Binary variables in linear regression [*]

Jean-Marie Dufour [†]
McGill University

First version: October 1979
Revised: April 2002, July 2011, December 2011
This version: December 2011
Compiled: December 8, 2011, 14:42

[*]This work was supported by the William Dow Chair in Political Economy (McGill University), the Bank of Canada (Research Fellowship), a Guggenheim Fellowship, a Konrad-Adenauer Fellowship (Alexander-von-Humboldt Foundation, Germany), the Canadian Network of Centres of Excellence [program on *Mathematics of Information Technology and Complex Systems* (MITACS)], the Natural Sciences and Engineering Research Council of Canada, the Social Sciences and Humanities Research Council of Canada, and the Fonds de recherche sur la société et la culture (Québec).

[†] William Dow Professor of Economics, McGill University, Centre interuniversitaire de recherche en analyse des organisations (CIRANO), and Centre interuniversitaire de recherche en économie quantitative (CIREQ). Mailing address: Department of Economics, McGill University, Leacock Building, Room 519, 855 Sherbrooke Street West, Montréal, Québec H3A 2T7, Canada. TEL: (1) 514 398 8879; FAX: (1) 514 398 4938; e-mail: jean-marie.dufour@mcgill.ca . Web page: http://www.jeanmariedufour.com

Contents

1. Notion of binary variable

Suppose we wish to estimate a consumption function

$$C_t = \alpha + \beta Y_t + \varepsilon_t\,,\ t = 1, \ldots, T, \tag{1.1}$$

where the errors ε_t satisfy the assumptions of the classical linear model, but we have reasons to think that the constant is not the same during war years (as opposed to normal or peace years):

$$C_t = \alpha_1 + \beta Y_t + \varepsilon_t \ \ ,\ \text{if } t \text{ is a normal year}, \tag{1.2}$$

$$C_t = \alpha_2 + \beta Y_t + \varepsilon_t \ \ ,\ \text{if } t \text{ is a war year}. \tag{1.3}$$

We maintain the assumption that the marginal propensity to consume β is the same for all observations (although this could also be relaxed).

We can then estimate β, α_1 and α_2 by considering the following linear regression:

$$\begin{aligned} C_t &= \alpha_1 + (\alpha_2 - \alpha_1) D_t + \beta Y_t + \varepsilon_t \\ &= \alpha_1 + \delta D_t + \beta Y_t + \varepsilon_t,\ t = 1, \ldots, T, \end{aligned} \tag{1.4}$$

where

$$D_t = \begin{cases} 1, \text{ if } t \text{ is a war year}, \\ 0, \ \text{otherwise}. \end{cases} \tag{1.5}$$

We call D_t a "binary variable".

An equivalent way to proceed consists in considering the regression:

$$C_t = \alpha_1 D_{1t} + \alpha_2 D_{2t} + \beta Y_t + \varepsilon_t\,,\ t = 1, \ldots, T, \tag{1.6}$$

where

$$D_{1t} = \begin{cases} 1, \text{ if } t \text{ is a normal year}, \\ 0, \ \text{otherwise}, \end{cases} \tag{1.7}$$

$$D_{2t} = 1 - D_{1t} = \begin{cases} 1, \text{ if } t \text{ is a war year}, \\ 0, \ \text{otherwise}. \end{cases} \tag{1.8}$$

An advantage of the second approach comes from the fact that the linear regression directly yields the values of α_1 and α_2 (and their standard errors).

2. Seasonal dummy variables

Another important use of dummy variables consists in taking into account seasonal variation. For example, consumption C_t may depend on income Y_t and the season (first, second, third or fourth quarter):

$$C_t = \alpha + \beta Y_t + \lambda_1 D_{1t} + \lambda_2 D_{2t} + \lambda_3 D_{3t} + \varepsilon_t,\ t = 1, \ldots, T, \tag{2.1}$$

where

$$D_{1t} = \begin{cases} 1, \text{ if } t \text{ is a first quarter} \\ 0, \text{ otherwise} \end{cases}, \tag{2.2}$$

$$D_{2t} = \begin{cases} 1, \text{ if } t \text{ is a second quarter} \\ 0 \text{ otherwise} \end{cases}, \tag{2.3}$$

$$D_{3t} = \begin{cases} 1, \text{ if } t \text{ is a third quarter} \\ 0 \text{ otherwise} \end{cases}. \tag{2.4}$$

Equivalently, we can consider the regression:

$$C_t = \beta Y_t + \lambda_1 D_{1t} + \lambda_2 D_{2t} + \lambda_3 D_{3t} + \lambda_4 D_{4t} + \varepsilon_t, \; t = 1, \ldots, T, \tag{2.5}$$

where

$$D_{4t} = 1 - D_{1t} - D_{2t} - D_{3t}\,. \tag{2.6}$$

However, if we tried to estimate the model

$$C_t = \alpha + \beta Y_t + \lambda_1 D_{1t} + \lambda_2 D_{2t} + \lambda_3 D_{3t} + \lambda_4 D_{4t} + \varepsilon_t, \; t = 1, \ldots, T, \tag{2.7}$$

the matrix $X'X$ would not be invertible (exact multicollinearity). So this should be avoided.

3. Qualitative explanatory variables

Another use of binary variables consists in representing "qualitative variables". For example, the consumption C_i of a product by individual i may depend on the income Y_i (of the individual) and sex S_i:

$$C_i = \alpha_0 + \alpha_1 S_i + \alpha_2 Y_1 + \varepsilon_i\,, \; i = 1, \ldots, n,$$

where

$$S_i = \begin{cases} 1, \text{ if } i \text{ is a woman,} \\ 0, \text{ otherwise,} \end{cases} \tag{3.1}$$

Equivalently, we can also write:

$$C_i = \beta_1 S_{1i} + \beta_2 S_{2i} + \alpha_2 Y_i + \varepsilon_i\,, \; i = 1, \ldots, n, \tag{3.2}$$

where

$$S_{1i} = S_i\,, \quad S_{2i} = 1 - S_{1i}\,, \tag{3.3}$$

so that

$$\begin{aligned} \mathsf{E}(C_i) &= \beta_1 + \alpha_2 Y_i = (\alpha_0 + \alpha_1) + \alpha_2 Y_i\,, && \text{if } S_i = 1\,, \\ \mathsf{E}(C_i) &= \beta_2 + \alpha_2 Y_i = \alpha_0 + \alpha_2 Y_i\,, && \text{if } S_i = 0\,. \end{aligned} \tag{3.4}$$

One can also include several binary variables which represent different characteristics. For example, consumption may be a function of income Y_i, sex (M or F) and age (less than 25 years, between 25 and 50, more than 50 years):

$$C_i = \alpha + \beta Y_i + \gamma_1 S_{1i} + \gamma_2 A_{1i} + \gamma_3 A_{2i} + \varepsilon_i\,,\; i = 1, \ldots, n, \tag{3.5}$$

where

$$S_{1i} = \begin{cases} 1, \text{ if } i \text{ has sex M,} \\ 0, \text{ otherwise,} \end{cases} \tag{3.6}$$

$$A_{1i} = \begin{cases} 1, \text{ if } i \text{ is less than 25 years old,} \\ 0, \text{ otherwise,} \end{cases} \tag{3.7}$$

$$A_{2i} = \begin{cases} 1, \text{ if } i \text{ has age between 25 and 50 years,} \\ 0, \text{ otherwise.} \end{cases} . \tag{3.8}$$

If a constant is included in the regression, one must leave out binary variable for each characteristic.

4. Bibliographic notes

Dummy variables can also be used to compute predictions and prediction errors, as well as to perform tests for structural change; see Dufour (1980, 1981, 1982*a*, 1982*b*) For more details on binary variables in econometrics, the reader may consult Maddala (1977) and Johnston (1984).

References

Dufour, J.-M. (1980), ‘Dummy variables and predictive tests for structural change’, *Economics Letters* **6**, 241–247.

Dufour, J.-M. (1981), ‘Variables binaires et tests prédictifs contre les changements structurels: une application á l’équation de St.-Louis’, *L’Actualité économique* **57**, 376–385.

Dufour, J.-M. (1982*a*), ‘Generalized Chow tests for structural change: A coordinate-free approach’, *International Economic Review* **23**, 565–575.

Dufour, J.-M. (1982*b*), ‘Recursive stability analysis of linear regression relationships: An exploratory methodology’, *Journal of Econometrics* **19**, 31–76.

Johnston, J. (1984), *Econometric Methods*, third edn, McGraw-Hill, New York.

Maddala, G. S. (1977), *Econometrics*, McGraw-Hill, New York.