BMS and the Fixed Effects Estimator - A Tutorial

Martin Feldkircher*

This version: June 2011

Abstract

This tutorial should illustrate how to use Bayesian Model Averaging (BMA) in R with panel data.

1 Introduction

Methods for estimating econometric models with panel data have been frequently discussed in the literature (see eg Mundlak, 1978). Two estimators seem to resurface most often: the fixed effects estimator (FE) and the random effects estimator (RE). Both have their separate virtues and underlying assumptions (for an exposition see Bartels, 2008). Since the FE estimator can be easily cast into the linear regression framework that is used for BMS it will be our focus in this tutorial. For an application of Bayesian model averaging employing the RE estimator please refer to Moral-Benito (2011). Furthermore, a great deal of the literature seems to pivot around the question of how to calculate standard errors (Bartels, 2008). At the current stage, we abstract from the calculation of so-called *clustered* standard errors since we stay in a pure Bayesian framework in BMS (and standard errors are a classical concept).

For the purpose of illustration we will use the data put forward in Moral-Benito (2011) and made publicly available at Moralbenito.com. The data contains $K = 35$ variables (including the dependent variable, the growth rate of per capita GDP) for $N = 73$ countries and for the period 1960-2000. The appendix lists the variables together with a short description. The dependent variable, GDP growth, is calculated for five year averages resulting into eight observations per country. Moral-Benito (2011) argues in favor of calculating averages of flow variables, while stock variables have been measured at the first year of each five-year period. The data can be downloaded here http://www.moralbenito.com/research.htm ('download data for the paper Determinants of Economic Growth: A Bayesian Panel Data Approach).

```
> library(BMS)
```

After having started R and loaded the BMS we can read in the data file:
For that purpose I have simply saved the 'Dataset.xls' file as a 'Dataset.csv' file and have replaced all missing values by 'NA' in excel. You can also download the data here 'panel-Dat.rda'.

*martin.feldkircher@gzpace.net.

```
              GRW_PWT  GDP_PWT      POP       PGRW       IPR      OPEM
DZA_1960  0.06187711 8.254050 10909.29 0.01644121  59.15831 1.2944342
DZA_1965  0.05152103 8.315927 11963.09 0.03053423  62.97262 0.8557407
DZA_1970 -0.06349568 8.367448 13931.85 0.02917826  78.94137 0.9270065
DZA_1975  0.23215850 8.303953 16140.25 0.03087119 105.67677 1.1867272
DZA_1980  0.07489113 8.536111 18861.62 0.03249203  98.74844 1.0174890
DZA_1985  0.01557414 8.611002 22182.25 0.02804192 101.39928 0.7784194
               CSH       GSH        ISH     LBF      LEX      PDE
DZA_1960 0.5099434 0.2166622 0.14821260 0.29712 48.32439 4.597672
DZA_1965 0.5172348 0.1652750 0.09507138 0.27192 51.42439 5.006004
DZA_1970 0.5728414 0.1742702 0.16151614 0.25738 54.47561 5.771411
DZA_1975 0.7239094 0.2314359 0.25317528 0.25883 57.47561 6.725335
DZA_1980 0.7621881 0.2547458 0.23459953 0.26404 60.47561 7.838458
DZA_1985 0.6667698 0.2093508 0.17095402 0.27489 65.71951 9.186141
             URBP NWTR LND ADIS    TAR    TPOP     LAR OPEI INDP SOC
DZA_1960 0.30440    5   0 1675 0.1618 0.12865 2381740    0    2   1
DZA_1965 0.37628    5   0 1675 0.1618 0.12865 2381740    0    2   1
DZA_1970 0.39500    5   0 1675 0.1618 0.12865 2381740    0    2   1
DZA_1975 0.40330    5   0 1675 0.1618 0.12865 2381740    0    2   1
DZA_1980 0.43542    5   0 1675 0.1618 0.12865 2381740    0    2   1
DZA_1985 0.47969    5   0 1675 0.1618 0.12865 2381740    0    2   1
         WAR CLI EU SAFR LATM EAS       MAL      P15       P65   PED
DZA_1960   1   0  0    0    0   0 0.7667722 0.437300 0.0385185 0.820
DZA_1965   1   0  0    0    0   0 0.7667722 0.459700 0.0328776 0.515
DZA_1970   1   0  0    0    0   0 0.9044099 0.483700 0.0413211 0.671
DZA_1975   1   0  0    0    0   0 0.9044099 0.474000 0.0418279 0.875
DZA_1980   1   0  0    0    0   0 0.0003254 0.464900 0.0393276 1.235
DZA_1985   1   0  0    0    0   0 0.0003254 0.454239 0.0394254 1.607
           SED PR CL
DZA_1960 0.139  6  6
DZA_1965 0.118  6  6
DZA_1970 0.141  6  6
DZA_1975 0.189  6  6
DZA_1980 0.280  6  6
DZA_1985 0.458  6  6
```

The rownames of the data are a combination of the country code and the year of the observation. The data is provided in a `data frame` consisting of stacked observations per column. That is, the first column containing the dependent variable consists

$$Y_1 = (y_{1,1}, y_{1,2}, \ldots, y_{1,T=8}, \ldots, y_{N,1}, y_{N,2}, \ldots, y_{N,T})$$

This is also the format we can use later on when calling the `bms` function.
We will use two approaches to estimate a fixed effects (FE) panel (country / time fixed effects). The first approach makes use of the Frisch-Waugh-Lovell theorem (see eg Lovell, 2008) boiling down to demeaning the data accordingly. That is, in the case of country fixed effects, subtract from each observation (dependent and independent variable) the within

country mean. For the case of time fixed effects, subtract from each observation the mean across countries per time period. We will start with the country fixed effects first.
For that purpose we will have to re-shape the data frame and put it into the form of a three dimensional array ($T \times K \times N$). That can be achieved with the function `panel_unstack`. Since `bms` uses data in its stacked form, we have to make use of `panel_stack` as well. Both functions are not part of the `BMS` library and thus have to be copy and pasted into your `R` console by yourself:

```
> panel_unstack = function(stackeddata, tstep = NULL) {
+     bigT = nrow(stackeddata)
+     K = ncol(stackeddata)
+     if (is.null(tstep))
+         tstep = bigT
+     X1 = aperm(array(as.vector(t(as.matrix(stackeddata))),
+         dim = c(K, tstep, bigT/tstep)), perm = c(2, 1,
+         3))
+     try(dimnames(X1)[[1]] <- unique(sapply(strsplit(rownames(stackeddata),
+         "_"), function(x) x[[2]])), silent = TRUE)
+     try(dimnames(X1)[[2]] <- colnames(stackeddata), silent = TRUE)
+     try(dimnames(X1)[[3]] <- unique(sapply(strsplit(rownames(stackeddata),
+         "_"), function(x) x[[1]])), silent = TRUE)
+     return(X1)
+ }
> panel_stack = function(array3d) {
+     x1 = apply(array3d, 2, rbind)
+     try(rownames(x1) <- as.vector(sapply(dimnames(array3d)[[3]],
+         FUN = function(x) paste(x, dimnames(array3d)[[1]],
+             sep = "_"))), silent = TRUE)
+     return(as.data.frame(x1))
+ }
```

We can now easily transform the data from its stacked form into the three-dimensional array via:

```
> dat.array = panel_unstack(panelDat, tstep = 8)
```

where we have set `tstep=8` since we have eight time periods per country. The advantages of the three-dimensional array are that we can easily access each dimension of the data:

```
> dat.array[, , "ZWE"]
> dat.array["1965", , ]
> dat.array[, "GSH", ]
```

2 Fixed Effects Estimation by Demeaning the Data

The function `demean` (again not part of the `BMS` library, so copy and paste the following lines into your `R` console) demeans the data to estimate individual (eg country), time and individual and time fixed effects. It takes as argument the three dimensional data array we have created above (`dat.array`) and via `margin` we can specify over which dimension we want to demean the data (country / time).

```
> demean = function(x, margin) {
+     if (!is.array(x))
+         stop("x must be an array/matrix")
+     otherdims = (1:length(dim(x)))[-margin]
+     sweep(x, otherdims, apply(x, otherdims, mean))
+ }
```

Demeaning is now easily accomplished by:

```
> timeDat = panel_stack(demean(dat.array, 3))
> countryDat = panel_stack(demean(dat.array, 1))
```

where we have used `panel_stack` to re-transform the demeaned data into its stacked form that can be passed to the `bms` function.
Since in the data frame only the first 12 explanatory variable show variation over time, we will restrict estimation to these variables only.

```
> modelCd = bms(countryDat[, 1:13], user.int = F)
> modelTd = bms(timeDat[, 1:13], user.int = F)
```

We will briefly discuss the results in the next section (see Table 1 and Table 2). Note that demeaning the data yields the same posterior estimates for the coefficients as with incorporating the FE directly, the approach we opt for in the next section. However, the posterior variance for the coefficient estimates is not identical (though very similar). Also note that demeaning does not save you from the degrees of freedom problem when incorporating the large set of fixed effects by a set of dummy variables. For an application using BMA with country FEs see for example Crespo Cuaresma et al. (2009).

3 Fixed Effects Estimation with Dummy Variables

We will now turn to the second possibility of estimating FEs, which is the dummy variable approach. The advantage of the dummy variable approach is also that it yields estimates for the FEs which can be important for some applications. For the dummy approach we will make use of the new `BMS` feature of holding variables constant (not sampling) them by the `bms` argument `fixed.reg`. Please make sure that you have installed `BMS` $\geq$ version 0.3. We start now with creating the country dummies:

```
> bigT = nrow(panelDat)
> tstep = 8
> countryDummies = kronecker(diag(bigT/tstep), rep(1, tstep))
> colnames(countryDummies) = dimnames(dat.array)[[3]]
> countryDummies = countryDummies[, -1]
```

In a same fashion we can easily create a set of time dummies:

```
> timeDummies = matrix(diag(tstep), bigT, tstep, byrow = T)
> colnames(timeDummies) = dimnames(dat.array)[[1]]
> timeDummies = timeDummies[, -1]
> modelTdummy = bms(cbind(panelDat[, 1:13], timeDummies),
+     fixed.reg = colnames(timeDummies), user.int = F)
```

Running the two regressions (for the first 13 elements of the demeaned data frame only):

```
> modelCdummy = bms(cbind(panelDat[, 1:13], countryDummies),
+     fixed.reg = colnames(countryDummies), user.int = F)
> modelTdummy = bms(cbind(panelDat[, 1:13], timeDummies),
+     fixed.reg = colnames(timeDummies), user.int = F)
```

should yield the same results as with demeaning. Type `coef(modelCdummy)` / `coef(modelTdummy)` to get the results in `R`. These are summarized in the Table below:

	PIP	Post Mean	Post SD	PIP	Post Mean	Post SD
GDP_PWT	1.00	-0.24	0.02	1.00	-0.24	0.02
POP	1.00	0.00	0.00	1.00	0.00	0.00
PGRW	0.29	-0.53	0.99	0.29	-0.53	0.99
IPR	0.33	-0.00	0.00	0.33	-0.00	0.00
OPEM	1.00	0.16	0.03	1.00	0.16	0.03
CSH	1.00	-0.30	0.07	1.00	-0.30	0.07
GSH	0.98	-0.46	0.14	0.98	-0.46	0.14
ISH	0.52	0.13	0.15	0.52	0.13	0.15
LBF	0.51	0.28	0.32	0.51	0.28	0.32
LEX	0.13	0.00	0.00	0.13	0.00	0.00
PDE	0.07	-0.00	0.00	0.07	-0.00	0.00
URBP	0.08	-0.01	0.04	0.08	-0.01	0.04

Table 1: Estimation of country fixed effects: Left panel based on demeaning the data, right panel on the dummy variable estimation approach.

As one can see the results are very similar to each other. Since we have used 'full enumeration' no stochastic variability should be expected for the two approaches. However, when using large data sets and thus the MCMC sampler in turn, please bear in mind that there might be some stochastic variation of results when running differen `bms` chains. Posterior coefficients for the model employing country fixed effects are to be interpreted with respect to the within variation: A positive coefficient on the variable measuring the country's openness (`OPEM`) means that if openness increases *within* a country GDP growth is incraesing. On the other hand, time fixed effects in the particular example look at the between variation of the data. That is, if openness *across countries* (at once) increases, does this affect GDP growth? From Table 2 we see that this effect is smaller compared to that for the within transformed data.

	PIP	Post Mean	Post SD	PIP	Post Mean	Post SD
GDP_PWT	1.00	-0.07	0.01	1.00	-0.07	0.01
POP	0.43	0.00	0.00	0.43	0.00	0.00
PGRW	0.99	-2.54	0.65	0.99	-2.54	0.65
IPR	0.77	-0.00	0.00	0.77	-0.00	0.00
OPEM	0.38	0.01	0.02	0.38	0.01	0.02
CSH	0.10	-0.00	0.02	0.10	-0.00	0.02
GSH	0.14	-0.01	0.04	0.14	-0.01	0.04
ISH	0.86	0.21	0.11	0.86	0.21	0.11
LBF	0.06	0.00	0.03	0.06	0.00	0.03
LEX	1.00	0.01	0.00	1.00	0.01	0.00
PDE	0.35	0.00	0.00	0.35	0.00	0.00
URBP	0.10	-0.00	0.02	0.10	-0.00	0.02

Table 2: Estimation of time fixed effects: Left panel based on demeaning the data, right panel on the dummy variable estimation approach.

References

[1] Bartels, Brandon (2008). Beyond 'Fixed versus Random Effects': A Framework for Improving Substantive and Statsitcal Analysis of Panel, Time-Series Cross-Sectional, and Multilevel Data. *Mimeo, Stony Brook University, New York.*

[2] Crespo Cuaresma, J. and Doppelhofer, G. and Feldkircher, M. 2009: The Determinants of Economic Growth in European Regions. *CESifo Working Paper Series, No. 2519.*

[3] Lovell, M., 2008. A Simple Proof of the FWL (Frisch,Waugh,Lovell) Theorem. *Journal of Economic Education.*

[4] Moral Benito, Enrique (2011). Determinants of Economic Growth: A Bayesian Panel Data Approach. *The Review of Economics and Statistics, forthcoming.*

[5] Mundlak, Yair, 1978. On the Pooling of Time Series and Cross Section Data. *Econometrica, Vol. 46, p. 69-85.*

4 Appendix

GRW_PWT	Dependent variable (source: PWT 6.2) Growth of per capita GDP over 5-year periods (2000 US dollars at PPP)
GDP_PWT	log Initial GDP (PWT 6.2)(IN) Logarithm of initial real GDP per capita (2000 US dollars at PPP)
POP	Population (source PWT 6.2)(IN) Population in thousands of people
PGRW	Population Growth (source PWT 6.2)(AV) Average annual growth rate of population
IPR	Investment Price (source PWT 6.2)(AV) Average investment price level
OPEM	Opennes measure (source PWT 6.2)(AV) Export plus imports as a share of GDP
CSH	Consumption Share (source PWT 6.2)(AV) Consumption as a share of GDP
GSH	Government Share (source PWT 6.2)(AV) Government consumption as a share of GDP
ISH	Investment Share (source PWT 6.2)(AV) Investment as a share of GDP
LBF	Labor Force (source PWT 6.2)(IN) Ratio of workers to population
LEX	Life Expectancy (source WDI 2005)(IN) Life expectancy at birth
PDE	Population Density (source WDI 2005)(IN) Population divided by land area
URBP	Urban Population (source WDI 2005)(IN) Fraction of population living in urban areas
NWTR	Navigable Water (source Gallup et. al) Fraction of land area near navigable water
LND	Landlocked Country (source Gallup et. al) Dummy for landlocked countries
ADIS	Air Distance (source Gallup et. al) Logarithm of minimal distance in km from New York, Rotterdam, or Tokio
TAR	Tropical Area (source Gallup et. al) Fraction of land area in geographical tropics
TPOP	Tropical Population (source Gallup et. al) Fraction of population in geographical tropics
LAR	Land Area (source Gallup et. al) Area in km^2
OPEI	Openness Index (source: Sachs and Warner) Index of trade openness from 1 (highest) to 0
INDP	Independence (source Gallup et. al) Timing of national independence measure: 0 if before 1914; 1 if between 1914 and 1945; 2 if between 1946 and 1989 and 3 if after 1989
SOC	Socialist (source Gallup et. al) Dummy for countries under socialist rule for considerable time during 1950 to 1995
WAR	War Dummy (source: Barro and Lee) Dummy for countries that participated in external war between 1960 and 1990
CLI	CLimate (source Gallup et. al) Fraction of land area with tropical climate
EU	Europe Dummy for EU countries
SAFR	Sub-Saharan Africa Dummy for Sub-Sahara African countries
LATM	Latin America Dummy for Latin American countries
EAS	East Asia Dummy for East Asian countries
MAL	Malaria (source Gallup et. al) (IN) Fraction of population in areas with malaria
P15	Population under 15 (source: Barro and Lee)(IN) Fraction of population younger than 15 years
P65	Population over 65 (source: Barro and Lee)(IN) Fraction of population older than 65 years
PED	Primary Education (source: Barro and Lee)(IN) Stock of years of primary education
SED	Secondary Education (source: Barro and Lee)(IN) Stock of years of secondary education
PR	Political Rights (source: Freedom House)(IN) Index of political rights from 1 (highest) to 7
CL	Civil Liberties (source: Freedom House)(IN) Index of civil liberties from 1 (highest) to 7

Table 3: Source: `Moralbenito.com`. Notes: 1.-(IN)̅ refers to initial value for the 5-year period. 2.-(AV) refers to 5-year average. 3.-Variables without neither (IN) nor (AV) are the same for all the years.