Facilitating Reproducibility:
Open Data and Code in Economics

Victoria Stodden
Department of Statistics
Columbia University

Open Economics International Workshop
Dec 17, 2012
An Illustrious History

• Diewald, Thursby, and Anderson, “Replication in Empirical Economics: The Journal of Money Credit and Banking Project,” AER 78(4), Sep 1986

• two year study to replicate published results - “The confirmation of research findings through replication by other researchers is an essential part of scientific methodology.”

• code/data requested for all papers published in 1980-1986 - “We note that NSF Policy Number 754.2 requires that computer programs and data which have been produced with the assistance of NSF grants be made available to other researchers either by publication, duplication, or loan to the researcher.”
<table>
<thead>
<tr>
<th></th>
<th>Published before Data Requested</th>
<th>Accepted before Data Requested</th>
<th>Under Review when Data Requested</th>
</tr>
</thead>
<tbody>
<tr>
<td>Requests</td>
<td>62</td>
<td>27</td>
<td>65</td>
</tr>
<tr>
<td>Responses</td>
<td>42</td>
<td>26</td>
<td>49</td>
</tr>
<tr>
<td>Response Rate (Percent)</td>
<td>66</td>
<td>96</td>
<td>75</td>
</tr>
<tr>
<td>Mean Response Time (Days)</td>
<td>217</td>
<td>125</td>
<td>130</td>
</tr>
<tr>
<td>Not Submitted:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Confidential Data</td>
<td>2</td>
<td>1(^b)</td>
<td>0</td>
</tr>
<tr>
<td>Lost or Destroyed Data</td>
<td>14</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Data Available, But Not Sent(^c)</td>
<td>4</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Nonrespondents</td>
<td>20</td>
<td>1</td>
<td>16</td>
</tr>
<tr>
<td>Total Not Submitted</td>
<td>40</td>
<td>6</td>
<td>18</td>
</tr>
<tr>
<td>Nonsubmission Rate (Percent)</td>
<td>66</td>
<td>22</td>
<td>28</td>
</tr>
</tbody>
</table>

\(^a\)Includes all requests made through December 1984, and excludes authors whose papers were rejected.

\(^b\)Two data sets were partially confidential.

\(^c\)This category includes authors who (i) stated that their data were available from published sources, but did not send their data; and (ii) authors who claimed to have their data but were unwilling to sort through their papers to find the data.
Journal Policy Changes

• “Authors of accepted papers that contain empirical work, simulations, or experimental work must provide to the Review, prior to publication, the data, programs, and other details of the computations sufficient to permit replication. These will be posted on the AER Web site.”

RunMyCode.org

The concept

As simple as 1, 2, 3:
1. A researcher has an idea.
2. The researcher writes a paper based on this idea.
3. Using RunMyCode, the researcher creates a companion website associated with this paper. The companion website allows people to implement the methodology presented in the paper.

1. Idea
2. Paper
3. Companion website

Create your own companion website >>
The Companion Page

Copula-Based Models for Financial Time Series

By Andrew J. Patton

Coder:
Andrew J. Patton
Duke University
United States
Coder Page

The code estimates a dozen constant and time-varying copula functions for bivariate time-series (e.g. Normal, Clayton, Rotated Clayton, Hackett, Frank, Gumbel, Rotated Gumbel, Student, Symmetrized Joe-Clayton). These copulas are then compared by relying on criteria such as Log-likelihood, AIC or BIC. Besides, the code reports the plots for exceedance correlations, quantile dependence and the graphical comparison of the constant and the time-varying versions of three copulas, i.e. Normal, Gumbel and SJC. For the constant copulas, the level of tail dependence (Joe and Ubeda) is also indicated.

Click for download

Number of volatility frequencies

Starting values for optimization (optional)

runmycode
RunMyCode.org

• inform research on sharing, scientific transparency, impact of computation on discovery and validation:
  ▸ facilitate code and data sharing, alongside published articles,
  ▸ longevity and persistence of digital scholarly objects - 10 year guarantee (via partnerships) including metadata,
  ▸ recognize data, code, and reimplementation contributions,
  ▸ execution of code in the cloud, or locally,
  ▸ public interaction/access, community engagement, large scale validation, acceleration of discoveries,
  ▸ understand the data lifecycle, reuse, best practices.
Reproducible Research Standard

The *Reproducible Research Standard (RRS)* (Stodden, 2009)

- A suite of license recommendations for computational science:
  - Release media components (text, figures) under CC BY,
  - Release code components under Modified BSD or similar,
  - Release data to public domain or attach attribution license.

➡ Remove copyright’s barrier to reproducible research and,

➡ Realign the IP framework with longstanding scientific norms.

Winner of the Access to Knowledge Kaltura Award 2008
References

• “The Scientific Method in Practice: Reproducibility in the Computational Sciences”

• “Open Science: Policy Implications for the Evolving Phenomenon of User-led Scientific Innovation”

• “Enabling Reproducible Research: Open Licensing for Scientific Innovation”

• Reproducible Research: Tools and Strategies for Scientific Computing, July 2011

• Reproducible Research in Computational Science: What, Why and How, Community Forum, July 2011

available at http://www.stodden.net
Tools for Computational Science

- **Dissemination Platforms:**
  - RunMyCode.org
  - MLOSS.org
  - IPOL
  - Madagascar
  - MLOSS.org
  - thedatahub.org
  - Open Science Framework

- **Workflow Tracking and Research Environments:**
  - VisTrails
  - Kepler
  - Galaxy
  - GenePattern
  - Sumatra
  - Taverna
  - CDE
  - Paper Mâché
  - Pegasus

- **Embedded Publishing:**
  - Verifiable Computational Research
  - Collage Authoring Environment
  - Sweave
  - SHARE
## Credibility Crisis

<table>
<thead>
<tr>
<th>JASA June</th>
<th>Computational Articles</th>
<th>Code Publicly Available</th>
</tr>
</thead>
<tbody>
<tr>
<td>1996</td>
<td>9 of 20</td>
<td>0%</td>
</tr>
<tr>
<td>2006</td>
<td>33 of 35</td>
<td>9%</td>
</tr>
<tr>
<td>2009</td>
<td>32 of 32</td>
<td>16%</td>
</tr>
<tr>
<td>2011</td>
<td>29 of 29</td>
<td>21%</td>
</tr>
</tbody>
</table>

Generally, data and code not made available at the time of publication, insufficient information captured in the publication for verification, replication of results.

⇒ A Credibility Crisis