.6	0.5896
1976	7	10633.3	0.6052
1977	8	10944.3	0.5762
1978	9	13625.8	0.6033
1979	10	16814.6	0.4386
1980	11	15246.0	0.3970
1981	12	8972.8	0.7625
1982	13	8301.7	1.0293
1983	14	8180.3	1.1783
1984	15	8654.3	1.0574
1985	16	8817.9	1.0044
1986	17	9746.2	1.1291
1987	18	13033.7	0.9871
1988	19	13287.9	0.9798
1989	20	11768.3	0.9798
1990	21	11521.1	0.9081
1991	22	2899.8	0.8976
1992	23	3387.8	0.9656
1993	24	2296.9	1.0464
1994	25	2839.6	0.9356

6.Analyzing the Relation of The Consumption Ratio and The National Income Without Using Dummy Variables:

A simple linear regression model has been used to represent the data that are given in table (1) which is given by the following

$$Y_i=\alpha+\beta X_i+\varepsilon_i \quad ; \quad i=1,2,\ldots,25$$

where α, β are the parameters of regression model, Y_i represents the ratio of consumption to the national income for the period (1970-1994), X_i represents the notional income for

the period (1970-1994) and ε_i represents the variable of random error.

The program Minitab is used to analyze these data and the results are shown below

The regression equation is

$Y_i = 1.07 - 0.000026\ X_i$ and $R^2 = 21.9\%$

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	0.24752	0.24752	6.47	0.018
Residual Error	23	0.88031	0.03827		
Total	24	1.12783			

We note that all statistical experiments indicates that the simple linear regression model is not suitable to represent the data and this is normal because the data is considered inconsistent for some stages of the mentioned period due to the hard conditions in Iraq. We can see that the value of R^2 is small and equal to (21.9), Thus the predictor variable (national income) explains (21.9%) from the total variance of the consumption, the calculated value F indicates that the model is not significant.

7. Analyzing Data Using Dummy Variables:

After we have shown that the simple linear regression model does not represent the relation between national income and the ratio of consumption in a logical way we can use dummy variables to apply regression analysis on the qualitative variables and these are 1970's, 1980's and 1990's and the data is divided according to these periods and represented by two dummy variables as follows:

$$(D_{i1}, D_{i2}) = \begin{cases} (0,1) & for \quad 1970's \\ (1,0) & for \quad 1980's \\ (0,0) & for \quad 1990's \end{cases}$$

Hence the regression model that describes the phenomena has become as follows:

$$Y_i = \alpha + \beta_1 X_i + \beta_2 D_{i1} + \beta_3 D_{i2} + \varepsilon_i$$

Where α, β, X_i, Y_i and ε_i defined similar to the previous and D_{i1}, D_{i2} represent the dummy variables.

Table (2) represents the data of national income, the ratio of consumption and the suggested dummy variables for the three stages.

Table (2)
Consumption ratio and the national income with the suggested dummy variables.

Year	*i*	*The national income (dinar)*	*Consumption ratio*	D_1	D_2
1970	1	5171.6	0.9271	0	1
1971	2	5425.8	0.8981	0	1
1972	3	5582.9	0.8168	0	1
1973	4	8062.3	0.6022	0	1
1974	5	7585.4	0.7464	0	1
1975	6	9094.6	0.5896	0	1
1976	7	10633.3	0.6052	0	1
1977	8	10944.3	0.5762	0	1
1978	9	13625.8	0.6033	0	1
1979	10	16814.6	0.4386	0	1
1980	11	15246.0	0.3970	0	1
1981	12	8972.8	0.7625	1	0
1982	13	8301.7	1.0293	1	0
1983	14	8180.3	1.1783	1	0
1984	15	8654.3	1.0574	1	0
1985	16	8817.9	1.0044	1	0
1986	17	9746.2	1.1291	1	0
1987	18	13033.7	0.9871	1	0
1988	19	13287.9	0.9798	1	0
1989	20	11768.3	0.9798	1	0
1990	21	11521.1	0.9081	0	0
1991	22	2899.8	0.8976	0	0
1992	23	3387.8	0.9656	0	0
1993	24	2296.9	1.0464	0	0
1994	25	2839.6	0.9356	0	0

By analyzing the data from table (2) using the multiple regression model we obtained the following results

The regression equation is

$Y_i = 1.08 - 0.000028\ X_i + 0.214\ D_{i1} - 0.150\ D_{i2}$

and $R^2 = 80.1\%$

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	3	0.90379	0.30126	28.24	0.000
Residual Error	21	0.22404	0.01067		
Total	24	1.12783			

From the results above it indicates that the multiple regression model represents the relation in a consistent way that R^2= 80% which means that the variables (predictor and dummies) explains 80.1% from the variable in the consumption year, the increment of R^2 came from considering the abnormal conditions in Iraq which are represented by the dummy variables.

8. Analyzing Data Using Power Transformation Method:

After we have used the dummy variables in improving the simple linear regression model which represents the relation between the consumption ratio and the national income in Iraq for the period (1970-1994) the power transformation will be used to obtain better results and compare them with the ones that we obtained by using dummy variables where the transformation will be made on the response variable y_i as shown in the following model:

$$\frac{Y^{\lambda}-1}{\lambda} = \alpha + \beta X_i + \varepsilon_i$$

The parameter λ is estimated by using Draper & Smith method which is introduced earlier in this paper and the possible value for R^2 obtained was (34.38) we got this value at λ equal to (-2.7) this shows that the best model we can have by using power transformation is the following:

$$\frac{Y^{-2.7}-1}{-2.7} = 0.7869 - 0.0002X_i$$

We can see from the value of R^2 obtained after improving the model that it did not increase significantly and it was not effective unlike the dummy variables this shows that the dummy variables are more efficient than the power transformation in improving the model of consumption ratio to the national income in Iraq during the studied period.

9. Conclusions and Recommendations:

1. The Iraqi national income for the period (1970-1994) does not give an accurate and logical explaining as a predictor variable for the consumption ratio for that period considering the consumption ratio as a response variable in the simple linear regression model because of the unstable situations in that period.
2. Using the dummy variables in dividing the period into three stages helped improving the regression model significantly which made the notional income explains almost 80% of the consumption ratio in that period.
3. Using the power transformation in improving the model did not help a lot in explaining a larger ratio of the results.
4. The results showed that using dummy variables is more efficient than power transformation in improving the linear regression model of the consumption ratio to the national income in Iraq for the period (1970-1994).

References

[1] Abrevaya, Jason (1997), "The Equivalence of Two Estimators of the Fixed- Effects Logit Models", Economics Letters 55:41-43.

[٢] Allison, Paul D. and Richard P. Waterman (2002), "Fixed-Effects Negative Binomial Regression Models", Forthcoming in Sociological Methodology 2002.

[3] Box, G.E.P. and Cox, D.R. (1964), "An Analysis of Transformation", with discussion, J.R. Statistical Society, Ser. B, 26, 11-252.

[4] Cameron, A. Colin and Pravin K. Trivedi. (1998), "Regression Analysis of Count Data", Cambridge, UK, Cambridge University Press.

[5] Chamberlain, Gary A. (1980), "Analysis of Covariance with Qualitative Data", Review of Economic Studies 47: 225-238.

[6] Kalbfleisch, Jhon D. and David A. Sprott (1970), "Applications of Likelihood Methods to Models Involving Large Numbers of Parameters" (with discussion), Journal of the Royal Statistical Society Series B, 32:175-208.

[7] Miller, Jeff and Haden, Patricia (2006), "Statistical Analysis with the General Linear Model", Creative Commons Attribution, USA.

[8] Skrivanek, Smita (2009), "The Use of Dummy Variables in Regression Analysis", More Steam, LLC.

[9] آفان سفر الصفار (٢٠٠٨)، "استخدام تحويلات القوى في منهجية سطوح الاستجابة دراسة تطبيقية لتعظيم القدرة التنبؤية للتجارب الحياتية"، رسالة ماجستير/ قسم الاحصاء والمعلوماتية/ جامعة الموصل.

الخلاصة

تم في هذا البحث عرض مفهوم المتغيرات الصماء واهمية استخدامها في التحليل الاحصائي وذلك بتحويل المتغيرات النوعية الى متغيرات كمية قابلة للقياس وتطبيق هذه المتغيرات في تحليل الانحدار بشقيه الخطي البسيط والمتعدد. تم اجراء مقارنة بين استخدام المتغيرات الصماء مع استخدام منهجية تحويلات القوى هذه المقارنة تهدف الى بيان اي من الطريقتين المذكورتين الاحسن في تحسين نموذج الانحدار الخطي وذللك بتطبيق الطريقتين على بيانات نسبة الاستهلاك الى الدخل القومي في العراق للفترة (١٩٧٠-١٩٩٤). ان البيانات المتاحة في الفترة المذكورة بينت المتغيرات الصماء هي الاكثر كفاءة من تحويلات القوى في تحسين نموذج الانحدار لنسبة الاستهلاك الى الدخل القومي. ان المتغيرات الصماء ساعدت في تفسير ٨٠% من نسبة الاستهلاك في الفترة المذكورة في العراق ذلك انها جعلت البيانات اكثر وضوحا وتجانسا في النموذج.