Chapter 4

The econometric problem

Since the start of the decade, researchers have been granted general access to the individual data underlying the NES. This gives rise to two difficulties. Firstly, the size of the dataset is a very significant drawback to using it. Secondly, this public access is limited by various confidentiality restrictions. This chapter discusses these problems and outlines the solution path taken.

4.1The size thing

Econometric packages calculate the data needed for each particular regression whenever a regression is run. Data is held as a set of vectors to manipulated in appropriate ways. This has several advantages. It clearly allows much more flexibility in the type of regression allowed than if the data was held in an aggregate form, such as a cross-product matrix. It also makes the creation of new variables simple, including predictions for instruments. It allows for the analysis of the residuals and the creation of estimated covariance matrices. Finally, non-linear models and solution methods are feasible.

Given the speed and capacity of modern computers, holding raw data and manipulating it at will is a sensible way to approach most datasets. However, the NES panel dataset is very large. The raw ASCII files for the first sixteen years are some 600Mb in total, and conversion to binary format does not reduce this as most of the variables are qualitative and so take small values. Reading the data can take some hours.

This causes problems for standard statistical analysis. First of all, some packages are not able to cope with such a large dataset, or only with extreme hardware requirements. However, even if the admissible size of the raw data is unlimited, running regressions becomes an extremely tedious and time consuming process. Misspecification of an equation is heavily

penalised, if the regression has to be run more than once; and so is "exploratory" analysis, where the researcher would hope to try a variety of specifications.

This limits severely the advantage of traditional econometric packages in analysing the data. While SPSS has been used by a number of researchers to create cross-tabulations, running regressions is still a slow business. As the matrices used in a standard analysis are created anew for every regression, each estimation is likely to involve a complete run through the data. If the program automatically creates diagnostic statistics involving residuals, then a second run will have to take place.

This is the only practical solution for non-linear models, those requiring numerical optimisation, or those which need several steps to estimate a final model. However, for one-stage, linear models (in other words, OLS), this is hugely inefficient. This is because linear combinations of variables can be stored very compactly in aggregate form and manipulated to produce further linear combinations. The implications of this are considered in section 4.3.

4.2The confidentiality thing

While access to the data is complete, the amount of information which can be taken "outside" the DE is not. The NES contains information on individuals, their work histories, and their wages, and is subject to, among others, the 1947 Statistics of Trade and 1986 Data Protection Acts. The legal position of the DE is that no information can be removed from its premises which would allow the wages of an individual to be identified. This applies to both computer media and printed materials[1].

This means that the NESPD must remain on the DE's computers only. The practical upshot is

[1] In the mid-1980s the DE constructed a small dataset by aggregating indviduals into cells of three and allowed general access. However, although this got round the confidentiality restrictions and led to some analysis, it did not prove as popular as the DE hoped and was dropped in favour of the panel dataset.

that standard regressions (or almost any statistical analysis using standard packages) have to be run at the DE's head office by DE staff or visiting researchers. This leads to time being wasted, either by DE staff being recalled from other duties to service the dataset, or by researchers needing to travel to London. It also limits the scope for exploratory analysis by introducing a further delay between requesting information and receiving results.

Finally, no information on residuals, for example, or other disaggregated data can be removed. Therefore, diagnostic tests and second-stage regressions have to be run at the DE too. Although researchers can be given summary statistics from running the regression, any further analysis has to return to the DE for processing.

4.3Outline of the solution methods

The main solution taken at the University of Stirling is, as hinted at earlier, to create linear combinations of the variables and then to analyse them using programs devised for this purpose. This method is extremely efficient because OLS is a simple multiplication of two summation terms. If the sums of several variables are calculated at the same time, then multiple regressions can be run simply by multiplying the appropriate sums together.

Consider the simple OLS regression:

$$y = X\beta + u \qquad \textbf{(4.1)}$$

with the normal equations

$$\hat{\beta} = (X'X)^{-1} X'y \qquad \textbf{(4.2)}$$

and the calculation of errors which needs

$$TSS = y'y \quad ESS = \hat{\beta}'X'y \quad RSS = TSS - ESS \qquad \textbf{(4.3)}$$

Just three matrices, X'X, X'y, and y'y, are sufficient to estimate β and calculate R^2, σ^2, and the standard error of the coefficients. Define a vector W≡[X y]. Then the moment matrix

W'W contains all the information needed for the OLS estimation of (4.1):

$$W'W = \begin{bmatrix} X'X & X'y \\ y'X & y'y \end{bmatrix} \qquad \textbf{(4.4)}$$

Suppose X and y are NxK and Nx1 matrices respectively, so that W is a Nx(K+1) matrix. W'W could be created as the moment of the W matrix; however, W'W could also be constructed without having to create the W matrix, for W'W is merely a sum of the moments of each individual row w_i of W:

$$W = \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_N \end{bmatrix} = \begin{bmatrix} x_1\ y_1 \\ x_2\ y_2 \\ \vdots \\ x_N\ y_N \end{bmatrix} \quad \_ \quad W'W = \sum_i^N w_{i'}\, w_i = \begin{bmatrix} \sum x_{i'}\, x_i & \sum x_{i'}\, y_i \\ \sum y_{i'}\, x_i & \sum y_{i'}\, y_i \end{bmatrix} \qquad \textbf{(4.5)}$$

Therefore W'W can be created without any need to store anything bigger than the (K+1)x(K+1) moment matrix.

Thus, the fact that N is enormous for the NES is no longer a restriction. The size of the cross-product matrix to be calculated depends only on the number of variables, not the number of observations. Moreover, the moment matrix is flexible in its definition of variables. The choice of X and y as explanatory and dependent variables, respectively, is for notational convenience. As far as the moment matrix is concerned, there is no difference between the two.

Suppose W is an Nx4 matrix:

$$W = \begin{bmatrix} 1\ x_1\ y_1\ z_1 \\ 1\ x_2\ y_2\ z_2 \\ \vdots\ \ \vdots \\ 1\ x_N\ y_N\ z_N \end{bmatrix} \qquad \textbf{(4.6)}$$

Then W'W is a 4x4 matrix

$$W'W = \begin{bmatrix} N & \sum x_i & \sum y_i & \sum z_i \\ \sum x_i & \sum x_i x_i & \sum x_i y_i & \sum x_i z_i \\ \sum y_i & \sum y_i x_i & \sum y_i y_i & \sum y_i z_i \\ \sum z_i & \sum z_i x_i & \sum z_i y_i & \sum z_i z_i \end{bmatrix} \quad \textbf{(4.7)}$$

This gives a range of possible regressions to run. Any of x, y, or z could be used as the dependent variable with any or all of the others as explanatory variables, including constants if so wished. For example, to regress x on y only requires

$$\hat{\beta} = \left(\sum y_i y_i\right)^{-1} \sum y_i x_i \quad \textbf{(4.8)}$$

while the regression of z on x, y, and a constant gives the estimate

$$\hat{\beta} = \begin{bmatrix} N & \sum x_i & \sum y_i \\ \sum x_i & \sum x_i x_i & \sum x_i y_i \\ \sum y_i & \sum y_i x_i & \sum y_i y_i \end{bmatrix}^{-1} \begin{bmatrix} \sum z_i \\ \sum x_i z_i \\ \sum y_i z_i \end{bmatrix} \quad \textbf{(4.9)}$$

Even the simple 4x4 matrix in (4.7) allows for twenty-one possible regressions (excluding using the constant as a dependent variable). Moreover, adding further variables to this matrix involves relatively little extra space. In this compact form, an extra variable increases the size of the matrix by 2K+1 elements, where K is the previous number of variables. In contrast, adding a new variable to be stored in raw form requires the space to store an extra N observations, and in the NES N could be almost three million.

The efficiency of this approach is due to the fact that each moment matrix contains the information for a number of regressions. If a moment matrix was created for the purposes of running one regression only, then this method has no significant advantage over other solutions in producing the estimates. However, when several regressions are to be run then the time saving of this method is large. The time taken to create moment matrices containing more variables is small relative to the time saved when regressions are run. Moment matrices calculated for the University of Stirling typically include between sixty and one hundred variables for each year, and experience has shown that there is relatively little difference in the

time taken to create these matrices[2].

This is especially important if two-stage methods are being considered. Suppose predicted values of the dependent variable are to be included in the regression. Which predictions are to be used? In this case, the solution is to include all the various predicted values in the moment matrix, as the time and space cost is usually small. Then all the predicted values are available for use at estimation time.

A second advantage of this solution relates to the confidentiality issue. Although these cross-product matrices can be used to run OLS regressions as if they contained disaggregated data, because the data is actually aggregated the matrices can be taken out of the DE and used by researchers at their own institutions[3]. This gives external researchers the chance to do their own analysis, rather than having to inform the DE of their instructions. It also relieves the pressure on the DE's staff resources.

Finally, the cross-product matrix contains useful information other than that needed to run regressions. For example, the W'W matrix in (4.7) could be used to discover, amongst other things, the number of observations N, the mean of all of the variables, the variance and covariances of the variables, and hence the correlation between variables.

In this chapter a simple OLS example has illustrated the main method of analysis used in this regression. In fact, a number of models can be estimated from properly constructed moment matrices, and the next chapter describes those currently available. Analysis of the moment matrix requires some particular software, described in chapter six. However, some standard statistical packages (for example, STATA) can work with moment matrices instead of raw

[2] The time to create the fixed-effects models of sections 5.2 and 5.3 is much greater, but this is because the moment matrix is of size TK, rather than K, and T is the main factor affecting speed.

[3] Some confidentiality checks have to be made in case some individuals can still be identified, but the effect of this is negligible when the full dataset is used.

data, and so the software supplied by Stirling University is not a necessity for analysis.

Two other forms of analysis have been applied to the NES by the researcher team at Stirling. One is the construction of cohort matrices, containing information on cohorts of people of particular ages and in the panel at particular times. The second is the construction of observation histories, to be described in further detail later in chapter seven. Some other researchers have also made use of cross-tabulations (cross-product matrices in all but name), which satisfy the confidentiality requirement and so can be removed from the DE. All these are also subject to confidentiality checks.

The cross-product route is not appropriate for non-linear models or solution methods, and although there is some non-linear analysis currently being carried out, this is necessarily limited to a sample of the NES[4]. Some interesting results have been achieved from linear approximations to the non-linear models, and further work on this is under way. However, it seems that non-linear analysis exploiting the full potential of the NES remains impractical for the time being.

4.3Varying coefficients and fixed-effects

The models used for the software package all allow coefficients to vary over time. They allow for fixed time effects, and the "fixed effects" and differencing models allow for fixed individual effects. The time-varying coefficients are "fixed" in a similar way to the fixed individual effect, in that the estimation of fixed (T sets of K) parameters rather than the characteristics of a distribution is involved.

The choice of a fixed-effects specification for individual heterogeneity is twofold. From a practical perspective, the two stages necessary for GLS estimation of a random-effects

[4] Source: DE

specification constitute a serious drawback. The covariance method is a single-stage effort and so is far more practical given the nature of the data. This does not stop random-effects models being estimated; however, the onus of calculating the first-stage component estimates, recreating the cross-product matrix and re-estimating is shifted onto the user.

There are also theoretical considerations for using fixed-effects, the most prominent being the potential for collinearity between the individual effects and explanatory variables. Given the self-selecting nature of many labour force decisions, the assumption that unchanging personal characteristics are not correlated with the choice of occupation, industry, and so on seems unlikely. For example, Vella(1994) argues strongly that a significant factor in job choice is due to something called "attitude". As was discussed in chapter two, these potential correlations can lead to the random-effects estimator being biased or inconsistent while the fixed-effects estimator remains valid. Although the fixed-effects estimator is relatively inefficient for random-effects models, the large number of individuals in the NES make the trade of efficiency for robustness appealing.

The choice of varying coefficients likewise has a practical and theoretical component. From a practical point of view, the time penalty for estimating cross-sectional models with varying coefficients as opposed to pooled CS models is relatively small. For the fixed-effects models the penalty is more significant; but the extra data requirements are relatively minor and the programming is straightforward. Therefore the practical cost of allowing for varying-coefficient models is not prohibitively high, and given the compelling theoretical arguments for allowing coefficients to vary over time, the trade-off seems reasonable.

Whilst many models have been developed which allow coefficients to vary over individuals (Hsiao (1992) surveys these), relatively few authors have considered coefficients which vary over time. The theoretical work in this area is largely limited to the work of Chamberlain (1984), and applied work with time-varying coefficients is almost non-existent. However, the assumed stability of the structure of labour market is open to doubt. Surveys of the UK labour

market (for example, Robinson (1994)) lead to the overwhelming conclusion that the employment in the UK has changed considerably in recent years. To assert that these changes have not been reflected at a micro-level, without testing the alternative hypothesis of varying coefficients, seems an unjustifiable presumption.

Given the lack of information about changes in the parameters of the market, imposing a particular structural form on changing coefficients (for example, evolutionary coefficients or random-effects about a stable mean) is unlikely to improve on the constant-coefficient assumption except by luck. Instead, the approach taken here is to allow for a different set of slope parameters for each year. This allows any changes in the labour market to occur in a relatively unrestricted way, by letting the responses of the population vary without reference to any specific structure. The default models therefore condition on time-varying slope coefficients in the same manner as the fixed-effects estimator conditions on individual heterogeneity. Just as the fixed-effects specification is less efficient than the random-effects specification, this slope conditioning is less efficient than a properly-specified model which takes account of the structure of coefficients over time (for example, models with systematic evolution of the coefficients). However, the lack of evidence on trends in the parameters of the labour market would seem to justify this approach for the moment[5].

[5] Chapters eight and nine consider this issue of structural change in more detail, and at this point it may be noted that F-tests on varying-coefficients model comprehensively reject the hypothesis of parametric stability.