A Reproducible Benchmark of Fixed Effects Estimation

Florian Oswald*

May 6, 2025

Abstract

We illustrate a workflow which tries to address several pitfalls when creating a reproducible research project. We focus on preserving raw data, documenting data sources, creating a folder structure, writing stata and R code in a way which helps to preserve the package version environment, outputting results to disk and referencing them in a final output document. As a by-product, we report timings of a typical two-way fixed effects estimation exercise on a large dataset.

1 Introduction

One could be forgiven to think that reproducibility is as simple as following a few simple steps:

1. Preserve raw data
2. Document data origin
3. Preserve code and document how to use it
4. Run everything again before submitting the package.

While this is a good start, this list if far from exhaustive, and a more complete version is available under https://datacodestandard.org. Whatever the list, however, the devil is in the details, and *in practice* achieving reproducibility is far from trivial. We want to use this fictitious research project to illustrate one potential strategy when setting up code, and data, and a few associated pitfalls.

*RES Data Editor. You can find the source code generating the entire workshop at https://github.com/floswald/ReproData.jl

2 Computational Task

In this paper, we want to estimate the following linear regression with two fixed effects:

$$y_{ij} = \beta X_{ij} + \alpha_i + \gamma_j + u_{ij} \tag{1}$$

where X_{ij} is a matrix which stacks the 1 by 7 vectors $[x_{ij1}, \ldots, x_{ij7}]$. The indices (i, j) group observations along two ad-hoc dimensions: imagine person and time, or worker and firm specific effects. Those α_i, γ_j are unobservable.

We generated the data such that the first x is a function of both fixed effects, $x_{it1} = g(\alpha_i, \gamma_t)$, the second a function only of γ_t, $x_{it2} = h(\gamma_t)$, and we set the true values for coefficients to $\beta = [3, 3, 1, 1, 1, 1, 1]$. Now let me show you the first result in table 1. Observe that models (1) and (2) exhibit bias, and only after we account for both fixed effects, we get the correct results. Overall this seems to work.[1]

Let us also have a table produced by R. We show the results in table 2 and in figure 1.

3 Timings

We found that this leads to the following result in terms of run time between stata and R, which are displayed in table 3.

[1] The interested reader may consult the data generating process here.

	(1) y	(2) y	(3) y	(4) y
x1	3.982***	3.494***	3.001***	2.998***
	(0.00108)	(0.00553)	(0.00778)	(0.00318)
x2	3.019***	3.497***	3.003***	3.001***
	(0.00151)	(0.00553)	(0.00779)	(0.00318)
x3				1.000***
				(0.000318)
x4				1.000***
				(0.000318)
x5				1.001***
				(0.000318)
x6				1.000***
				(0.000318)
x7				1.000***
				(0.000318)
Constant	0.00114	0.000955	-0.000234	-0.00164***
	(0.000775)	(0.000775)	(0.000775)	(0.000316)
FE 1	No	Yes	Yes	Yes
FE 2				
Observations	10000000	10000000	10000000	10000000

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 1: This is done with stata. I couldn't figure out why the FE2 row does not display a "yes" in columns 3 and 4. My bad, sorry!

Dependent Variable:			y	
Model:	(1)	(2)	(3)	(4)
Variables				
Constant	0.0011			
	(0.0008)			
x1	3.982***	3.494***	3.001***	2.998***
	(0.0011)	(0.0055)	(0.0078)	(0.0032)
x2	3.019***	3.497***	3.003***	3.001***
	(0.0015)	(0.0055)	(0.0078)	(0.0032)
x3				0.9996***
				(0.0003)
x4				0.9998***
				(0.0003)
x5				1.001***
				(0.0003)
x6				1.000***
				(0.0003)
x7				1.000***
				(0.0003)
Fixed-effects				
id1		Yes	Yes	Yes
id2			Yes	Yes
Fit statistics				
Observations	10,000,000	10,000,000	10,000,000	10,000,000
R^2	0.84518	0.84685	0.84698	0.97449
Within R^2		0.80515	0.02917	0.83816

*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Table 2: This is done with R.

Operation	Stata	R
CSV read	62.43	1.493
FE estimation	85.68	5.5

Table 3: Timing of operations in different languages in seconds.

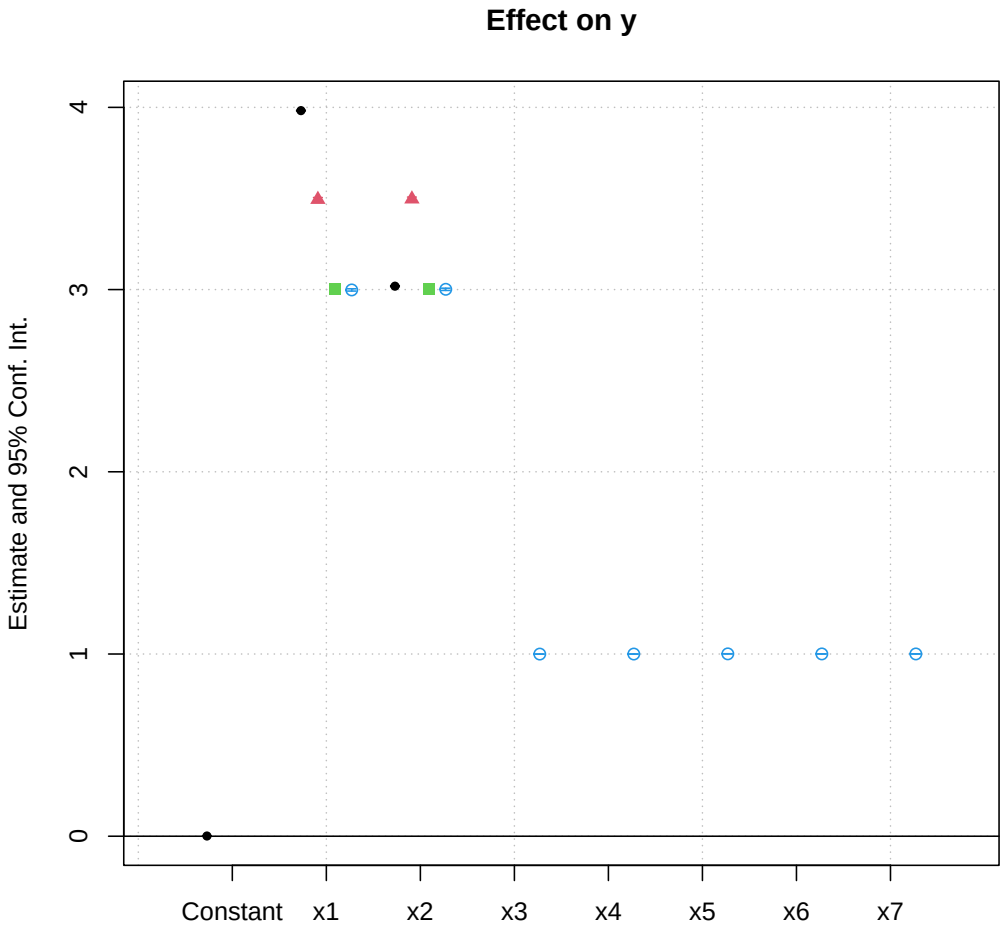

Figure 1: The coef plot corresponding to table 2.

References

Tyson Barrett, Matt Dowle, Arun Srinivasan, Jan Gorecki, Michael Chirico, Toby Hocking, Benjamin Schwendinger, and Ivan Krylov. *data.table: Extension of 'data.frame'*, 2025. URL https://CRAN.R-project.org/package=data.table. R package version 1.17.0.

Laurent Bergé. Efficient estimation of maximum likelihood models with multiple fixed-effects: the R package FENmlm. *CREA Discussion Papers*, (13), 2018.

Laurent Berge. *dreamerr: Error Handling Made Easy*, 2023. URL https://CRAN.R-project.org/package=dreamerr. R package version 1.4.0.

Laurent R Berge. *stringmagic: Character String Operations and Interpolation, Magic Edition*, 2024. URL https://CRAN.R-project.org/package=stringmagic. R package version 1.1.2.

Gergely Daróczi and Hadley Wickham. *logger: A Lightweight, Modern and Flexible Logging Utility*, 2024. URL https://daroczig.github.io/logger/. R package version 0.4.0.

Dirk Eddelbuettel. *Seamless R and C++ Integration with Rcpp*. Springer, New York, 2013. doi: 10.1007/978-1-4614-6868-4. ISBN 978-1-4614-6867-7.

Dirk Eddelbuettel and James Joseph Balamuta. Extending R with C++: A Brief Introduction to Rcpp. *The American Statistician*, 72(1):28–36, 2018. doi: 10.1080/00031305.2017.1375990.

Dirk Eddelbuettel and Romain François. Rcpp: Seamless R and C++ integration. *Journal of Statistical Software*, 40(8):1–18, 2011. doi: 10.18637/jss.v040.i08.

Dirk Eddelbuettel, Romain Francois, JJ Allaire, Kevin Ushey, Qiang Kou, Nathan Russell, Iñaki Ucar, Doug Bates, and John Chambers. *Rcpp: Seamless R and C++ Integration*, 2025. URL https://CRAN.R-project.org/package=Rcpp. R package version 1.0.14.

Paul Gilbert and Ravi Varadhan. *numDeriv: Accurate Numerical Derivatives*, 2019. URL https://CRAN.R-project.org/package=numDeriv. R package version 2016.8-1.1.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2024. URL https://www.R-project.org/.

Kevin Ushey and Hadley Wickham. *renv: Project Environments*, 2024. URL https://CRAN.R-project.org/package=renv. R package version 1.0.11.

Achim Zeileis. Econometric computing with HC and HAC covariance matrix estimators. *Journal of Statistical Software*, 11(10):1–17, 2004. doi: 10.18637/jss.v011.i10.

Achim Zeileis. Object-oriented computation of sandwich estimators. *Journal of Statistical Software*, 16(9):1–16, 2006. doi: 10.18637/jss.v016.i09.

Achim Zeileis and Yves Croissant. Extended model formulas in R: Multiple parts and multiple responses. *Journal of Statistical Software*, 34(1):1–13, 2010. doi: 10.18637/jss.v034.i01.

Achim Zeileis and Gabor Grothendieck. zoo: S3 infrastructure for regular and irregular time series. *Journal of Statistical Software*, 14(6):1–27, 2005. doi: 10.18637/jss.v014.i06.

Achim Zeileis, Susanne Köll, and Nathaniel Graham. Various versatile variances: An object-oriented implementation of clustered covariances in R. *Journal of Statistical Software*, 95(1): 1–36, 2020. doi: 10.18637/jss.v095.i01.