

# Reproducible econometrics

Mahmood Arai and Michael Lundholm

October 16, 2009

# Outline

1. Terminology: Reproduction vs. replication
2. Econometric research has not been reproducible!
3. Levels of reproduction
4. Technologies for reproduction
  - Data
  - Code documentation
  - Weaving and tangling
  - DOCSTRIP
  - Version control software
5. Cost–benefit considerations

# 1. Reproduction vs. replication

Hammermesh [6]:

**Reproduction** There exists data and codes such that published results can be duplicated. (Hammermesh: Pure replication)

“It attempts at replication that check whether a genuine advance in knowledge has been made or a puzzle encountered, or whether either mistake or fraud lies behind the results.” O’Brian [15, pp. 262f]

**Replication** Using new data (other samples, other populations) and existing methods or existing data and new methods to reproduce published results. (Hammermesh: Scientific replication)

Terminology is a mess: Terms like replication and reproduction are sometimes used as synonyms, sometimes with qualifying adjectives, sometimes they are used for distinctly different concepts.

## 2. Econometric research has not been reproducible!

Two “metastudies” from AER (1986 and 2003):

**1986: Dewald, Thursby & Anderson [1]** Published articles in JMCB 1980–1982

- took on average 217 days for authors to respond to a request for data
- 66% of authors did not submit data for reproduction
- the first 54 submitted data sets (published and accepted papers) were checked whether they would allow reproduction; only 8 did

**2003: McCullough & Vinod [14]** Attempted to replicate all (8) articles in the June 1999 AER issue

- 4 authors submitted organised data and code
- 2 authors provided neither data or code (one had lost the files, one would do it “next semester”)
- 1 author submitted 400 undocumented files after several months
- 1 author submitted numerous data files which would not run with his code

## 2. Econometric research is not reproducible! (cont.)

**Example 1 (JPE):** Feldstein [3, 4] vs. Leimer & Lesnoy [10]

- Claim: Serious coding error
- Response: “I am embarrassed by the programming error that Leimer & Lesnoy uncovered . . .”

## 2. Econometric research is not reproducible! (cont.)

**Example 2 (AER):** Hoxby [7, 8] vs. Rothstein [16]

- Claim: “Despite several requests, Hoxby has not provided the precise data set from which her published results were derived. She has, however, made available a corrected data set . . .”
- Response: “Rothstein’s claim . . . is a misinterpretation. Rather, the original data set do not exist . . .”

<http://www.princeton.edu/~jrothst/replication/hoxbydocumentation/index.html>

## 2. Econometric research is not reproducible! (cont.)

**Example 3 (AER):** Levitt [11, 12] vs. McCrary [13]

- Claim: (i) Levitt supplied code and data which did not replicate published results, (ii) coding error which reverses the result (iii) no source for timing of elections made impossible to reconstruct the major instrument in the paper
- Response: Levitt did not provide the source of the instrument and could not document his “primary innovation”

## 2. Econometric research is not reproducible! (cont.)

**Example 4 (QJE):** Donohue & Levitt [2] vs. Foote & Goetz [5]

- Claim: “First . . . a coding mistake in the concluding regressions, which identify abortion’s effect on crime by comparing the experiences of different age cohorts within the same state and year. Second, correcting this error and using a more appropriate per capita specification for the crime variable generates much weaker results. Third, earlier tests in the paper, are not robust to allowing differential state trends based on statewide crime rates that pre-date the period when abortion could have had a causal effect on crime.”
- Response: “While embarrassing because it pointed out that the wrong numbers got into one of our tables, it doesn’t offer a fundamental challenge to our original findings. When you measure abortion more carefully – in ways that we showed back in our 2004 paper – the results are as strong or stronger than ever.”

# 3. Levels of reproduction

What can I reproduce?

**Core** Documented code and data to reproduce estimations

**Level 1** Documented code to reproduce graphs etc

**Level 2** Documented code to reproduce the complete article  
with merged text, inline statistics graphs and tables as well as software  
versions (weaving/tangling)

**Level 3** Documented code to reproduce the complete set of project  
documents (articles, presentations, technical documentations etc)  
such that there is an internal consistency regarding estimations  
and reported numbers (DOCSTRIP)

**Level 4** Documented code to reproduce the entire history of codes for the  
complete set of project documents (version control)

# 4. Technologies for reproduction: Data

“Enough information should be presented so that the reader could, in principle, obtain the data and redo your analysis. In particular, all public data sources should be included in the references, and short data sets can be listed in an appendix.” Woolridge [[17](#), p. 682]

# 4. Technologies for reproduction: Code comments

**Defining a variable; two different implementations of the same instruction:** *English spoken at home (older)*

Implementation 1:

“Dummy variable taking the value one if English is the language normally spoken at home by the respondent to members of the family who are older.”

```
1 gen esold=0;  
2 replace esold=1 if __s12f1==1 ;
```

# 4. Technologies for reproduction: Code comments (cont)

## Implementation 2:

“There are several language variables measuring whether the respondent is speaking English with different individuals; at home with older, at home with younger, at work and with friends. We construct these variables using question s12a which asks whether the respondent regularly speak to anyone in Britain in any other language than English and ..s12f, ..s12g, ..s12h and ..s12i which asks which language is spoken to the above mentioned categories of individuals. Each of ..s12f, ..s12g, ..s12h and ..s12i comes in 18 versions (e.g., ..s12f1,...,..s12f18) where each question 1–15 is coded yes if the respondent speaks the language. Question 16 is “Never speaks to these people/Not ap[plicable]”, question 17 NA and question 18 “None of the above answered positive”. Question 1 is always regarding English.

The respondent is coded as English speaker if either s12a is answered negatively or ..s12aX1, where X=f,g,h,i, is answered positively. Only respondents for which either s12a is NA or all of ..s12X1,...,..s12X16 are answered negatively are coded as NA. Below is the code for *English Spoken At Home With Older*:

# 4. Technologies for reproduction: Code comments (cont)

Implementation 2 (cont):

```
1 is.na(U$s12a) <- U$s12a== 8 | U$s12a == 9
2 s12flist <- c(paste("..s12f", 2:16, sep=""))
3 U$oOLD <- apply(U[s12flist]=="yes", 1, sum)
4 U$English.Spoken.at.Home.with.Older <-
5   ((U$s12a==1 | is.na(U$s12a)) &
6   U$..s12f1=="yes") | U$s12a==2
7 is.na(U$English.Spoken.at.Home.with.Older) <-
8   U$oOLD==0 &
9   U$English.Spoken.at.Home.with.Older==FALSE
10 U$English.Spoken.at.Home.with.Older <-
11   replace(U$English.Spoken.at.Home.with.Older,
12   U$oOLD>0 &
13   is.na(U$English.Spoken.at.Home.with.Older),FALSE)
14 U$DO.NOT.SPEAK.WITH.OLDER <- ifelse(U$..s12f16=="no",0,1)
```

# 4. Technologies for reproduction: Weaving/tangling

**Weaving** Code and code documentation are weaved in documents readable to humans (but in modern applications also to econometric software)

**Tangling** Code is extracted to a document readable to the econometric software  
Code should be understandable and

Implementations of *literate programming*

# 4. Technologies for reproduction: Weaving/tangling

“The practitioner of literate programming can be regarded as an essayist, whose main concern is with exposition and excellence in style. Such an author, with thesaurus in hand, chooses the names of variables carefully and explains what each variable mean.” Knuth [9, p. 97].

- Code should be understandable and readable to humans
- Code is marked up in *code chunks* and comments in *text chunks* using the so called `noweb` markup syntax

## 4. Technologies for reproduction: Weaving/tangling (cont.)

Software integrating various econometric and typesetting software exists:

- Sweave: R with  $\text{\LaTeX}$
- odfWeave: R with Open Document Format (Open Office)
- SASWeave: SAS with  $\text{\LaTeX}$
- StatWeave; Stata, R and SAS with  $\text{\LaTeX}$  and Open Document Format (Open Office)

# 4. Technologies for reproduction: Weaving/tangling (cont.)

Example where the R function Sweave produces L<sup>A</sup>T<sub>E</sub>X-code:

```
1 \section{Which came first; Chicken or Egg?}
2 Thurman and Fischer (1988) test which came first, the chicken or the egg? Their data is available
3 as the data set \code{ChickEgg} in package \code{lmtest}. We reproduce their results by implementing the
4 \code{grangertest} function in the same library using lag order  $\$ \input{lagorder.txt}$ .
5 <<echo=FALSE>>=
6 library(lmtest)
7 lagorder <- 4
8 test1 <- grangertest(chicken~egg,data=ChickEgg,order=lagorder)
9 test2 <- grangertest(egg~chicken,data=ChickEgg,order=lagorder)
10 write(lagorder,file="lagorder.txt")
11 @
12 The null that eggs do not Granger cause chicken is rejected ( $p=\$ \Sexpr{round(test1$`Pr(>F)`[2],3)}$ ),
13 but not the null that chicken do not Granger cause eggs ( $p=\$ \Sexpr{round(test2$`Pr(>F)`[2],3)}$ ). In
14 the latter case the complete results are as in Table~\ref{tab:chick}.
15
16 <<echo=FALSE,results=tex>>=
17 library(xtable)
18 print(xtable(test2,caption="Egg vs. chicken.\\label{tab:chick}"))
19 @
20 Hence, the egg came first!
```

# 4. Technologies for reproduction: Weaving/tangling (cont.)

```
1 \section{Which came first; Chicken or Egg?}
2 Thurman and Fischer (1988) test which came first, the chicken or the egg? Their data is available
3 as the data set {ChickEgg} in package {lmtest}. We reproduce their results by implementing the
4 {grangertest} function in the same library using lag order  $\text{\input{lagorder.txt}}$ .
5 The null that eggs do not Granger cause chicken is rejected ( $p=0.006$ ),
6 but not the null that chicken do not Granger cause eggs ( $p=0.813$ ). In
7 the latter case the complete results are as in Table~\ref{tab:chick}.
8
9 % latex table generated in R 2.9.2 by xtable 1.5-5 package
10 % Wed Oct 14 09:14:10 2009
11 \begin{table}[ht]
12 \begin{center}
13 \begin{tabular}{lrrrr}
14 \hline
15 & Res.Df & Df & F & Pr(>F) \\ \hline
16 1 & 41 & & & \\ \hline
17 2 & 45 & -4 & 0.39 & 0.8125 \\ \hline
18 \end{tabular}
19 \end{center}
20 \caption{Egg vs. chicken.\label{tab:chick}}
21 \end{table}Hence, the egg came first!
```

# 4. Technologies for reproduction: Weaving/tangling (cont.)

## 1 Which came first; Chicken or Egg?

Thurman and Fischer (1988) test which came first, the chicken or the egg? Their data is available as the data set `ChickEgg` in package `lmtest`. We reproduce their results by implementing the `grangertest` function in the same library using lag order 4. The null that eggs do not Granger cause chicken is rejected ( $p = 0.006$ ), but not the null that chicken do not Granger cause eggs ( $p = 0.813$ ). In the latter case the complete results are as in Table 1.

|   | Res.Df | Df | F    | Pr(>F) |
|---|--------|----|------|--------|
| 1 | 41     |    |      |        |
| 2 | 45     | -4 | 0.39 | 0.8125 |

Table 1: Egg vs. chicken.

Hence, the egg came first!

# 4. Technologies for reproduction: DOCSTRIP

- DOCSTRIP is the literate programming environment of  $\text{\LaTeX}$
- Designed to weave documentation and code of  $\text{\LaTeX}$  document classes and packages
- Material in files are tagged
- Different tags in one or more files can be exported to selected files (one tag for the article, one for the presentation etc)

# 4. Technologies for reproduction: revision control

- Concurrent Version Systems (CVS)
- Subversion (SVN)
- Git (a distributed version control system)

# 5. Cost–benefit considerations

What are the cost and benefits of use “reproducible econometrics”?

- Costs
  - Invest time in new methods
  - Programming becomes more complex
- Benefits
  - Easier and faster to replicate
  - Save time when variable definitions and specifications are changed

## References

- [1] W. G. Dewald, J. G. Thursby, and R. G Anderson. Replication in empirical economics: The *journal of money, credit and banking project*. *American Economic Review*, 76:557–603, 1986.
- [2] J. J. Donohue and S. D. Levitt. The impact of legalized abortion on crime. *Quarterly Journal of Economics*, 116:379–420, 2001.
- [3] M. Feldstein. Social security, induced retirement and aggregate capital accumulation. *Journal of Political Economy*, 82:905–926, 1976.
- [4] M. Feldstein. Social security and private saving: Reply. *Journal of Political Economy*, 90:630–642, 1982.
- [5] C. L. Foote and C. F. Goetz. The impact of legalized abortion on crime: Comment. *Quarterly Journal of Economics*, 123:407–23, 2008.
- [6] D. S. Hammermesh. Viewpoint: Replication in economics. *Canadian Journal of Economics*, 40:715–733, 2007.
- [7] C. Hoxby. Does competition among public schools benefit students and taxpayers? *American Economic Review*, 90:1209–1238, 2000.
- [8] C. Hoxby. Does competition among public schools benefit students and taxpayers? A reply to Rothstein. NBER Working Paper Series, Working Paper 11216, 2005.
- [9] D. E. Knuth. Literate programming. *The Computer Journal*, 27:97–111, 1984.
- [10] D. Leimer and S. Lesnoy. Social security and private saving. *Journal of Political Economy*, 90:606–629, 1982.
- [11] S. D. Levitt. Using electoral cycles in police hiring to estimate the effect of police on crime. *American Economic Review*, 87:1230–1250, 1997.
- [12] S. D. Levitt. Using electoral cycles in police hiring to estimate the effect of police on crime: Reply. *American Economic Review*, 92:1244–1250, 2002.
- [13] J. McCrary. Using electoral cycles in police hiring to estimate the effect of police on crime: A reply. *American Economic Review*, 92:1236–1243, 2002.
- [14] B. D. McCullough and H. D. Vinod. Verifying the solution from a nonlinear solver: A case study. *American Economic Review*, 93:557–603, 2003.
- [15] D. P. O’Brian. Economists and data. *British Journal of Industrial Relations*, 30:253–285, 1992.
- [16] J. Rothstein. Does competition among public schools benefit students and taxpayers? A comment on Hoxby (2000). NBER Working Paper Series, Working Paper 11216, 2005.
- [17] J. M. Woolridge. *Introductory econometrics*. South–Western, 4 edition.