Some Consequences of Including Impulse-Indicator Dummy Variables in Econometric Models

David E. Giles

Department of Economics
University of Victoria, Canada

August, 2020

Abstract

Suppose that a regression model includes a regressor that is a dummy variable that takes a non-zero value for *only one* observation. Then the least squares estimates of the coefficients of the other regressors are the same as would be obtained by dropping that observation from the sample and omitting the dummy variable. This is well-known, but is frequently overlooked by practitioners. In this note we extend this result to the case of instrumental variables estimation, and to models for both count data and duration data. These extensions also allow for the inclusion of many such "impulse-indicator" variables in the model, not just one.

Keywords Dummy variables; impulse-indicator variables; instrumental variables; count data; duration data

JEL Classifications C20; C25; C26; C41

Acknowledgement

I am very grateful to Ryan Godwin and Jacob Schwartz for helpful discussions relating to an earlier version of this paper.

Author Contact:
David E. Giles, 58 Rock Lake Court, L'Amable, ON, K0L 2L0, CANADA
e-mail: dgiles@uvic.ca; Phone: +1-613-332-6833

1. Introduction

Dummy variables are used frequently as regressors in regression analysis. These binary variables take a value of unity (or zero) if the corresponding sample observation is (is not) associated with some qualitative attribute. Following Hendry and Santos (2005), a dummy regressor takes its non-zero value for only a *single sample observation* will be referred to as an "impulse-indictor" regressor. As is well known (*e.g.*, Salkever, 1976) the OLS estimates of the coefficients of the other regressors, and their standard errors, are identical to those obtained if an impulse-indicator variable is omitted from the model, and the observation associated with its unitary value is deleted from the sample for all of the model's variables. In addition, the OLS residual corresponding to the non-zero value of the impulse-indicator variable is identically zero.

These results emerge essentially from the geometry of the least squares set-up. Surprisingly, there appears to have been no discussion of this problem when other common regression estimators are used. In this paper we extend the principal results of Salkever (1976), for the case where there may be several such impulse-indicator regressors, to cover a wide range of other models and estimator choices. In particular, we show that similar results hold for the family of instrumental variables (IV) estimators; and for the maximum likelihood (ML) estimators for various count data and duration (survival) data models

It should be emphasized that our results are purely algebraic in nature, and do not speak to the statistical properties of the estimators of the coefficients in the model. These properties have been discussed to some extent elsewhere. For example, Hendry and Santos (2005) show that, among other things, the OLS estimator of the coefficient of an impulse-indicator regressor is inconsistent under standard assumptions, although the OLS estimator for the remaining coefficient vector is consistent. Giles (2017) shows that this also holds in the case of stochastic regressors and IV estimation.

The rest of the paper is structured as follows. In the next section we provide the basic model and associated assumptions. Section 3 deals with our new results for the family of IV estimators; and section 4 presents our results for ML estimation in the context of count data models. Some models for duration data are discussed in section 5; and section 6 concludes.

2. Model and assumptions

In the next section we will be concerned with the following linear regression model:

$$y_{(n+r)} = X_{(n+r)}\beta + \varepsilon \quad ; \qquad \varepsilon \sim [0, \sigma^2 I] \quad , \tag{1}$$

where $X_{(n+r)}$ is $((n+r)\times k)$, of full rank, and includes an intercept variable as one of its columns. Typically, $X_{(n+r)}$ will be random, thus motivating the use of an IV estimator to estimate β, but the details of the stochastic nature of $X_{(n+r)}$ are not relevant to our algebraic results.

Now, consider a set of *r* zero-one dummy variables. As the estimators that we are considering are invariant to the ordering of the data in the sample, suppose that the i^{th} impulse-indicator variable takes its sole non-zero value in observation (n + i), for i = 1, 2,…., r; and let Z be the $(r \times k)$ matrix for the additional observations on the non-dummy regressors.

Then, let Q be the $((n+r)\times(k+r))$ matrix defined as:

$$Q = \begin{pmatrix} X & 0 \\ Z & I \end{pmatrix} , \tag{2}$$

where X is $(n \times k)$ and corresponds to the first n rows of $X_{(n+r)}$; Z is $(r \times k)$ and corresponds to the last r rows of $X_{(n+r)}$; 0 is an $(n \times r)$ null matrix; and I is an identity matrix of order r. Partition $y_{(n+r)}$ conformably:

$$q = \begin{pmatrix} y \\ z \end{pmatrix} , \tag{3}$$

so y is $(n \times 1)$, and z is $(r \times 1)$. Finally, define $\alpha = (\beta , \gamma)'$, say, where γ is $(r \times 1)$.

We will be concerned with the estimation of two forms of our model:

$$\text{M}_1: \qquad q = Q\alpha + v \qquad ; \tag{4}$$

and

M_2: $y = X\beta + u$. (5)

In M_1 the model includes the r impulse-indicator variables and is to be estimated using all $(n + r)$ observations, while in M_2 the impulse-indicator variables and the last r observations are omitted. In the next section we show that the IV estimator of β is the same in both models.

3. IV estimation

We will consider the case where there are g ($\geq k$) instruments for the regressors in X, and the r impulse-indicator variables are also included in the instrument set. The IV estimator of α in (4) is

$$\tilde{\alpha} = (Q'MQ)^{-1}Q'Mq, \tag{5}$$

where $M = W(W'W)^{-1}W'$, and W is the $((n+r)\times(g+r))$ instrument matrix:

$$W = \begin{pmatrix} X^{**} & 0 \\ Z^{**} & I \end{pmatrix}, \tag{6}$$

and the columns of $\begin{pmatrix} X^{**} \\ Z^{**} \end{pmatrix}$ are the g instruments.

Using the standard partitioned inverse formula,

$$(W'W)^{-1} = \begin{pmatrix} (X^{**\prime}X^{**})^{-1} & -(X^{**\prime}X^{**})^{-1}Z^{**\prime} \\ -Z^{**}(X^{**\prime}X^{**})^{-1} & I + Z^{**}(X^{**\prime}X^{**})^{-1}Z^{**\prime} \end{pmatrix}, \tag{7}$$

and

$$M = \begin{pmatrix} X^{**}(X^{**\prime}X^{**})^{-1}X^{**\prime} & 0 \\ 0 & I \end{pmatrix}. \tag{8}$$

So, from (8) and (2),

$$Q'MQ = \begin{pmatrix} X'M^{**}X + Z'Z & Z' \\ Z & I \end{pmatrix}, \tag{9}$$

and

$$Q'Mq = \begin{pmatrix} X'M^{**}X + Z'y \\ z \end{pmatrix} \quad , \tag{10}$$

where $M^{**} = X^{**}(X^{**\prime}X^{**})^{-1}X^{**\prime}$.

Using the result that

$$(Q'MQ)^{-1} = \begin{pmatrix} (X'M^{**}X)^{-1} & -(X'M^{**}X)^{-1}Z' \\ -Z(X'M^{**}X)^{-1} & I + Z(X'M^{**}X)^{-1}Z' \end{pmatrix} \quad , \tag{11}$$

we have:

$$\tilde{\alpha} = \begin{pmatrix} \tilde{\beta} \\ \tilde{\gamma} \end{pmatrix} = \begin{pmatrix} (X'M^{**}X)^{-1}X'M^{**}y \\ Z'z - Z(X^{*}X)^{-1}X'M^{**}y \end{pmatrix} \quad , \tag{12}$$

and so $\tilde{\beta}$ is identical to the IV estimator of β in (5), with X^{**} as the instrument matrix.

4. Models for count data

In this section we consider some distributions that are commonly used for modeling count data – that is, data whose values are non-negative integers. Typically, covariates are introduced through the conditional mean of the underlying distribution, and the parameters are estimated by maximum likelihood (ML).

It will be helpful to use the following notation. Let y_i be the i^{th} observation on the variable to be modeled; let x_i' be the $(1\times k)$ vector for the i^{th} observation on the k non-dummy covariates; and let β be the corresponding $(k\times 1)$ coefficient vector. As before, there are r impulse-indicator dummy variables, each taking the value unity for only one (different) observation, $(n + i)$, for $i = 1, 2, \ldots, r$. Let d_i' be the $(1\times r)$ vector for the i^{th} observation on the r impulse-indicator variables. Note that d_i' will be a null vector for all $i \le n$. The $(r\times 1)$ vector, γ, contains the coefficients of the impulse-indicator variables. We will use the notation $f(y_i;\theta,x_i)$ to denote the p.m.f. for y_i, conditional on x_i and some $(m\times 1)$ parameter vector, $\theta = (\beta,\gamma,\phi)'$, where $m \ge (k + r)$. The sub-vector, ϕ, which may be null, allows for additional parameters (such as a scale parameter) in the p.m.f. for y_i.

We begin by considering the p.m.f. for a Poisson-distributed random variable, *Y*:

$$f(Y = y_i; \lambda) = \lambda^{y_i} e^{-\lambda} / y_i! \qquad ; \quad y_i = 0, 1, 2, \ldots\ldots \quad ; \ \lambda > 0 \,. \tag{13}$$

In this model $E(Y) = \lambda$. So, to ensure the positivity of the mean, a natural way to introduce covariates (including the impulse-indicator variables) into the model is to assign $\lambda = \lambda_i = \exp(x_i'\beta + d_i'\gamma)$.

For a sample of $(n + r)$ i.i.d. observations, the log-likelihood function is

$$\begin{aligned} l(\beta; y, x) &= \sum_{i=1}^{n+r} [-\exp(x_i'\beta + d_i'\gamma) + y_i(x_i'\beta + d_i'\gamma) - ln(y_i!)] \\ &= \sum_{i=1}^{n} [-\exp(x_i'\beta) + y_i x_i'\beta] + \sum_{j=n+1}^{n+r} [-\exp(x_j'\beta + \gamma_j) + y_j(x_j'\beta + \gamma_j)] + c \\ &= l_n + l_r + c \,, \end{aligned} \tag{14}$$

where $c = -\sum_{i=1}^{n+r} ln(y_i!)$ is a constant that can be ignored for the purposes of maximization.

So,

$$\frac{\partial l}{\partial \gamma_j} = [y_j - \exp(x_j'\beta + \gamma_j)] \quad ; \qquad j = n + 1, \ldots., n + r \tag{15}$$

and

$$\frac{\partial l}{\partial \beta} = \frac{\partial l_n}{\partial \beta} - \sum_{j=n+1}^{n+r} [y_j x_j - \exp(x_j'\beta + \gamma_j) x_j] \,. \tag{16}$$

When every equation in (15) is zero, as required for maximization, we see that the second term on the RHS of (16) will also be zero. So, the ML estimator for β, based on the full $(n + r)$ observations and a model that includes the impulse-indicator variables, is identical to the ML estimator for β, based on the model that excludes the impulse-indicator variables and the last *r* observations.

As is well known, a major weakness of the Poisson regression model is that it holds only for the equi-dispersed case, and often more general models are needed to allow for over-dispersion in the

data. In this respect, the most popular econometric model is the so-called "NegBin2 model" (Cameron and Trivedi, 1986), whose p.m.f. is

$$f(Y = y_i;\ \theta, \lambda_i) = \Gamma(\theta + y_i)\ r_i^{\theta}\ (1+r_i)^{y_i}\ ] \ /\ [\Gamma(1 + y_i)\Gamma(\theta)] \quad ;\quad y_i = 0, 1, 2, \ldots\ldots \quad ;\ \theta > 0$$

$$r_i = \theta / (\theta + \lambda_i)$$

.

Again, $E(Y) = \lambda_i$, and so it is natural to introduce covariates (including the impulse-indicator variables) into the model by assigning $\lambda_i = \exp(x_i'\beta + d_i'\gamma)$.

Setting up the log-likelihood function in this case, and proceeding in exactly the same way as we did above for the Poisson model, it is easy to establish once again that the ML estimator for β, based on the full $(n + r)$ observations and a model that includes the impulse-indicator variables, is identical to the ML estimator for β, based on the model that excludes the impulse-indicator variables and the last r observations.

It is also straightforward to show that the same result holds for the ML estimator in the more general family of Negbin-P models suggested by Cameron and Trivedi (1986), and discussed by Greene (2008). This includes the so-called Negbin1 model as a special case.

5. Parametric models for "survival" ("duration") data

"Survival" (or "duration") data are encountered frequently in empirical economics. For example, see Lancaster (1992), *inter alia*. Such data can be modelled by non-parametric, semi-parametric methods, or parametric methods. The first of these is exemplified by the product-limit estimator of Kaplan and Meier (1958); while the most popular semi-parametric method is the proportional hazard model of Cox (1972). Only fully parametric methods that incorporate covariates lend themselves to the discussion in this paper, so we focus on the so-called "accelerated failure time" (AFT) model. The following discussion assumes uncensored data, but the results still hold in the (likely) case of censoring.

The simplest AFT model, with a constant hazard rate, is the one in which the survival times follow an exponential distribution. The p.d.f. for an exponentially distributed random variable, Y, is:

$$f(y_i;\phi) = \frac{1}{\phi} exp(-y_i/\phi) \qquad ; \; y_i > 0 \qquad ; \; \phi > 0 \tag{17}$$

and $E(Y) = \phi$, so to allow for covariates in the model we set $\phi = \phi_i = \exp(x_i'\beta + d_i'\gamma)$. Based on a sample of $(n + r)$ i.i.d. observations, the log-likelihood function is

$$\begin{aligned} l(\beta; y, x) &= -\sum_{i=1}^{n+r}[(x_i'\beta + d_i'\gamma) + y_i / \exp(x_i'\beta + d_i'\gamma)] \\ &= -\sum_{i=1}^{n}[(x_i'\beta) + y_i / \exp(x_i'\beta)] - \sum_{j=n+1}^{n+r}[(x_j'\beta + \gamma_j) + y_j / \exp(x_j'\beta + \gamma_j)] \\ &= l_n + l_r \; . \end{aligned} \tag{18}$$

Now,

$$\frac{\partial l}{\partial \gamma_j} = [y_j / \exp(x_j'\beta + \gamma_j)] - 1 \quad ; \qquad j = n + 1, \ldots., n + r \tag{19}$$

and

$$\frac{\partial l}{\partial \beta} = \frac{\partial l_n}{\partial \beta} - \sum_{j=n+1}^{n+r}\{[y_j / \exp(x_j'\beta + \gamma_j)] - 1\}x_j \; . \tag{20}$$

To maximize the log-likelihood function, the expressions in (19) and (20) are equated to zero, and solved simultaneously for the unknown parameters. When (every equation in) (19) is zero, the expression in (20) becomes

$$\frac{\partial l}{\partial \beta} = \frac{\partial l_n}{\partial \beta} \quad ,$$

and we see immediately that the ML estimator for β, based on the full $(n + r)$ observations and a model that includes the impulse-indicator variables, is identical to the ML estimator for β, based on the model that excludes the impulse-indicator variables as well as the last r observations.

By the same argument, this result can readily be shown to hold if the survival times follow other standard distributions that are used in the AFT literature. These include, for example, the Weibull, gamma, log-normal, and log-logistic distributions.

6. Conclusions

In this paper we have established an important consequence of including "impulse-indicator" variables in a variety of econometric models, and under several different estimation methods. Specifically, the estimates of the coefficients of the model's other explanatory variables are the same as we obtain by if we dropping the "impulse observations" from the sample, and omit the "impulse-indicator" variables themselves from the model. Our results generalize a standard result, which is for one such variable in the context of least squares estimation of a linear regression model, in several important directions.

Our findings are important in the context of general-to-specific model selection based on "impulse-indicator" variables (*e.g.*, Hendry and Santos, 2005; Santos *et al.*, 2008). Specifically, this type of model selection can be applied in situations that are far more general than least squares linear regression.

Declaration

This research was not funded. There are no conflicts of interest/competing interest.

References

Cameron, A. C. and P. K. Trivedi, 1986. Econometric models based on count data: Comparisons and applications of some estimators and tests. *Journal of Applied Econometrics*, 1, 29–54.

Cameron, A. C. and P. K. Trivedi, 1998. *Regression Analysis of Count Data*. Cambridge University Press, New York.

Cameron, A. C. and P. K. Trivedi, 2005. *Microeconometrics: Methods and Applications*. Cambridge University Press, Cambridge.

Cox, D. R., 1972. Regression models and life tables (with discussion). *Journal of the Royal Statistical Society, B,* 34, 187–220.

Giles, D. E., 2017. On the inconsistency of instrumental variables estimators for the coefficients of certain dummy variables. *Journal of Quantitative Economics*, 15, 15-26.

Greene, W., 2008, Functional forms for the negative binomial model for count data. *Economics Letters*, 99, 585-590.

Hendry, D. F. and C. Santos, 2005. Regression models with data-based indicator variables. *Oxford Bulletin of Economics and Statistics*, 67, 571-595.

Kaplan, E. L. and P. Meier, 1958. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53, 457–481.

Lancaster, T., 1992. *The Econometric Analysis of Transition Data*. Cambridge University Press, Cambridge.

Salkever, D. S., 1976. The use of dummy variables to compute predictions, prediction errors, and confidence intervals. *Journal of Econometrics*, 4, 393-397.

Santos, C., D. F. Hendry, and S. Johansen, 2008. Automatic selection of indicators in a fully saturated regression. *Computational Statistics*, 23, 317-335.