Universität Innsbruck

On Reproducible Econometric Research

Achim Zeileis

http://eeecon.uibk.ac.at/~zeileis/

Overview

Joint work with **Roger Koenker** (University of Urbana-Champaign).

Koenker R, Zeileis A (2009). "On Reproducible Econometric Research." *Journal of Applied Econometrics*, **24**(5), 833–847. `doi:10.1002/jae.1083`

And with **Christian Kleiber** (Universität Basel).

Zeileis A, Kleiber C (2005). "Validating Multiple Structural Change Models – A Case Study." *Journal of Applied Econometrics*, **20**(5), 685–690. `doi:10.1002/jae.856`

Kleiber C, Zeileis A (2008). *Applied Econometrics with R*. Springer-Verlag, New York. URL `http://CRAN.R-project.org/package=AER`

Kleiber C, Zeileis A (2011). "Reproducible Econometric Simulations." *Working Paper 2011-02*, Working Papers in Economics and Statistics, Research Platform Empirical and Experimental Economics, Universität Innsbruck. URL `http://EconPapers.RePEc.org/RePEc:inn:wpaper:2011-02`

Overview

- Forensic econometrics: Case studies
  - Cross-country growth regressions
  - Multiple structural change models
- Software tools
  - Version control
  - Data technologies and data archiving
  - Programming environments
  - Document preparation systems
  - Literate programming
- Challenges and conclusions

Forensic econometrics: Growth regressions

Investigation: Cross-country growth behavior based on extended Solow model.

- Durlauf and Johnson (1995, *Journal of Applied Econometrics*) extend analysis by Mankiw, Romer, Weil (1992, *The Quarterly Journal of Economics*).
- Of interest: Output (GDP per capita) growth from 1960 to 1985 for 98 non-oil-producing countries.
- Variables: Real GDP per capita; fraction of real GDP devoted to investment; population growth; fraction of population in secondary schools; and adult litercy rate.
- Data taken from MRW. DJ added literacy rate. Available as `data.dj` in JAE data archive.

Models: OLS regressions for full sample and breaks based on initial output and literacy.

Forensic econometrics: Growth regressions

Dependent variable: $\log(Y/L)_{i,1985} - \log(Y/L)_{i,1960}$.

	Full sample	$(Y/L)_{i,1960} < 1950$ $LR_{i,1960} < 54\%$	$(Y/L)_{i,1960} \geq 1950$ $LR_{i,1960} \geq 54\%$
Observations	98	42	42
Constant	3.040	1.400	0.450
	(0.831)	(1.850)	(0.723)
$\log(Y/L)_{i,1960}$	-0.289	-0.444	-0.434
	(0.062)	(0.157)	(0.085)
$\log(I/Y)_i$	0.524	0.310	0.689
	(0.087)	(0.114)	(0.170)
$\log(n+0.05)_i$	-0.505	-0.379	-0.545
	(0.288)	(0.468)	(0.283)
$\log(SCHOOL)_i$	0.233	0.209	0.114
	(0.060)	(0.094)	(0.164)

Forensic econometrics: Growth regressions

Replication: Data is available from JAE archive, and OLS regression should be trivial . . . right?

Data: Read, code missing values, and select non-oil countries.

```
R> dj <- read.table("data.dj", header = TRUE,
+    na.strings = c("-999.0", "-999.00"))
R> dj <- subset(dj, NONOIL == 1)
```

Model: R formula (converting percentages to fractions).

```
R> f1 <- I(log(GDP85) - log(GDP60)) ~ log(GDP60) +
+    log(IONY/100) + log(POPGRO/100 + 0.05) + log(SCHOOL/100)
```

Regression: OLS fit for full sample and subsamples.

```
R> mrw  <- lm(f1, data = dj)
R> sub1 <- lm(f1, data = dj, subset = GDP60 <  1950 & LIT60 <  54)
R> sub2 <- lm(f1, data = dj, subset = GDP60 >= 1950 & LIT60 >= 54)
```

Forensic econometrics: Growth regressions

Full sample results: Success! Only minor deviations.

```
R> mrw   <- lm(f1, data = dj)
R> coeftest(mrw)
```

	Durlauf & Johnson	Replication
Observations	98	98
Constant	3.040	3.022
	(0.831)	(0.827)
$\log(Y/L)_{i,1960}$	-0.289	-0.288
	(0.062)	(0.062)
$\log(I/Y)_i$	0.524	0.524
	(0.087)	(0.087)
$\log(n + 0.05)_i$	-0.505	-0.506
	(0.288)	(0.289)
$\log(SCHOOL)_i$	0.233	0.231
	(0.060)	(0.059)

Forensic econometrics: Growth regressions

Subsample results: Failure! Not even sample size is correct.

```
R> sub2  <- lm(f1, data = dj, subset = GDP60 >= 1950 & LIT60 >= 54)
R> coeftest(sub2)
```

	Durlauf & Johnson	Replication
Observations	42	39
Constant	0.450	3.952
	(0.723)	(1.337)
$\log(Y/L)_{i,1960}$	-0.434	-0.425
	(0.085)	(0.104)
$\log(I/Y)_i$	0.689	0.653
	(0.170)	(0.187)
$\log(n + 0.05)_i$	-0.545	-0.587
	(0.283)	(0.361)
$\log(SCHOOL)_i$	0.114	0.137
	(0.164)	(0.180)

Forensic econometrics: Growth regressions

Problem 1: Grid search plus educated guessing leads to different breaks.

```
R> sub2b <- lm(f1, data = dj, subset = GDP60 >= 1800 & LIT60 >= 50)
R> coeftest(sub2b)
```

	Durlauf & Johnson	Replication
Observations	42	42
Constant	0.450	4.147
	(0.723)	(1.230)
$\log(Y/L)_{i,1960}$	-0.434	-0.435
	(0.085)	(0.096)
$\log(I/Y)_i$	0.689	0.689
	(0.170)	(0.178)
$\log(n+0.05)_i$	-0.545	-0.545
	(0.283)	(0.345)
$\log(SCHOOL)_i$	0.114	0.114
	(0.164)	(0.171)

Forensic econometrics: Growth regressions

Problem 2: Population growth and schooling not fractions but percent.

```
R> sub2c <- update(sub2b, . ~ log(GDP60) +
+    log(IONY) + log(POPGRO/100 + 0.05) + log(SCHOOL))
```

	Durlauf & Johnson	Replication
Observations	42	42
Constant	0.450	0.450
	(0.723)	(0.899)
$\log(Y/L)_{i,1960}$	−0.434	−0.435
	(0.085)	(0.096)
$\log(I/Y)_i$	0.689	0.689
	(0.170)	(0.178)
$\log(n + 0.05)_i$	−0.545	−0.545
	(0.283)	(0.345)
$\log(SCHOOL)_i$	0.114	0.114
	(0.164)	(0.171)

Forensic econometrics: Growth regressions

Problem 3: Robust sandwich standard errors.

```
R> coeftest(sub2c, vcov = sandwich)
```

	Durlauf & Johnson	Replication
Observations	42	42
Constant	0.450	0.450
	(0.723)	(0.723)
$\log(Y/L)_{i,1960}$	-0.434	-0.435
	(0.085)	(0.085)
$\log(I/Y)_i$	0.689	0.689
	(0.170)	(0.170)
$\log(n + 0.05)_i$	-0.545	-0.545
	(0.283)	(0.283)
$\log(SCHOOL)_i$	0.114	0.114
	(0.164)	(0.164)

Forensic econometrics: Growth regressions

Summary:

- Cutoffs actually used did not match those indicated.
- Usage of standard errors inconsistent.
- Scaling of variables (and hence intercepts) inconsistent.
- Other models in DJ paper: Similar problems, and some inference not reproducible at all.

Implications:

- Casts doubt results. (Even though – in this case, so far – qualitative results remain unchanged.)
- Very hard to track down without original code.
- Might have been impossible for less standard models.
- *Hence:* Provide replication code even for simple things and details.

Forensic econometrics: Structural change models

Investigation: Multiple structural change model for level of US ex-post real interest rate (Jan 1961–Jul 1986).

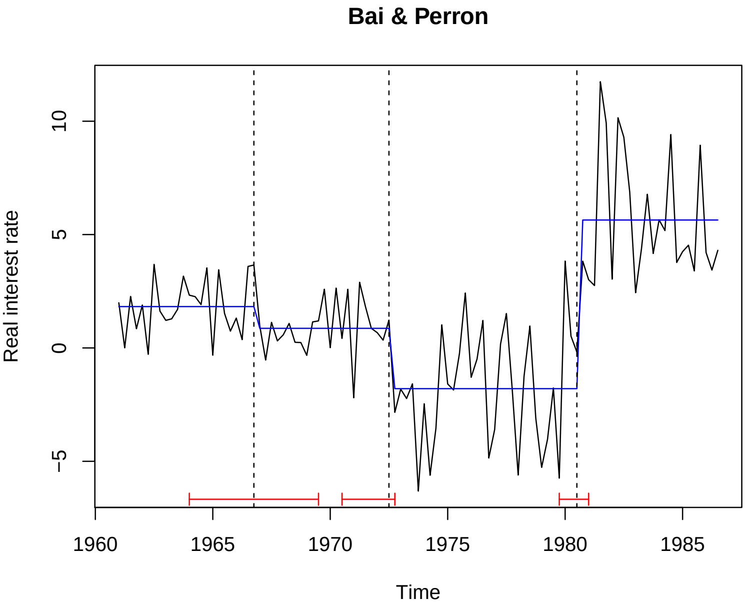
Source: Bai and Perron (2003, *Journal of Applied Econometrics*).

- Comprehensive discussion of computational aspects of multiple structural change models.
- Empirical examples, with data in JAE archive.
- GAUSS software and replication code!

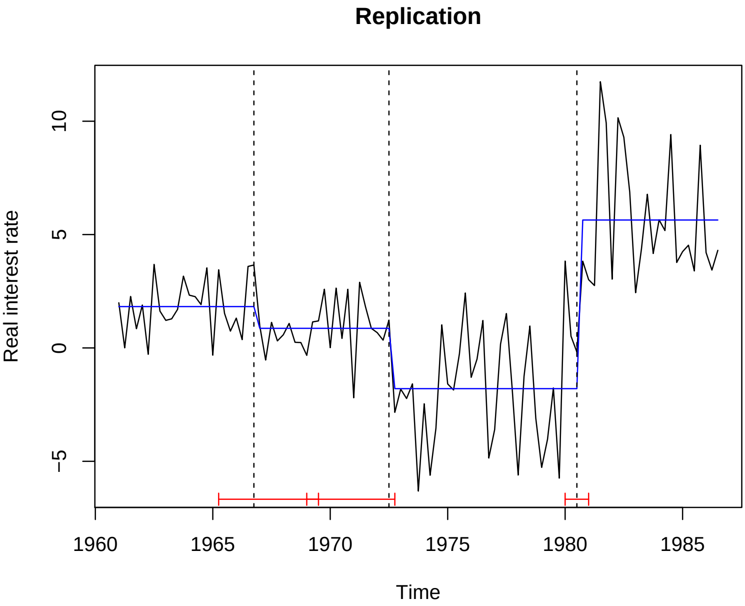
Replication:

- Re-implementation of methods in R (package *strucchange*).
- Successful replication of: Breakpoint estimates (OLS), coefficient estimates (OLS), coefficient standard errors (quadratic spectral kernel HAC with prewhitening).
- Problems: Confidence intervals of breakpoints.

Forensic econometrics: Structural change models



Forensic econometrics: Structural change models

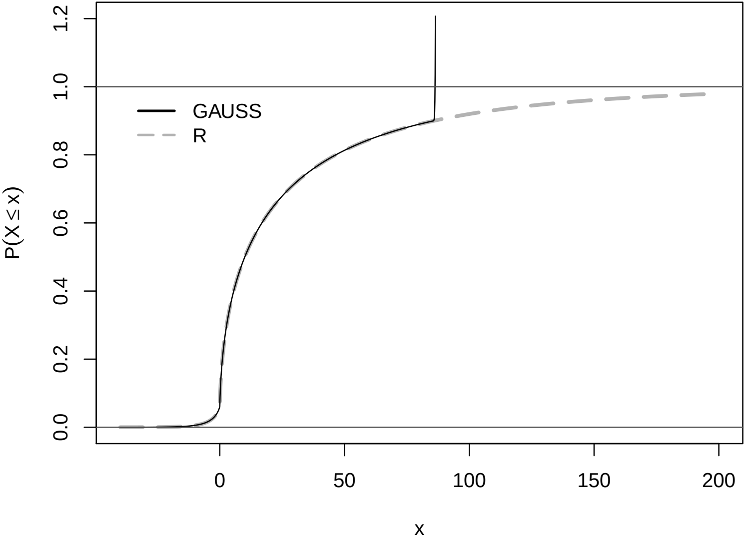


Forensic econometrics: Structural change models

What is going on?

- Computation of confidence intervals is based on asymptotic theory, leading to nonstandard distribution function.
- Quantiles need to be computed from functional of a two-sided Brownian motion with different scales and drifts.
- This also involves the term: $\exp(ax) \cdot \Phi(-b\sqrt{x})$.
- For second breakpoint: $a \approx 8.31$, $b \approx 4.08$, and $x \in [0, 300]$.
- Product of a huge and a tiny number, numerically very instable.
- Better: $\exp\{ax \ + \ \log \Phi(-b\sqrt{x})\}$ and compute $\log \Phi$ directly.
- But still: GAUSS 3.2.38 (and even up to 6.0.8) chokes on this.

Forensic econometrics: Structural change models



Forensic econometrics: Structural change models

Computation of $\Phi(y)$ and $\log(\Phi(y))$ with $y = -4.08\sqrt{x}$ in GAUSS 3.2.38 and R (all versions at least since 2.1.1).

	GAUSS	R	GAUSS	R
`x`	`cdfn(y)`	`pnorm(y)`	`lncdfn(y)`	`pnorm(y, log.p=TRUE)`
82	$4.23e-299$	$4.23e-299$	-687.03	-687.03
84	$2.46e-306$	$2.46e-306$	-703.69	-703.69
86	$2.23e-308$	0	-720.35	-720.35
88	$2.23e-308$	0	-737.01	-737.01
90	$2.23e-308$	0	$-\infty$	-753.66

Forensic econometrics: Structural change models

Conclusions:

- Be careful about numerical precision of your own code ...
- ... and also the functions of your programming environment.
- Replication would not have been successful without access to GAUSS code of Bai and Perron.

Epilogue:

- Aptech fixed `lncdfn()` in recent versions of GAUSS (after initial private e-mails to us claiming that our computations were wrong).
- Stata published a very good and openly available C implementation for normal log-probabilities.

Software tools

Typically: An econometric analysis encompasses the following.

- Data handling.
- Data analysis in some programming environment.
- Document preparation with results of the analysis.

Question: Which software can assist the researcher in making such an analysis reproducible?

Software tools: Version control

Often: Research is carried out

- over an extended period,
- by several authors,
- on several computers,
- and hence difficult to reconstruct *exactly*.

Problem: Files proliferate with inconsistent naming conventions, get overwritten or deleted or are ultimately archived upon paper acceptance . . . or next disk crash.

Idea: Employ version control tools.

- Only one current version of each file.
- But full history of all changes in database.
- Annotate changes in log files.
- Enable moving back and forth through revisions.

Software tools: Version control

Work flow:

- Initially, check out a repository of files.
- Subsequently, easily check out updates by other authors.
- Work on files and commit own changes.
- All changes, additions, removals stored in repository.

Software:

- Popularized through internet and open software development.
- Various packages available: CVS, SVN, Git, Mercurial, . . .
- Probably most popular for small to medium sized projects: *Subversion (SVN)*.
- Only “diffs” stored in each revision.
- On Windows: TortoiseSVN integrates with Explorer.

Software tools: Data technologies and data archiving

Typically:

- Data is *not* extremely large or complex.
- Flat plain text file ideal for reproducibility: Portable, easy to store and access.

Furthermore:

- Relational database management systems for complex data, e.g., open-source systems PostgreSQL or MySQL.
- New standards for web-based sharing of data, e.g., XML or PHP.

Software tools: Programming environments

In general: Many aspects drive choice of programming environment.

Here: Focus on aspects directly relevant to reproducibility.

Desirable:

- Command line interface (CLI) or at least script from some graphical user interface (GUI).
- Modular code that is easy to read, encapsulates conceptual tasks, and is reusable in other settings.
- Open sources to enable gradual refinement.

Software tools: Document preparation systems

Two approaches:

- WYSIWYG (what you see is what you get) text processors (e.g., Microsoft Word, OpenOffice.org, LibreOffice, . . .).
- Markup languages (like $\LaTeX$, HTML, . . .).

Again: Focus on reproducibility.

- Stable open standards preferred.
- Proprietary binary formats (such as Microsoft Word) problematic.
- Flat text files ideal for combination with version control.

Software tools: Literate programming

Idea: Merge text, documentation, and computer code to facilitate keeping everything in sync.

Literate programming:

- Single file contains documentation and computer code.
- *Tangling:* Extract computer code.
- *Weaving:* Produce documentation that optionally shows or hides the code.

Literate data analysis:

- Extend weaving step: Execute code to produce all numeric output, tables, figures, etc.
- `Sweave` in R: Combines R code with LaTeX (or HTML, ODF, . . .).
- Results in “dynamic” or “revivable” documents.

Challenges and conclusions

Real challenge: Better incentives from journals and funding agencies for archiving and distribution of details underlying empirical and/or computational work.

Goal: Convince authors that providing such details will enhance the chances for publication and citation of their work.

However: Advances in (open) software make it relatively easy to enhance reproducibility without too much extra effort.

At eeecon: New server and support by the admins provide web-based services, in particular web space, working paper series, SVN, …