Empirical exercise – The FE estimator as an application of the FWL theorem

1. Aim of the exercise

The fixed effects estimator is an estimator that is widely used to estimate the coefficients of the independent variables in a fixed effects panel data model. The derivation of this estimator makes direct use of the FWL theorem. Hence, this exercise not only serves to the purpose of learning to apply the FWL theorem, but it also gives the theoretical rationale for the fixed effects estimator which is a typical topic in a panel data course.

The empirical context is that we want to investigate the determinants of the hourly wage. We consider two variables that we suspect are affecting the hourly wage in general. The first is experience, measured in terms the number of years spent in the labor market. We expect that as workers cumulate experience, they are eligible for a more competitive wage. The second variable is union membership. We expect that union membership is a proxy for bargaining power in negotiating sectoral wages in the collective labour agreements between the employees and employers.

What is special about the data at hand is that we observe what the same individuals do each year over a period of eight years. Hence, in the data, for each variable, we have observations for multiple years for a given individual. This means that from this data we not only can learn about the earnings behaviour of individuals on average, but also learn from the variation in earnings over time within individuals. We can exploit this data using a panel data regression model.

2. Load the data

Let MATLAB read the supplied mat file.

```
clear;
load 'M:\exercisefe.mat';
```

3. Drop variables and define the dependent variable

Use the code presented at the end of the section to keep the variables `exper`, `expersq`, `union`, and `nr`. Define the dependent variable as `y`. `y` represents `lwage`.

```
clearvars -except exper expersq expersq lwage nr union year;
y = lwage;
```

4. Determine the number of observations and individuals in the panel

This exercise uses panel data. That is, we have multiple years of data available for each individual in the panel. In the Workspace, inspect the two variables `nr` and `year`. `nr` identifies an individual in the panel. `year` is the time variable. The two variables together identify a

unique observation in the data. Determine the number of individuals and the number of observations using the commands presented at the end of the section. These quantities will be used for calculations later in this exercise.

```
uniq = unique(nr);
N_ind = size(uniq,1);
N_obs = size(nr,1);
```

5. Create a matrix containing dummy variables for panel units

In Section 6 we will estimate a linear model that allows for an intercept term for each individual in the panel data regression model. We do this to exploit individual heterogeneity in the dependent variable we are trying to explain. This means that we need to create dummy variables that indicate individuals in the panel. The code at the end of the section serves to this purpose. In particular, `NaN(N_obs,N_ind)` creates an empty matrix with missing values that are to be inputted. The row and column dimensions of the matrix are `N_obs` and `N_ind` since there are `N_obs` observations for `N_ind` individuals in the data. We name this matrix as `D`.

The for loop presented at the end of the section adds a dummy variable to each column of the empty matrix `D` in an iterative process. In particular, the first line of the for loop specifies the index of the for loop. The index runs from `1` to `N_ind` because a program in the second line of the for loop is about to operate on the `N_ind` columns of the matrix `D`. In the second line of the for loop, in each iteration of the for loop, a small program fills column `i` of the matrix `D` with a value of 1 if a personal identification number stored in row `i` of the vector array `uniq` is found in the vector array `nr`, and with a value of 0 otherwise. Inspect the content of the matrix array `D` in relation to the vector array `nr` in the Workspace.

```
D = NaN(N_obs,N_ind);
for i = 1:N_ind
    D(:,i) = nr == uniq(i,1);
end
```

6. Create the systematic component of the Least Squares Dummy Variable model

Consider the linear model $\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \boldsymbol{\iota}\alpha_i + \boldsymbol{\varepsilon}_i$. Assumptions of the model are the following. There are T observations available for each individual i. For example, the panel data we use for this exercise contains eight years of data for each individual. There are N individuals. $\mathbf{y}_i$ is $T \times 1$. $\mathbf{X}_i$ is a matrix of independent variables. In this exercise it contains `exper`, `expersq`, and `union`. $\mathbf{X}_i$ has T rows. Row t contains the row vector $\mathbf{x}'_{it}$. $\mathbf{x}'_{it}$ contains k observations for k independent variables for individual i at time t. It is a row vector with dimension $1 \times k$. Since there are T observations for each individual i, $\mathbf{X}_i$ is $T \times k$. $\boldsymbol{\beta}$ represents the vector of coefficients. It is a column vector with dimension $k \times 1$. $\boldsymbol{\iota}$ (Greek letter 'iota') is a $T \times 1$ column vector of 1s. α_i is a time-invariant constant term specific to each individual in the panel, and it is potentially correlated with $\mathbf{X}_i$. $\boldsymbol{\varepsilon}_i$ meets the standard OLS assumptions so that it is not correlated with the independent variables and has constant variance conditional on the

independent variables. Outline this regression in matrices on a piece of paper. The dimensions of the matrices as specified above should help when doing this.

If we stack the N individuals, the model presented above for individual i becomes $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{D}\boldsymbol{\alpha} + \boldsymbol{\varepsilon}$. $\mathbf{y}$ is $NT \times 1$. $\mathbf{X}$ is $NT \times k$. $\mathbf{D}$ has N diagonal elements. Each element of the diagonal is a vector, is the same, and is given by the column vector $\boldsymbol{\iota}$. All of the off-diagonal elements are $\mathbf{0}$ column vectors of size $T \times 1$. Hence, $\mathbf{D}$ is $NT \times N$. $\boldsymbol{\alpha}$ is $N \times 1$ since there are N different α_is. This is the least squares dummy variable (LSDV) model.

Combine the matrices `X` and `D` into a larger matrix, and name this matrix as `Z`. `Z` is to be considered as the systematic component of the LSDV regression you are about to estimate. Note that no constant term is added to `Z`. Why? This is to avoid perfect collinearity since a column of ones can be expressed as a linear combination of the columns of `D`.

```
Z = [exper expersq union D];
```

7. Estimate the Least Squares Dummy Variable model

Use the OLS estimator to estimate the LSDV model. Name this estimator as `B_hat_Z`. Note that the calculation of `B_hat_Z` involves inversion of the matrix `Z'*Z` which has a dimension of $k + N \times k + N$. For large N, this inversion might demand a substantial amount of computer power. Furthermore, N intercept terms need to be estimated which requires a lot of degrees of freedom. Obtain the coefficient estimates for `exper`, `expersq`, and `union`.

```
B_hat_Z = (Z'*Z)\Z'*y;
B_hat_Z(1:3);
```

8. Obtain the transformation matrix

In this and the next section you will derive an estimator that serves to two purposes. First, it will avoid the inversion of a large matrix that the LSDV estimator had to deal with as described in the previous section. Second, it will avoid the estimation of a large number of intercept terms, and still provide the same coefficient estimates for the main variables of interest `exper`, `expersq`, and `union`. After all, we are usually interested in the effects of these variables, rather than the estimated coefficients of the intercept terms for all individuals in the panel.

Create the projection matrix for `D`, and name it as `P_D`. Create the projection matrix that is orthogonal to the projection matrix `P_D`, and name it as `M_D`. In the Workspace, double-click on `M_D` to view its content. Consider the block of entries at the top-left corner of the spreadsheet. This block, or any other block, can be represented in matrix form as $\mathbf{I}_T - \frac{1}{T}\iota\iota'$. $\mathbf{I}_T$ represents the identity matrix and it is $T \times T$. ι is as described in Section 6. $\mathbf{I}_T - \frac{1}{T}\iota\iota'$ creates deviations from time averages for each individual i. Convince yourself that this is true.

Premultiply `y` with `M_D`. In the resulting column vector, an entry represents the deviation of an observation of an individual from the time average of the observations of that individual: $y_{it} - \bar{y}_i$. Convince yourself that this is true. In the same manner, premultiply `X` with `M_D`. The resulting matrix has three columns associated with the three independent variables. In a given column of the matrix, an entry represents the deviation of an observation of an individual from the time average of the observations of that individual for a given variable: $x_{it} - \bar{x}_i$.

```
X = [exper expersq union];
I = eye(N_obs);
P_D = D*((D'*D)\D');
M_D = I-P_D;
```

9. The Fixed Effects model and the FWL theorem

The regression of the demeaned y_{it} (`M_D*y`) on the demeaned $\mathbf{x}_{it}$ (`M_D*X`) is the fixed effects regression and is given by $y_{it} - \bar{y}_i = (\mathbf{x}'_{it} - \bar{\mathbf{x}}'_i)\boldsymbol{\beta} + \varepsilon_{it} - \bar{\varepsilon}_i$. Use the OLS estimator to estimate the coefficients of this regression. Name this estimator as `B_hat_X`. This estimator is given by `((M_D*X)'*M_D*X)\(M_D*X)'*M_D*y`. This is the fixed effects estimator. It is nothing but the OLS estimator on the transformed variables `M_D*y` and `M_D*X`. The estimator can be simplified to `(X'*M_D*X)\X'*M_D*y`.

The calculations above have two important results. First, the fixed effects estimator you have used is an application of the FWL theorem. Second, the fixed effects regression you have carried out drops the time-invariant individual fixed effects. Convince yourself that it does. This means that you do not have to estimate an intercept term for each individual in the panel, as you did while estimating the Least Squares Dummy Variable model in Section 7.

The coefficient estimate of union membership indicates that, compared to workers who are not a union member, those who are a union member earn 8.33 percent more, on average, holding experience constant as well as holding individual specific time-invariant factors constant. These individual specific factors represent observable or unobservable individual heterogeneity. Interoperation of the coefficient estimate of experience is left as an exercise.

```
M_D*y;
M_D*X;
B_hat_X = (X'*M_D*X)\X'*M_D*y;
```