

“Econometric Computing with R”

B. D. McCullough

Departments of Decision Sciences and Economics
Drexel University

Introduction

“Econometric Computing” is just “statistical computing for econometrics”.

The Royal Statistical Society distinguishes between “statistical computing” and “computational statistics”:

- ▶ statistical computing - the development and application of statistical methods that rely on computation;
- ▶ computational statistics - the algorithms and software that underlie statistical methods.

Statistical Computing: numerical analysis for statistics.

Computational Statistics: computationally intensive methods for performing statistical analysis, *e.g.*, Monte Carlo, bootstrap.

Introduction

“The mathematical definition of a function, a computing formula, and a practical algorithm for its evaluation on a computer are all very different.” (p. 153)

John F. Monahan
Numerical Methods of Statistics
Cambridge Univ Press, 2001

Introduction

Why is econometric computing important?

Statistics has a long tradition of emphasizing accuracy; economics doesn't.

In general, most users don't know enough to avoid obvious mistakes.

- ▶ solving for $\hat{\beta}$ as $b = \text{inv}(X'X) * X'y$
- ▶ writing $\log(\text{norm}(x))$ instead of $\text{lognorm}(x)$
- ▶ taking $x**0.5$ instead of $\text{sqrt}(x)$

Introduction

Many times, neither do developers.

- ▶ $s^2 = \frac{\sum x_i^2 - n\bar{x}^2}{n-1}$ instead of $s^2 = \frac{(\sum x_i - \bar{x})^2}{n-1}$
- ▶ not offering `lognorm(x)`
- ▶ using bad algorithms – correlation, least squares, nls, distributions, RNGs.....

It's clear that many developers have never read a book on statistical computing or numerical analysis.

Exhibit One: algebra

Does it make a difference whether you code

$$y = x/2 + z/2$$

or

$$y = (x + z)/2$$

????

Think before you answer.

Exhibit One: algebra

The “Santa Fe Stock Market” (SFSM) by LeBaron *et al.* (1999) is a popular agent-based model that is used to generate “realistic” stock prices and volumes.

LeBaron, Blake, W. B. Arthur and R. Palmer (1999), “Time Series Properties of an Artificial Stock Market,” *Journal of Economic Dynamics & Control* **23**, 1487-1516

Polhill, Gary J., Luis R. Izquierdo and Nicholas M. Gotts (2005), “The Ghost in the Model (and Other Effects of Floating Point Arithmetic),” *Journal of Artificial Societies and Social Simulation* **8**(1)

Exhibit One: algebra

$$demand = -\left(\frac{trialprice * intratep1 - forecast}{divisor} + position\right),$$

which Polhill, *et al.* (2005) called the “Baseline Version”.

They rewrote this equation as “Version 1”:

$$demand = -\left(\frac{trialprice * intratep1}{divisor} - \frac{forecast}{divisor} + position\right).$$

Exhibit One: algebra

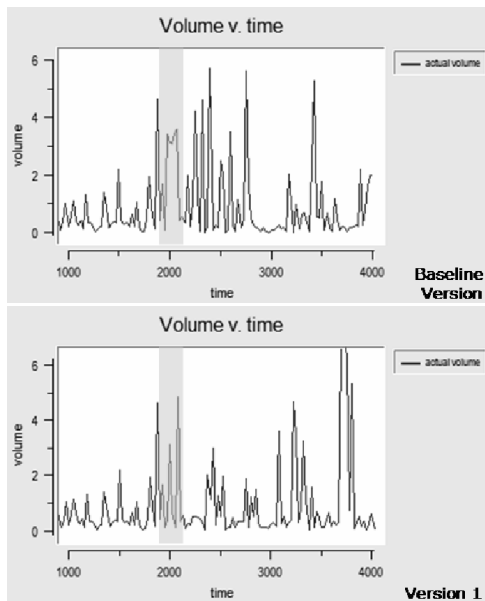


Exhibit Two: algorithms

Cointegration. Let x be a random walk. It is *integrated of order one* (the first difference is stationary).

Let y be a random walk.

If x and y cannot wander too far from each other, the x and y are said to be *co-integrated*.

Exhibit Two: algorithms

In the estimation of co-integrated systems, the likelihood is maximized by solving a generalized eigenvalue problem:

$$|\lambda S_{11} - S_{10} S_{00}^{-1} S_{01}| = 0$$

How would you solve this? Think for a moment.

Does the name “Cholesky” mean anything?

Exhibit Two: algorithms

$$\mathbf{y}_t = (y_{1t}, y_{2t}); y_{2t} = y_{1t} + u_t 10^{-m}; u_t \sim N(0, 1)$$

Alg.	$m = 3$	$m = 5$	$m = 10$
Cholesky-1	0.350919 40555	0.05949553867	failed
Cholesky-2	0.350919 40557	0.13273076746	failed
QR-1	0.35091938503	0.350919385 42	0.01335798065
QR-2	0.35091938503	0.350919385 40	failed
SVD	0.35091938503	0.350919384 94	0.00487801748

Table: Largest Eigenvalue of System

Doornik, J. A. and O'Brien, R. J. (2002), "Numerically Stable Cointegration Analysis," *Computational Statistics and Data Analysis* **41** 185-193

Exhibit Two: algorithms

Let X be the Hilbert matrix of dimension 12. Consider $XX^{-1} = I$.
Diagonal of I when X^{-1} computed using:

	Cholesky	SVD
1	1.0000002	1.0000000
2	0.9999807	1.0000012
3	1.0005020	0.9999702
4	0.9944312	1.0002072
5	1.0383301	0.9990234
6	0.8465960	0.9966422
7	1.3649726	0.9946361
8	0.3687544	0.9902124
9	1.6553078	1.0111723
10	0.5280704	1.0308642
11	1.1820953	1.0003664
12	0.9688452	1.0012815

Exhibit Three: Textbooks

Most econometrics texts convey the impression that one software package is as good as another, and that all are equally good.

Some “econometricians” do not even think it noteworthy when two packages give different answers to the same OLS problem.

Exhibit Three: Textbooks

Essentials of Econometrics, 4e
Damodar N. Gujarati and Dawn C. Porter

The equation is

$$CLFPR = \beta_0 + \beta_1 CUNR + \beta_2 AHE82 + \epsilon$$

28 observations, 1980-2007 from *Economic Report of the President, 2008*.

Condition number is 14,000.

Exhibit Three: Textbooks

package	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$
EViews/Excel	81.22673	0.638362	-1.444883
Minitab/Stata	81.286	0.63877	-1.4521
R (correct)	81.30500	0.63846	-1.45477

Table: Gujarati/Porter results

If Gujarati/Porter assume one is correct, then the other is accurate to two digits: shouldn't they warn students?

Three Exhibits

1. Algebra
2. Eigenvalues
3. Textbooks

The economics profession is in dire need of econometric computing!

How Can “R” Help?

The usual approach (FORTRAN, numerical analysis) won't work.

The “programming language” in the econometrics package won't work.

- ▶ These “languages” often are an afterthought; unwieldy.
- ▶ Lack features, *e.g.*, offer maybe a single (unspecified) matrix inversion routine.
- ▶ Gauss/Matlab? Even if they were correct, try getting source code; cost.

How Can “R” Help?

Let's see how easy it is to solve a standard statistical computing problems using R.

An Example: Longley's Data

y	x1	x2	x3	x4	x5	x6
60323	83.0	234289	2356	1590	107608	1947
61122	88.5	259426	2325	1456	108632	1948
60171	88.2	258054	3682	1616	109773	1949
61187	89.5	284599	3351	1650	110929	1950
63221	96.2	328975	2099	3099	112075	1951
63639	98.1	346999	1932	3594	113270	1952
64989	99.0	365385	1870	3547	115094	1953
63761	100.0	363112	3578	3350	116219	1954
66019	101.2	397469	2904	3048	117388	1955
67857	104.6	419180	2822	2857	118734	1956
68169	108.4	442769	2936	2798	120445	1957
66513	110.8	444546	4681	2637	121950	1958
68655	112.6	482704	3813	2552	123366	1959
69564	114.2	502601	3931	2514	125368	1960
69331	115.7	518173	4806	2572	127852	1961
70551	116.9	554894	4007	2827	130081	1962

An Example: Linear Regression

The “condition number” (κ) of the Longley data is about 15000 ($\approx 10^4$).

Folk Theorem: if the data in y and X are accurate to about s digits and $\kappa(X) \approx 10^t$, then the *computed solution* is accurate to about $s - t$ digits.

If we work in single precision, then if we're lucky we can calculate coefficients accurately to about 3 digits.

An Example: Linear Regression

But even when we use double precision on the Longley data and calculate the coefficients correctly to 10 digits, all those coefficients are meaningless!

An Example: Linear Regression

Hammarling, Sven (2005), “An Introduction to the Quality of Computed Solutions” in *Accuracy and Reliability in Scientific Computing* SIAM, Philadelphia, USA, pp. 43-76

Because of finite precision, we do not actually calculate with y and X but with $y^* = y + \delta y$ and $X^* = X + \delta X$.

Because of rounding error we do not actually calculate \hat{b} but instead some \hat{b}^* .

An Example: Linear Regression

What do we want from a linear regression?

We either want to predict y or estimate b .

If the residual e is small then we know that \hat{b}^* solves a problem that is close to $y = Xb + \epsilon$.

But small residuals do not guarantee that \hat{b}^* is close to \hat{b} (the value we would calculate with infinite precision).

An Example: Linear Regression

The typical use of the condition number:

$$\frac{\|\hat{\mathbf{b}} - \hat{\mathbf{b}}^*\|}{\|\hat{\mathbf{b}}\|} \leq \kappa \frac{\|\hat{\mathbf{y}} - \hat{\mathbf{y}}^*\|}{\|\hat{\mathbf{y}}\|}$$

Or something a little more refined (Monahan Eq. 5.2.3):

$$\frac{\|\hat{\mathbf{b}} - \hat{\mathbf{b}}^*\|}{\|\hat{\mathbf{b}}\|} \leq 2\kappa \frac{\|P_X E\|}{\|X\|} + 4\kappa^2 \frac{\|(I - P_X)E\| \|\hat{\mathbf{e}}\|}{\|X\| \|\hat{\mathbf{y}}\|} + 8\kappa^3 \frac{\|(I - P_X)E\|^2}{\|X\|^2}$$

An Example: Linear Regression

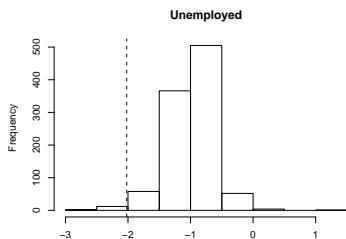
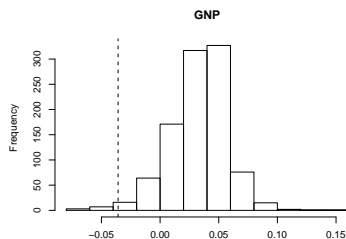
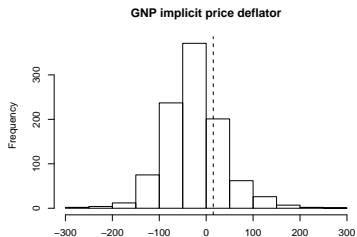
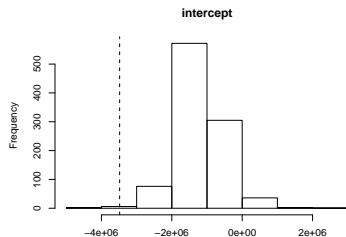
Why are well-calculated Longley coefficients meaningless?

It shouldn't matter if we add rounding error to the data, e.g., if we add a random uniform $[-0.4999, 0.4999]$ to each observation.

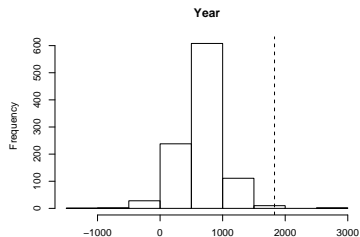
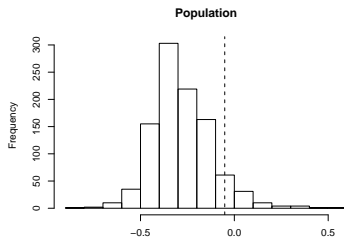
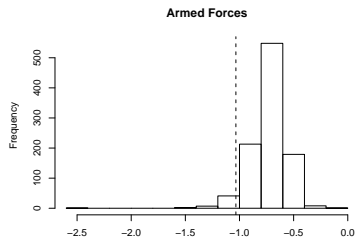
We should still get roughly the same coefficients, shouldn't we?

A.E. Beaton, D.B. Rubin, J.L. Barone (1976), "The acceptability of regression solutions: another look at computational accuracy", *JASA* **71** (1976) 158-168

An Example: Linear Regression



An Example: Linear Regression



An Example: Linear Regression

“[T]he numerically accurate solution in this example [Longley’s data] was probably an unreasonable estimate of the regression coefficients. This is true because the accuracy of the data and appropriateness of the model may affect the solution more than the computational method.”

“They [the data] are inadequate for the solution of this model.”

Beaton, Rubin and Barone p. 158

An Example: Linear Regression

Donohue, J.J. and S.D. Levitt (2001), “The Impact of Legalized Abortion on Crime”, *Quarterly Journal of Economics* 116(2):379-420.

Model I: $X \sim 663 \times 72$, $\kappa = 329,930$, $\frac{\|\hat{b} - \hat{b}^*\|}{\|\hat{b}\|} \leq 530 \times 10^9$

“The Impact of Legalized Abortion on Crime: Comment,” with Christopher F. Goetz. *Quarterly Journal of Economics* **123**:1 (February 2008)

Model II: $X \sim 6,724 \times 1,251$, $\kappa = 8141$, $\frac{\|\hat{b} - \hat{b}^*\|}{\|\hat{b}\|} \leq 530 \times 10^6$
[eigenvalues of $X'X$ are complex-valued!]

William Anderson, Martin T Wells (2008), “Numerical Analysis in Least Squares Regression with an Application to the Abortion-Crime Debate” *Journal of Empirical Legal Studies* **5**, 647-681.

An Example: Linear Regression

Let's just see if all these “linear” regressions had a linear relationship between the dependent variable (log of violent crime) and the key independent variable (effective abortion rate).

Conclusions


The economics profession is largely ignorant of the elements of accurate computing.

- ▶ solving OLS using normal equations
- ▶ calculator formula for standard deviation
- ▶ algebra - simulation
- ▶ algorithms - cointegration
- ▶ algorithms - vector autoregression
- ▶ algorithms - Yule-Walker
- ▶ textbooks - authors don't know
- ▶ journal articles - researchers don't know

Conclusions

The economics profession is largely ignorant of the elements of accurate computing.

Traditional econometrics packages are not capable of being used effectively for teaching the elements of econometric computing.

Only  can save economics!