LECTURE 4

Multiple Regression 2

Dummy Variables and Categorical Data

The era of modern econometrics began shortly after the war at a time when there was a paucity of reliable economic data. The data consisted mostly of annual observations; and the number of years spanned by the usable data were few. Nowadays, we benefit from data which is collected on a quarterly and, sometimes, on a monthly basis.

In spite of the belief that the structure of the economy and the behaviour of its agents had changed under the impact of the war, it seemed imperative to the pioneer investigators to include wartime data and prewar data in their series. Therefore it was often necessary to incorporate in the econometric equations devices which were designed to accommodate changes in structure. The simplest of such devices entails a dummy variable which enables one to calculate an intercept term which takes different values in different epochs.

Let us imagine, for the sake of simplicity, that we are concerned only with the difference between wartime and the succeeding peacetime. Then our structural equation might take the form of

$$y_t = d_{t1}\gamma_1 + d_{t2}\gamma_2 + x_{t1}\beta_1 + \cdots + x_{tk}\beta_k + \varepsilon_t. \tag{185}$$

Here we are replacing the usual intercept term β_0 by the term $d_{t1}\gamma_1 + d_{t2}\gamma_2$ wherein d_1 and d_2 are the so-called dummy variables whose values are specified by

$$d_{t1} = \begin{cases} 1, & \text{if } t \in \text{ Wartime;} \\ 0, & \text{if } t \in \text{ Peacetime,} \end{cases} \qquad d_{t2} = \begin{cases} 0, & \text{if } t \in \text{ Wartime;} \\ 1, & \text{if } t \in \text{ Peacetime.} \end{cases} \tag{186}$$

To understand how this scheme is represented in terms of the matrix notation, let us recall that the intercept term β_0 is accompanied in equation (103) by a vector $i = [1, \ldots, 1]'$ of T units. In place of this unit vector, we now have two

vectors which constitute the following matrix:

$$\begin{bmatrix} 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ \cdots & \cdots \\ 0 & 1 \\ \vdots & \vdots \\ 0 & 1 \end{bmatrix}. \tag{187}$$

The submatrix above the horizontal dots corresponds to wartime whilst the submatrix below the dots corresponds to peacetime.

There is an alternative but equivalent way of constructing this mechanism which gives rise to the equation

$$\begin{aligned} y_t &= \mu + d_{t2}\delta + x_{t1}\beta_1 + \cdots + x_{tk}\beta_k + \varepsilon_t, \\ &\text{where} \quad \mu = \gamma_1 \quad \text{and} \quad \mu + \delta = \gamma_2. \end{aligned} \tag{188}$$

This scheme is associated with the following vectors of dummy variables:

$$\begin{bmatrix} 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ \cdots & \cdots \\ 1 & 1 \\ \vdots & \vdots \\ 1 & 1 \end{bmatrix}. \tag{189}$$

We can conceive of yet another scheme in which there is a constant intercept term κ together with two parameters γ_1 and γ_2 which are intended to reflect the peculiar circumstances of the two epochs. In that case, the structural equation would be of the form

$$y_t = \kappa + d_{t1}\gamma_1 + d_{t2}\gamma_2 + x_{t1}\beta_1 + \cdots + x_{tk}\beta_k + \varepsilon_t, \tag{190}$$

whilst the associated vectors of dummy variables would be

$$\begin{bmatrix} 1 & 1 & 0 \\ \vdots & \vdots & \vdots \\ 1 & 1 & 0 \\ \cdots & \cdots & \cdots \\ 1 & 0 & 1 \\ \vdots & \vdots & \vdots \\ 1 & 0 & 1 \end{bmatrix}. \tag{191}$$

The summary notation for this matrix is $[i, d_1, d_2]$.

A problem besetting this formulation is immediately apparent, for there is now an exact linear relationship between the three columns of dummy variables such that $i = d_1 + d_2$. This feature conflicts with a necessary condition for the practicability of linear regression which is that the columns of the data matrix X must be linearly independent to ensure that the inverse matrix $(X'X)^{-1}$ exists.

It is instructive to investigate the consequence of forming the matrix $X'X$ from the three columns in (191) and attempting to invert it via the *Matrix-Invert* function of *Lotus 1-2-3*. Let the number of observations be $T = T_1 + T_2$ where T_1 is the length of the wartime period and T_2 is the length of the peacetime period. Then the matrix product in question is

$$X'X = \begin{bmatrix} T & T_1 & T_2 \\ T_1 & T_1 & 0 \\ T_2 & 0 & T_2 \end{bmatrix}. \tag{192}$$

It is clear that the first column of this matrix is the sum of the second and third columns, as was the case with the original matrix X of (191).

If there were no other explanatory variables apart from the dummy variables of the matrix of (191), then the problem of inversion could be easily averted. We should only need to add an extra row to the the matrix X in the form of [0, 1, 1] and to append an extra zero to the vector $y = [y_1, \ldots, y_T]'$ of the observations of the dependent variable and the regression would, in principle, become viable. This addition to the data corresponds to the wholly reasonable restriction that $\gamma_1 + \gamma_2 = 0$; and we should discover that this condition is fulfilled by the resulting estimates.

With the additional row in X, the matrix $X'X$ now takes the form of

$$X'X = \begin{bmatrix} T & T_1 & T_2 \\ T_1 & T_1 + 1 & 1 \\ T_2 & 1 & T_2 + 1 \end{bmatrix} \tag{193}$$

wherein the columns are linearly independent.

The device of supplementing the data will also work when the matrix X includes columns of genuine explanatory variables in addition to the dummy variables: we simply append zeros to the columns of explanatory variables and units to the columns of dummy varaibles. However, there is one drawback which should be guarded against. Imagine that the value of T becomes very large. Then the difference between the matrix of (193), which is formally invertible, and the matrix of (192), which is not invertible, tends to vanish with the effect that the matrix of (193) becomes almost singular or non-invertible. In the technical language of numerical analysis, we say that the latter matrix becomes

ill-conditioned. The consequence is that the numerical inversion of the matrix will be beset by rounding error.

One obvious recourse against this problem of near-singularity is to append to the data matrix a row vector whose elements have values which are non-negligible in comparison with the value of T. Thus the vector $[0, T, T]$ would serve the same purpose as the vector $[0, 1, 1]$; and it would result in a well-conditioned matrix of the form

$$X'X = \begin{bmatrix} T & T_1 & T_2 \\ T_1 & T_1 + T & T \\ T_2 & T & T_2 + T \end{bmatrix}. \tag{194}$$

Whilst the foregoing considerations are of some interest for the light that they cast upon problems of matrix inversion, they may have little practical significance in the present application. The reason is that, if we wish to estimate the parameters κ, γ_1 and γ_2 of equation (190) subject to the restriction that $\gamma_1 + \gamma_2 = 0$, then we might as well estimate the parameters μ and δ of equation (188) in the first instance and then proceed to find the alternative parameters by solving the equations

$$\begin{aligned} \mu &= \gamma_1 + \kappa, \\ \mu + \delta &= \gamma_2 + \kappa, \\ 0 &= \gamma_1 + \gamma_2. \end{aligned} \tag{195}$$

The solution is

$$\gamma_1 = -\frac{\delta}{2}, \quad \gamma_2 = \frac{\delta}{2}, \quad \kappa = \mu + \frac{\delta}{2}. \tag{196}$$

The advantage of this procedure is that it allows one to test for the constancy of the intercept term rather easily by testing the restriction that $\delta = 0$.

We can elaborate the device of dummy variables to accommodate more complicated effects such as the effects of the seasonal variations of economic activity. Imagine that we have quarterly data and that we decide that we must estimate an intercept term which varies over the seasons. We may do this with an equation of the form

$$y_t = d_{t1}\gamma_1 + \cdots + d_{t4}\gamma_4 + x_{t1}\beta_1 + \cdots + x_{tk}\beta_4 + \varepsilon_t. \tag{197}$$

In this case, the associated vector of dummy variables takes the form of

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}. \tag{198}$$

This is simply a partitioned matrix wherein the submatrix $I_4 = [e_1, e_2, e_3, e_4]$ is replicated as many times as the number of years spanned by the data.

As in the case of the dichotomous dummy variables, we can arrange matters in a variety of alternative ways. Thus, in place of equation (197), we may take the equation

$$y_t = \mu + d_{t2}\delta_2 + d_{t3}\delta_3 + d_{t4}\delta_4 + x_{t1}\beta_1 + \cdots + x_{tk}\beta_4 + \varepsilon_t, \tag{199}$$

which is associated with the following matrix of dummy variables:

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{bmatrix}. \tag{200}$$

Here it is the matrix $[i, e_2, e_3, e_4]$ which is replicated in each year. From the estimated values of μ, δ_2, δ_3 and δ_4, we can derive estimates of the alternative parameters $\gamma_1, \ldots, \gamma_4$.

Two-Way Classifications of Qualitative Factors

So far, we have considered qualitative factors which vary only in time. These factors might be accompanied by other factors which vary in a spatial or geographical dimension. In discussing further elaborations of this nature, let us confine our attention to a model which contains only categorical data of a sort which is encoded by dummy variables which take binary values.

Consider an equation of the form

$$y_{tj} = \mu + \gamma_t + \delta_j + \varepsilon_{tj} \tag{201}$$

wherein $t = 1, \ldots, T$ and $j = 1, \ldots, M$. This represents the model which underlies a so-called two-way analysis of variance. For a concrete interpretation, we may imagine that y_{tj} is an observation taken at time t in the jth region. Then the parameter γ_t represents an effect which is common to all observations taken at time t, whilst the parameter δ_j represents a characteristic of the jth region which prevails through time.

As an illustration, we may consider the case where $T = M = 3$. Then the equation (201) gives rise to the following structure:

$$\begin{aligned}\begin{bmatrix} y_{11} & y_{12} & y_{13} \\ y_{21} & y_{22} & y_{23} \\ y_{31} & y_{32} & y_{33} \end{bmatrix} &= \mu \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix} + \begin{bmatrix} \gamma_1 & \gamma_1 & \gamma_1 \\ \gamma_2 & \gamma_2 & \gamma_2 \\ \gamma_3 & \gamma_3 & \gamma_3 \end{bmatrix} \\ &\quad + \begin{bmatrix} \delta_1 & \delta_2 & \delta_3 \\ \delta_1 & \delta_2 & \delta_3 \\ \delta_1 & \delta_2 & \delta_3 \end{bmatrix} + \begin{bmatrix} \varepsilon_{11} & \varepsilon_{12} & \varepsilon_{13} \\ \varepsilon_{21} & \varepsilon_{22} & \varepsilon_{23} \\ \varepsilon_{31} & \varepsilon_{32} & \varepsilon_{33} \end{bmatrix}.\end{aligned} \tag{202}$$

Here there seems to be a large number of parameters—7 in all—in comparison with the number of observations on the variable y. However, it is likely that there will be several observations for each cell of the two-way classification. Thus we might observe n individuals in each region j at each point in time t.

In order to assimilate the two-way model to the ordinary regression model, we may rewrite it in the form of

$$\left[\begin{array}{c} y_{11} \\ y_{21} \\ y_{31} \\ y_{12} \\ y_{22} \\ y_{32} \\ y_{13} \\ y_{23} \\ y_{33} \end{array}\right] = \left[\begin{array}{c|ccc|ccc} 1 & 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 1 & 0 & 0 \\ \hline 1 & 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 & 1 & 0 \\ \hline 1 & 1 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 & 0 & 1 \end{array}\right] \left[\begin{array}{c} \mu \\ \hline \gamma_1 \\ \gamma_2 \\ \gamma_3 \\ \hline \delta_1 \\ \delta_2 \\ \delta_3 \end{array}\right] + \left[\begin{array}{c} \varepsilon_{11} \\ \varepsilon_{21} \\ \varepsilon_{31} \\ \varepsilon_{12} \\ \varepsilon_{22} \\ \varepsilon_{32} \\ \varepsilon_{13} \\ \varepsilon_{23} \\ \varepsilon_{33} \end{array}\right]. \tag{203}$$

Here the matrix X consisting of zeros and ones is called the design matrix. This format can easily accomodate n observations on the variable y taken in the tjth cell. The observations are simply arrayed one below another in the vector y. The corresponding rows of the matrix X are n replicas of the same vector.

A close inspection of the design matrix in (203) will show that there are two degrees of linear dependence amongst its columns. Thus column 1, which is associated with the intercept term μ, is the sum of the columns 2—4, which are associated with the temporal effects γ_1, γ_2 and γ_3. The first column is also the sum of the columns 5—7, which are associated with the spatial effects δ_1, δ_2 and δ_3. To ensure that the parameters are amenable to estimation, we must impose two restrictions:

$$\gamma_1 + \gamma_2 + \gamma_3 = 0, \quad \delta_1 + \delta_2 + \delta_3 = 0. \tag{204}$$

These restrictions may be assimilated to the regression equations of (203) by adding the following rows to the bottom of the design matrix

$$\left[\begin{array}{c|ccc|ccc} 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 \end{array}\right] \tag{205}$$

and by appending two zeros to the bottom of the y vector on the LHS and to the vector ε on the RHS.