Session 7 – High Dimensional Fixed Effects

Data Skills for Research
Kellogg Research Support

August 17, 2023

Northwestern | Kellogg

Fixed Effects

Fixed effects / group-specific intercepts are frequently used in econometric models:

- Get rid of omitted variable bias
- Handle potential unobserved heterogeneity
- Increase precision

Roadmap

- **Why use Fixed Effects?**

- **How to deal with “too many” fixed effects**
  - de -meaning variables

- **Data Examples and Computational Horse Races**
  - Stata 
  - R 
  - Python 

- **Parallelizing Fixed Effect Code on KLC**

But that’s what we have computers for, isn’t it?

- An important equation:

$$\widehat{\beta} = (X'X)^{-1} X'Y$$

- X’X is a k by k matrix, where k is the number of regressors
- What does this mean if you have tens of thousands of fixed effects??

Frisch, Waugh and Lovell to the rescue

- It turns out that including a full set of fixed effects to your predictor variables is equivalent to subtracting the group mean from each observation for each dependent and (‘other’) predictor variable, and running a regression using the demeaned variables

- Many of the efficient fixed-effects algorithms use this trick

Which command/software is right for you?

- Is the researcher interested in the point estimates/standard errors of the fixed effects themselves?
- Are multiple levels of fixed effects used?
- Do you need clustered standard errors? Clustered at what level?

Package Comparison

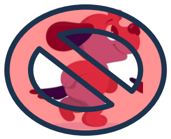
Feature/Aspect	reghdfe (Stata)	felm (R)	feols (R)	fixedeffect (Python)
Developer	Sergio Correia (ftools)	Simen Gaure (felm)	-	-
Core Efficiency	Efficient reprogramming using Mata	Efficient C++ implementation	Efficient C implementation	Efficient Python implementation
Performance	Faster for one-level fixed effects	Fast for one- and two-level fixed effects	Fast for one- and two-level fixed effects	Fast for high-dimensional fixed effects
Multiple Fixed Effects	Allows multiple levels of fixed effects	Limited to one- and two-level fixed effects	Limited to one- and two-level fixed effects	Limited to one- and two-level fixed effects
Clustering	Supports various clustering options	Limited clustering options	Limited clustering options	Limited clustering options

Data Example

- Dataset on Madison, WI home sales (Hendel et al. 2009)
- Research Question: Are realized prices higher when using a realtor, or when sold by owner themselves?
  - Potential for selection?
  - Solution: home fixed effects – only use within-home comparisons!
  - Contains 22,000 home sales and **16,000 fixed effects**

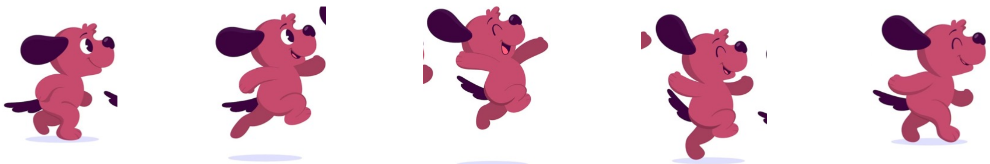
Stata Fixed Effect Models – KLC Run Time

```
reg log_sale_price list_fsbo age_home new i.month year i.home_id,
     cluster(home_id)
```



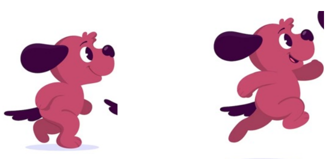
Did not Run

```
areg log_sale_price list_fsbo age_home new i.month year, absorb(home_id)
     cluster(home_id)
```



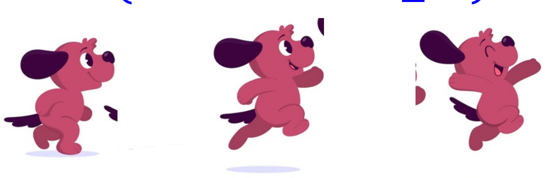
0.32 seconds

```
xtreg log_sale_price list_fsbo age_home new i.month year, fe cluster(home_id)
```



0.09 seconds

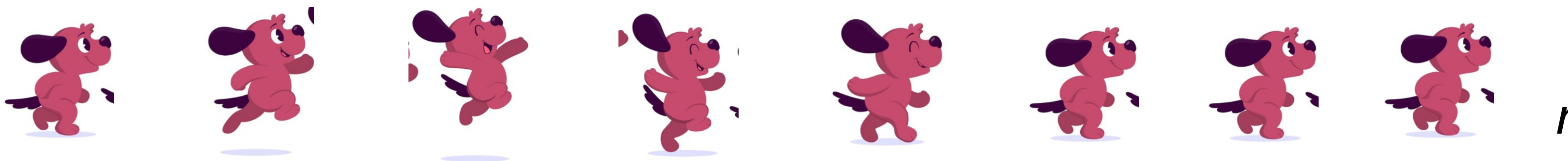
```
reghdfe log_sale_price list_fsbo age_home new i.month year, absorb(home_id)
     vce(cluster home_id)
```



1.00 seconds

R Fixed Effect Models - Run Time

```
lm( log_sale_price~ list_fsbo+ new + as.factor(month) + as.factor(home_id),
       data=homes_data)
```

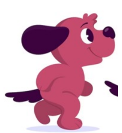


57 minutes

```
library(lfe)
felm(log_sale_price ~ list_fsbo + age_home + new + as.factor(month) + year
                    | home_id | 0 | home_id ,  data=homes_data)
```

0.16 seconds

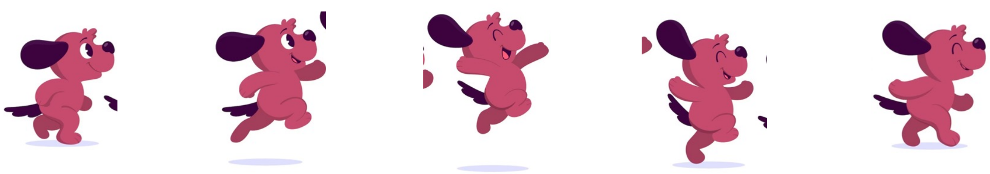
```
library(fixest)
feols(log_sale_price ~ list_fsbo + age_home + new + as.factor(month) + year |
home_id)
```



0.05 seconds

Python Fixed Effect Model - Run Time

```
fixedeffect(data_df = homes_data, dependent = log_sale_price, exog_x =
['list_fsbo', 'age_home', 'new'], category = ['home_id', 'month'], cluster =
home_id)
```



0.37 seconds

Creating a Fixed Effect Conda Environment

Using a text editor, create the following file on KLC called `fe_env.yml`

```yaml
name: fe_env
channels:
  - conda-forge
  - defaults
dependencies:
  # R packages
  - r-base=4.1.1
  - r-felm
  - r-lfe

  # Python packages
  - python=3.9
  - matplotlib
  - statsmodels
  - fixedeffect
```

You can run this file using the mamba module on KLC to create the environment

```
module load mamba
mamba env create -f r_env.yml
```

Activating a Conda Environment

To activate the environment in the future, either load mamba:

```
module load mamba
```

OR any version of conda:

```
module load python-anaconda3/2019.10
```

Then run this line to activate the environment:

```
source activate fe_env
```

To leave the environment:

```
source deactivate fe_env
```

To output the yaml file so you can share your environment with others:

```
conda env export > fe_env.yml
```

HD Fixed Effect Parallelization in R

Overview:

parallel library accelerates R computations by enabling simultaneous execution of tasks on multiple CPU cores.

Key Features:

parLapply: parallelizing tasks that require significant computation or when data needs to be shared across cores.

mclapply: Suited for scenarios with minimal data sharing and when independent tasks can be executed in parallel.

```r
# Load the required packages
library(parallel)

# Define a function to estimate fixed effects model
estimate_fe_model <- function(df) {
  library(felm)
  model <- felm(y ~ x | id, data = df)
  return(summary(model))
}

# Using parLapply for parallel fixed effects estimation
fe_model_results <- parLapply(data, estimate_fe_model)
```

HD Fixed Effect Example Studies

Environmental Economics

Study: Impact of Green Regulations on Firm Emissions

Method: Fixed effects control for unobserved firm-specific characteristics and regulatory effects.

Labor Economics

Study: Gender Wage Gap Over Time

Method: Fixed effects capture gender-specific wage differentials by controlling for time-invariant factors.

Health Economics

Study: Effects of Universal Healthcare on Health Outcomes

Method: Fixed effects account for unobserved health-related characteristics and policy impacts.

Finance

Study: Determinants of Stock Returns Across Industries

Method: Fixed effects control for industry-specific factors influencing stock performance.

Takeaways:

Accounting for Rich Heterogeneity

Challenge: Complex interactions in multi-dimensional data.

Solution: High-dimensional fixed effects capture diverse unobserved factors.

Efficiently Handling Big Data

Challenge: Large datasets with numerous variables.

Solution: High-dimensional fixed effects manage high-dimensional control variables.

Uncovering Granular Insights

Challenge: Exploring nuanced effects in detailed contexts.

Solution: High-dimensional fixed effects reveal fine-grained patterns and relationships.

Proof

Proof:

$$Y = X\beta + Z\gamma + U$$

Let $\widehat{\beta}$, $\widehat{\gamma}$ denote the OLS estimates of β and γ in the above model. Since $Y = \widehat{Y} + \widehat{U}$, we can write $Y = X\widehat{\beta} + Z\widehat{\gamma} + \widehat{U}$ wlog.

The procedure in 3, regressing $M_X Y$ on $M_X Z$, means that these two take the place of the dependent variable (normally, Y) and the the regressor matrix (normally, X) in our expression for the OLS estimator:

$$\begin{aligned}
\tilde{\gamma} &= \left[(M_X Z)' M_X Z\right]^{-1} (M_X Z)' M_X Y \\
&= \left[Z' M_X' M_X Z\right]^{-1} Z' M_X' M_X Y \\
&= \left[Z' M_X Z\right]^{-1} Z' M_X Y \text{ (symmetry, idempotency)}
\end{aligned}$$

Proof – Cont'd

Now plug in our expression for Y to see what the estimator will be

$$
\begin{aligned}
\tilde{\gamma} &= \left[Z'M_XZ\right]^{-1} Z'M_XY \\
&= \left[Z'M_XZ\right]^{-1} Z'M_X\left(X\widehat{\beta} + Z\widehat{\gamma} + \widehat{U}\right) \\
&= \left[Z'M_XZ\right]^{-1} Z'\underbrace{M_XX}_{=0}\widehat{\beta} + \left[Z'M_XZ\right]^{-1} Z'M_XZ\widehat{\gamma} + \left[Z'M_XZ\right]^{-1} Z'M_X\widehat{U} \\
&= \underbrace{\left[Z'M_XZ\right]^{-1} Z'M_XZ}_{=I}\widehat{\gamma} + \left[Z'M_XZ\right]^{-1} Z'M_X\widehat{U} \\
&= \widehat{\gamma} + \left[Z'M_XZ\right]^{-1} Z'\left(I_n - X\left(X'X\right)^{-1}X'\right)\widehat{U} \\
&= \widehat{\gamma} + \left[Z'M_XZ\right]^{-1} \underbrace{Z'\widehat{U}}_{=0} - \left[Z'M_XZ\right]^{-1} Z'X\left(X'X\right)^{-1}\underbrace{X'\widehat{U}}_{=0} \\
&= \widehat{\gamma},
\end{aligned}
$$

due to FOC (normal equations) for Z and X.