The Political Methodologist

Newsletter of the Political Methodology Section
American Political Science Association
Volume 13, Number 1, Spring 2005

Editors:
Adam J. Berinsky, Massachusetts Institute of Technology
berinsky@mit.edu
Michael C. Herron, Dartmouth College
Michael.C.Herron@dartmouth.edu
Jeffrey B. Lewis, University of California Los Angeles
jblewis@ucla.edu

Editorial Assistant:
Hans Noel, University of California Los Angeles
hnoel@ucla.edu

Contents

Notes From the Editor

This time around we offer the first of what we hope will be several pieces of important original research. With journal space at a premium, more and more of what editors consider "technical details," but that researchers know to be central contributions, are relegated to so-called web appendices or can only be found in working papers. It is our intention to provide a broader forum for the interesting methodological advancements that ended up on the cutting room floors of the general interest journals.

The first of these is Jim Snyder's piece on the scaling of aggregate election returns. As part of a study of political representation that appeared in *Legislative Studies Quarterly* in 1996, Snyder proved conditions under which information about the distributions of voter ideal points within and across legislative districts could be inferred from district-level returns on ballot propositions. While aggregate quantities such as presidential support in the district are often used as proxies for the location of the mean or median voter in each district, the admissibility of such measures has remained relatively unexplored. We present Snyder's interesting findings on this topic as the lead article in this issue. If you have methodological or theoretical findings languishing in working papers or web appendices that you think would be of interest to the readers of *TPM* and the field more generally, please send them in.

This issue also contains an informative introduction to scientific computing under Apple's OS X operating system by Kevin Quinn, a review of undergraduate research methods textbooks by Richard Almeida, and a detailed tutorial on integrating LaTeX and *R* using the Sweave package by Peter Rudloff.

The final item in this issue is memorial for our friend John Williams, written by Patrick Brandt, Michael McGinnis, Burt Monroe and John Aldrich.

The fall issue of TPM is beginning to take shape, but we are always on the lookout for more material. As always, your submissions and ideas for topic to address are most welcome.

The Editors

Articles

Estimating the Distribution of Voter Preferences Using Partially Aggregated Voting Data

James M. Snyder, Jr.
Massachusetts Institute of Technology
millett@mit.edu

Partially aggregated voting returns, especially voting on ballot intitatives and referendums, are a potentially valuable source of data for identifying patterns in voter preferences and for studying questions of political representation. Deacon and Shapiro (1975), Kuklinski (1978), Snyder (1996), Kahn and Matsusaka (1997), and Lewis (1998) exploited such data in earlier work, and Ansolabehere et al (2000, 2002), and others are using similar data currently. In fact, it is arguable that aggregated data are even more relevant than individual level data for studying representation issues. Stimson (1991) states the argument clearly: "For a politician to pay attention to individual views is to miss the main game... The politician must, as a matter of image, appear concerned about individuals, but aggregate opinion is what matters (p. 12)." This point is especially important because aggregate data often exhibit starkly different patterns than individual level data. Aggregate opinion appears to be much more stable than individual opinion, and more predictable as well (Stimson, 1991; Page and Shapiro, 1992) It also appears to be more ideological, at least as measured by Converse's concept of "constraint" (Kuklinski, 1978; Snyder, 1996).

Scholars have developed a variety of empirical models for studying individual level voting data and survey data, which are well-grounded in a decision-theoretic framework (e.g., Poole and Rosenthal, 1997). Currently, however, we lack similar models suited for analyzing partially aggregated voting data. This note begins to fill the gap.

If voting data are aggregated, by legislative districts for example, then it is only possible to recover information about summary measures of voter preferences, such as district means and variances. Also, some assumption must be made about the general form of the within-district distribution of voter preferences, in addition to assumptions about voter behavior. The model below makes the following assumptions: (i) each proposition is viewed as two points in K-dimensional issue space, a Yea alternative and a Nay alternative; (ii) voters have Euclidean preferences, so each voter is characterized by an ideal point and prefers policies closer to this ideal point; (iii) voters are uncertain about the true location of alternatives on each proposition, and this uncertainty can be viewed as random noise added to voters' utilities; (iv) voter ideal points are normally distributed within each district. Assumptions (i)-(iii) are standard in the theoretical and applied literature on probabilistic voting, and are similar to assumptions in Poole and Rosenthal (1997), Heckman and Snyder (1997), Clinton, Jackman and Rivers (2004), and other work. Assumption (iv) is the main addition, and allows the application to aggregated data.

The most important result proved below is that under assumptions (i)-(iv) the appropriate model to fit aggregated vote data is a simple *linear* factor model (Proposition 2 and Corollary 2). Treating the propositions as variables, the factor loadings describe the propositions, and the factor scores describe the means and variances of the distribution of ideal points in each district. Ordinary principle components analysis of factor analysis may be used to estimate the dimensionality of the issue space, and to estimate linear combinations of the ideal point means.

Three features of the model deserve mention. First, the model allows voter uncertainty to vary across propositions. This is important because the general level of voter knowledge varies widely across propositions, and in most cases is probably more important than variation in the level of knowledge across voters for any given proposition. Few voters in California were unaware that Proposition 13 on the June 1978 ballot would cut property taxes, or that Proposition 10 on the November 1980 ballot, entitled "Smoking and Nonsmoking Sections," required restaurants to establish smoking and non-smoking sections. On the other hand, most voters probably knew little about the key issues surrounding Proposition 10 on the 1982 ballot, which allowed counties to merge their superior, municipal and justice courts. Second, if all districts are approximately equally heterogeneous (*i.e.*, the ideal point distributions have the same variance), then there will be exactly as many factors as issue dimensions. If the districts are not equally heterogeneous, then there will $K+1$ factors when the issue space has K dimensions. In this case, K of the factors describe the means of the districts' ideal point distributions scaled by the variances, and the remaining factor describes the variances of these distributions. Third, the distribution of voter ideal

points may include dimensions on which all voters in a district have the same ideal point (Proposition 3). This might be try for sectionally defined dimensions, such as "north vs. south." Also, the "quality" of each proposition can be treated as a special case of this type of dimension, which *all* voters have the same ideal point (all voters want higher quality). Sectional issues simply enter as additional linear factors, and quality enters as a constant. Voter "moods" (Stimson, 1991) can also be captured simply as an extra quality dimension.

Finally, I must mention the main limitation of the model. The model assumes that the distribution of voter preferences within each district is symmetric and normal. The assumption of normality can be relaxed (Remark 1 below), but symmetry is necessary to keep the problem simple. Thus, if the actual within-district distributions of voter ideal points are skewed, or if the distributions are unimodal along one dimension but bimodal along another, then the linear factor model is only an approximation of the true model. More work needs to be done to see how adequate this approximation is in practice.

The formal presentation of the model is as follows. Let I be a set of regions, let J be a set of ballot propositions, and let y_{ij} denote the fraction of voters in region i who vote Yea on proposition j. I begin with a basic model, then consider various extension. The basic model consists of the following assumptions:

(A.1) Each proposition j can be described by two points in $\Re^K$, a Yea alternative $\mathbf{x}_j$ and a Nay alternative $\mathbf{s}_j$.

(A.2) All voters have Euclidean preferences. Thus, the utility of a voter with ideal point at $\mathbf{z}$ can be described by $u(\mathbf{z},\mathbf{x}) = -(\mathbf{x}-\mathbf{z})'(\mathbf{x}-\mathbf{z})$.

(A.3) Voters vote for their most preferred alternative on each proposition.

(A.4) In each region i, the distribution of ideal points is a spherical multivariate normal with mean $\mathbf{z}_i$ and variance σ_i^2. A larger value of σ_i^2 means that voter preferences in region i are more heterogeneous.

Assumptions (A.1)-(A.4) imply that voting on ballot propositions is described by a *linear* factor model, as shown by the following proposition and corollary.

Proposition 1. Assume (A.1)-(A.4), let $\mathbf{b}_j = (\mathbf{x}_j - \mathbf{s}_j)/||\mathbf{x}_j - \mathbf{s}_j||$ and let $c_j = (\mathbf{x}_j'\mathbf{x}_j - \mathbf{s}_j'\mathbf{s}_j)/2||\mathbf{x}_j - \mathbf{s}_j||$. Then $y_{ij} = \Phi((\mathbf{z}_i'\mathbf{b}_j - c_j)/\sigma_j)$ for all i and j, where Φ is the standard normal cumulative distribution function.

Proof. By (A.1)-(A.3), the set of voters who vote Yea on proposition j is the half-space $Y_j = \{\mathbf{z} \mid \mathbf{z}'\mathbf{b}_j > c_j\}$. By ($A$.4), the ideal point distribution in region i is given by the joint density $f_i(\mathbf{z}) = (2\pi\sigma_i^2)^{-K/2} exp[-(\mathbf{z}-\mathbf{z}_i)'(\mathbf{z}-\mathbf{z}_i)/2\sigma_i^2]$. Thus, $y_{ij} = \int ... \int_{Y_j} f_i(\mathbf{z}) d\mathbf{z}$. Letting $\mathbf{v} = (\mathbf{z} - \mathbf{z}_i)/\sigma_i$ and changing variables, $y_{ij} = \int ... \int_{Y_{ij}} (2\pi)^{-K/2} exp[-\mathbf{v}'\mathbf{v}/2] d\mathbf{v}$, where $Y_{ij} = \{\mathbf{v} \mid \mathbf{v}'\mathbf{b}_j > (c_j - \mathbf{z}_i'\mathbf{b}_j/\sigma_i\}$. The integrand is now a multivariate standard normal density, so it is easily integrated (*e.g.*, Anderson, 1958) yielding $y_{ij} = 1 - \Phi((c_j - \mathbf{z}_i'\mathbf{b}_j)/\sigma_i) = \Phi((\mathbf{z}_i'\mathbf{b}_j - c_j)/\sigma_i)$. Q.E.D.

Corollary 1. Let $w_{ij} \equiv \Phi^{-1}(y_{ij})$ for all i and j. Then $\mathbf{w}_j = \mathbf{F}\mathbf{g}_j$ for all j, where $\mathbf{w}_j \equiv (w_{1j} \equiv (w_{1j}, ..., w_{Ij})', \mathbf{g}_j \equiv (b_{j1}, ..., b_{jK}, c_j)'$, and $\mathbf{F} = (\mathbf{f}_1, ..., \mathbf{f}_{K+1}) = ((z_{11}/\sigma_1, ..., z_{I1}/\sigma_I)', ..., (z_{K1}/\sigma_1, ..., z_{KI}/\sigma_I)', (1/\sigma_1, ..., 1/\sigma_I)')$.

Remark 1. Thus, inverting the vote shares y_{ij} using Φ, the resulting variables $\mathbf{w}_1, ..., \mathbf{w}_J$ are *linear* functions of the factors $\mathbf{f}_1, ..., \mathbf{f}_{K+1}$. If σ_i is the same for all i, then there are K factors when the issue space as K dimensions (the $1/\sigma_i$ "factor" is a constant). If the σ_i vary, then there are $K+1$ factors. Also, results similar to Proposition 1 and Corollary 1 hold for *any* spherical distributions of voter ideal points, not only for normal distributions (see Fang, Kotz and Ng, 1990).

In the basic model voters have perfect information about the propositions and make no "mistakes" when voting. I now assume voters have limited information, modeled as independent, normal, random noise added to voters' utilities. Replace assumption (A.3) above with the following assumption:

(A.3)$'$ The probability that a voter with ideal point $\mathbf{z}$ votes Yea on proposition j is $G_j(u(\mathbf{z},\mathbf{x}_j) - u(\mathbf{z},\mathbf{s}_j))$, where G_j is the cumulative distribution function of a normal random variable with mean 0 and variance θ_j^2.

Then we have the following proposition.

Proposition 2. Assume (A.1),(A.2),(A.3)$'$,(A.4), and let $\mathbf{b}_j = 2(\mathbf{x}_j - \mathbf{s}_j)/\theta_j$, $c_j = (\mathbf{x}_j'\mathbf{x}_j - \mathbf{s}_j'\mathbf{s}_j)/\theta_j$, and $\psi_j^2 = \mathbf{b}_j'\mathbf{b}_j$. Then $y_{ij} = \Phi((\mathbf{z}_i'\mathbf{b}_j - c_j)/(\sigma_i^2\psi_j^2 + 1)^{1/2}))$ for all i and j, where Φ is the standard normal cumulative distribution function.

Proof. By (A.1),(A.2), and (A.3)$'$, the set of voters who vote Yea on proposition j is $Y_j = \{(\mathbf{z}, \epsilon) \mid 2\mathbf{z}'(\mathbf{x}_j - \mathbf{s}_j) - \mathbf{x}_j'\mathbf{x}_j + \mathbf{s}_j'\mathbf{s}_j > \epsilon\}$, where $\epsilon \sim N(0, \theta_j^2)$. By ($A$.4), the distribution of ideal points in region i is given by the density $f_i(\mathbf{z}) = (2\pi\sigma_i^2)^{-K/2} exp[-(\mathbf{z}-\mathbf{z}_i)'(\mathbf{z}-\mathbf{z}_i)/2\sigma_i^2]$. Thus, $y_{ij} = \int ... \int_{Y_j} g_j(\epsilon) f_i(\mathbf{z}) d\epsilon d\mathbf{z}$, where $g_j \equiv G_j'$ is the density function associated with G_j. Substituting $\eta = \epsilon/\theta_j$ and $\mathbf{v} = (\mathbf{z} - \mathbf{z}_i)/\sigma_i$ yields $y_{ij} = \int ... \int_{Y_{ij}} (2\pi)^{-(K+1)/2} exp[-(\mathbf{v}'\mathbf{v} +$

$\eta^2)/2]d\eta d\mathbf{v}$, where $Y_{ij} = \{(\mathbf{v},\eta) \mid \sigma_i \mathbf{v}'\mathbf{b}_j + \mathbf{z}_i'\mathbf{b}_j - c_j > \eta\} = \{(\mathbf{v},\eta) \mid ((\sigma_i \mathbf{v}'\mathbf{b}_j - \eta)/(\sigma_i^2\psi_j^2+1)^{1/2}) > ((c_j - \mathbf{z}_i'\mathbf{b}_j)/(\sigma_i^2\psi_j^2+1)^{1/2})\}$. Since the integrand is now a multivariate standard normal density it is easily integrated, yielding $y_{ij} = \Phi((\mathbf{z}'_i\mathbf{b}_j - c_j)/(\sigma_i^2\psi_j^2+1)^{1/2})$. Q.E.D.

Corollary 2. Let $w_{ij} \equiv \Phi^{-1}(y_{ij})$ for all i and j, and let $\mathbf{w}_j \equiv (w_{1j},...,w_{Ij})'$.

(i) If $\sigma_i = \sigma$ for all i, then $\mathbf{w}_j$ is a linear function of the K factors $(z_{11},...,z_{I1})',...,(z_{K1},...,z_{KI})'$.

(ii) If $\psi_j^2 = \psi^2$ for all j, then $\mathbf{w}_j$ is a linear function of the $K+1$ factors $(z_{11}/\tilde{\sigma}_1,...,z_{I1}/\tilde{\sigma}_I)',...,(z_{K1}/\tilde{\sigma}_1,...,z_{KI}/\tilde{\sigma}_I)'$, $(1/\tilde{\sigma}_1,...,1/\tilde{\sigma}_I)'$, where $\tilde{\sigma}_i = (\sigma_i^2\psi^2+1)^{1/2}$.

(iii) If σ_i^2 and ψ_j^2 are "large", then $\mathbf{w}_j$ is approximately a linear function of the $K+1$ factors $(z_{11}/\sigma_1,...,z_{I1}\sigma_I)',...,(z_{K1}/\sigma_1,...,z_{KI}/\sigma_I)'$, $(1/\sigma_1,...,1/\sigma_I)'$.

Proof. Parts (i) and (ii) are obvious. The proof of (iii) follows by noting that if σ_i^2 and ψ_j are large, then $(\sigma_i^2\psi_j^2+1)^{1/2} \approx \sigma_i\psi_j$.

Remark 2. To see what "large" means in Corollary 2, note that ψ_j^2 measures how informed voters are about proposition j – larger values of ψ_j^2 mean fewer voter mistakes. Normalize by setting $(\mathbf{x}_j - \mathbf{s}_j)'(\mathbf{x}_j - \mathbf{s}_j) = 1$, and suppose voters with $\mathbf{z} = \mathbf{x}_j$ vote for $\mathbf{x}_j$ at least 85% of the time (other voters will make mistakes more often). Then $\psi_j^2 = 4/\theta_j^2 \geq 4\Phi^{-1}(.85) \approx 4.3$. Assuming that preference heterogeneity is at least as important a factor as voter information, $\sigma_i^2\psi_j^2 \geq 18.5$, and the discrepancy between $(\sigma_i^2\psi_j^2+1)^{1/2}$ and $\sigma_i\psi_j$ is only 2.5% or less.

The basic model above assumes that voters' preferences within each region vary across all of the K dimensions. I now extend the model to incorporate issues on which all voters within a given region have the *same* ideal point. This might be true for issues dealing with the geographic distribution of resources. Also, the "quality" of a proposition can be treated as a dimension on which all voters have the same ideal point; all voters want higher quality.

$(A.4)'$ In each region i, the distribution of ideal points with respect to dimensions $1,...,K-1$ is a spherical multivariate normal with mean $\hat{\mathbf{z}}_i$ and variance σ_i. With respect to dimension K, either all voters have $z_{iK} = 1$, or all voters have $z_{iK} = 0$.

Proposition 3. Assume $(A.1),(A.2),(A.3),(A.4)'$, and let $\hat{\mathbf{x}}_j \equiv (x_{j1},...,x_{j,K-1})$, $\hat{\mathbf{s}} \equiv (s_{j1},...,s_{j,K-1})$, $\mathbf{b}_j = ((\mathbf{x}_j - \mathbf{s}_j)/||\hat{\mathbf{x}}_j - \hat{\mathbf{s}}_j||)$ and $c_j = ((\mathbf{x}_j'\mathbf{x}_j - \mathbf{s}_j'\mathbf{s}_j)/(2||\hat{\mathbf{x}}_j - \hat{\mathbf{s}}_j||))$. Then $y_{ij} = \Phi((\mathbf{z}_i'\mathbf{b}_j - c_j)/\sigma_1)$ for all i and j, where Φ is the standard normal cumulative distribution function.

Proof. By $(A.1)$-$(A.3)$, the set of Yea voters on j is $Y_j = \{\mathbf{z} \mid \mathbf{z}'\mathbf{b}_j > c_j\}$. Thus, letting $\hat{\mathbf{z}} \equiv (z_1,...,z_{K-1})$ and $\hat{\mathbf{b}}_j \equiv (b_{j1},...,b_{j,K-1})$, for each region i with $z_{iK} = 1$ the set of Yea voters is $Y_{ij} = \{\mathbf{z} \mid \hat{\mathbf{z}}'\hat{\mathbf{b}}_j > c_j - b_{jK}\}$. By $(A.4)'$, the ideal point density in region i is: $f_i(\mathbf{z}) = (2\pi\sigma_i^2)^{-(K-1)/2}exp[-(\hat{\mathbf{z}}-\hat{\mathbf{z}})'(\hat{\mathbf{z}}-\hat{\mathbf{z}})/2\sigma_i^2]$ if $z_K = 1$, and $f_i(\mathbf{z}) = 0$ if $z_K \neq 1$. Thus, $y_{ij} = \int ... \int_{Y_{ij}} f_i(\mathbf{z})d\mathbf{z}$, or substituting $\hat{\mathbf{z}} \equiv (\hat{z_1},...,\hat{z_{K-1}})$ and $\mathbf{v} = (\hat{\mathbf{z}}-\hat{\mathbf{z}}_i)/\sigma_i$, $y_{ij} = \int ... \int_{Y_{ij}} (2\pi)^{-K/2}exp[-\mathbf{v}'\mathbf{v}/2]d\mathbf{v}$, where $Y'_{ij} = \{\mathbf{v} \mid \mathbf{v}'\hat{\mathbf{b}}_j > (c_j - b_{jK} - \hat{\mathbf{z}}_i'\hat{\mathbf{b}}_j)/\sigma_i\}$. Integrating, $y_{ij} = \Phi((\hat{\mathbf{z}}_i'\hat{\mathbf{b}}_j + b_{jK} - c_j)/\sigma_i) = \Phi((\mathbf{z}_i'\mathbf{b}_j - c_j)/\sigma_i)$. Similarly, for each region i with $z_{iK} = 0$ the set of Yea voters is $Y_{ij} = \{\mathbf{z} \mid \hat{\mathbf{z}}'\hat{\mathbf{b}}_j > c_j\}$ and the distribution of ideal points is: $f_i(\mathbf{z}) = (2\pi\sigma_i^2)^{-(K-1)/2}exp[-(\hat{\mathbf{z}}-\hat{\mathbf{z}})'(\hat{\mathbf{z}}-\hat{\mathbf{z}})/2\sigma_i^2]$ if $z_K = 0$, and $f_i(\mathbf{z}) = 0$ if $z_K \neq 0$. Changing variables and integrating yields $y_{ij} = \Phi((\hat{\mathbf{z}}_i'\hat{\mathbf{b}}_j - c_j)/\sigma_i) = \Phi((\mathbf{z}_i'\mathbf{b}_j - c_j)/\sigma_i)$. Q.E.D.

Remark 3. Inverting the y_{ij} using Φ yields a linear factor model, just as in corollary 1. Also, Proposition 3 and its corollary are easily generalized to "geographic" issues that have more than two values, and to issue spaces with more than one such issue. Finally, combining "geographic" issues and limited information yields results analogous to Proposition 2 and its corollary.

Remark 4. To see how quality can be treated as a dimension on which all voters have the same ideal point, let the Kth dimension represent quality, and label the dimension's axis so that a higher value means *lower* quality. The preferences of a voter whose ideal point along dimensions $1,...,K-1$ is at $\hat{\mathbf{z}} \equiv (z_1,...,z_{K-1})$ can then be represented by $u(\mathbf{z},\mathbf{x}) = -(\hat{\mathbf{x}}-\hat{\mathbf{z}})'(\hat{\mathbf{x}}-\hat{\mathbf{z}}) - x_K^2$, where $\hat{\mathbf{x}} \equiv (x_1,...,x_{K-1})$. Letting $z_K = 0, u(\mathbf{z},\mathbf{x}) = -(\mathbf{x}-\mathbf{z})'(\mathbf{x}-\mathbf{z})$. That is, it is as if each voter has Euclidean preferences in a K-dimensional space, with ideal point along the Kth dimension at zero. The quality dimension does not enter as a separate factor, however, since all regions have the same mean ideal point along this dimension.

References

Anderson, T.W. 1958. *An Introduction to Multivariate Statistical Analysis.* New York: John Wiley and Sons.

Ansolabehere, Stephen, James M. Snyder, Jr., and Jonathan Woon. 2000. "Why Did a Majority of Californians Vote to Limit Their Own Power?" Unpub-

lished manuscript, Massachusetts Institute of Technology.

Ansolabehere, Stephen, James M. Snyder, Jr., and Ruimin He. 2002. "Evidence of Virtual Representation: Reapportionment in California." Unpublished manuscript, Massachusetts Institute of Technology.

Clinton, Joshua, Simon Jackman, and Douglas Rivers. 2004. "The Statistical Analysis of Roll Call Data." *American Political Science Review* 98: 355-370.

Deacon, Robert, and Perry Shapiro. 1975. "Private Preference for Collective Goods Revealed through Voting on Referenda." *American Economic Review* 65: 943-955.

Fang, K.T., S. Kotz and K.W. Ng. 1990. *Symmetric Multivariate and Related Distributions.* New York: Chapman and Hall.

Heckman, James J., and James M. Snyder, Jr. 1997. "Linear Probability Models of the Demand for Attributes with an Empirical Application to Estimating the Preferences of Legislators." *RAND Journal of Economics* 28:S142-189.

Kahn, Matthew E., and John G. Matsusaka. 1997. "Demand for Environmental Goods: Evidence from Voting patterns on California Initiatives." *Journal of Law and Economics* 40: 137-173.f

Kuklinski, James H. 1978. "Representativeness and Elections: A Policy Analysis." *American Political Science Review* 72: 165-177.

Lewis, Jeffrey B. 1998. *Who Do Representatives Represent?* Unpublished dissertation, Massachusetts Institute of Technology.

Page, Benjamin I., and Robert Y. Shapiro. 1992. *The Rational Public.* Chicago: University of Chicago Press.

Poole, Keith T., and Howard Rosenthal. 1997. *Congress: A Political-Economic History of Roll Call Voting.* New York: Oxford University Press.

Snyder, James M., Jr. 1996. "Constituency Preferences: California Ballot Propositions, 1974-90." *Legislative Studies Quarterly* 21: 463-488

Stimson, James A. 1991. *Public Opinion in America: Moods, Cycles, and Swings.* Boulder, CO: Westview Press, 1991.

Computing and Software notes

Mac OS X for Political Methodologists

Kevin M. Quinn
Harvard University
kevin_quinn@harvard.edu

Introduction

Not so long ago, I regularly poked fun at friends and colleagues who were Macintosh users for working with an operating system that seemed designed more for editing digital photographs than for scientific computing. However, as fate would have it, I am now writing this very sentence on a Mac. What happened? Why would someone who has experience working with "real" operating systems such as the various flavors of Unix and who is comfortable using Unix tools decide to work on a Mac?

In my case, the reasons for the switch are multiple (and have relatively little to do with editing digital photographs). They include the following.

- Mac OS X is built on FreeBSD and thus has the Unix tools I use on a daily basis
- Mac OS X is easily maintainable– I don't have to spend much time administering my machines
- the Aqua user interface is well-designed and a joy to use
- the high quality of Apple hardware– particularly the laptops

In the remainder of this article, I flesh out these points and provide some additional information about tools and resources for Mac OS X that may be of interest to those in the methods community.

Mac OS X vs. Linux

The most important factor contributing to viability of Apple machines for scientific computing was the radical break from past versions of the Mac operating system that occurred with the introduction of Mac OS X in spring 2001. The introduction of Mac OS X was such a profound change of direction because this new operating system, targeted at typical home users, was built on a foundation of Unix technology. Like the NeXTStep operating system, Mac OS X is based on the Mach kernel and the BSD branch of Unix (OS X makes use of FreeBSD). Further, the Cocoa development environment (a key part of Mac OS X) is an implementation of the OpenStep API specification. This has led many observers to remark that Mac OS X is essentially a current version of the NeXTStep / OPENSTEP operating system. A good overview of Mac OS X, both in terms of its history and architecture, can be found at `http://www.kernelthread.com/mac/osx/`.

Mac OS X's lineage is apparent in its performance. Most typical users find that, in addition to the Unix functionality found in FreeBSD, Mac OS X provides the ease of use and administration found in previous versions of the Mac operating system. Unlike users of earlier versions of the Mac operating system, Mac OS X users can interact with the computer from a terminal using their choice of Unix shell. The usual Unix shell commands (`ls`, `cd`, `grep`, etc.) are available as well as specialized commands that allow a user to interact directly with applications written specifically for Mac OS X. For instance, if one has a version of Microsoft Word installed one could open the Word document called `someoneelsesfile.doc` by issuing

```
open someoneelsesfile.doc
```

at the shell prompt. The `open` command is completely generic and can be used to open any file in the same way as would occur by clicking on that file's icon. This command is a holdover from the NeXTStep operating system.

While users can work exclusively from a terminal, most users will want to make at least some use of a modern graphical user interface. Aqua is Mac OS X's graphical user interface. Aqua is the best desktop environment I've ever used. It is both highly functional and great looking. One nice feature found in Aqua is Exposé. This is a feature that allows users to tile the desktop with miniaturized versions of all open windows with the push of a hot-key. The user can then easily switch control to any open window by clicking on it. Aqua has several universal access features such as sticky keys and some limited voice recognition ability. Mac OS X and Aqua also provide good support for speakers of languages other than English. Finally, Aqua features a fast user switching feature that allows multiple users to securely share a single OS X machine. It is worth noting that users who do not like the Aqua interface can always install a different desktop environment such as KDE.

References to Mac OS X throughout the rest of this article will generally refer to the current (at the time this is being written) version of Mac OS X, Mac OS 10.3 Panther.

Apple Hardware

Apple's hardware is some of the best available today. Apple's desktop and laptop machines had the highest user ratings in PC Magazine's 17th Annual Reader Satisfaction Survey (published in July 2004). Several Apple machines have also been awarded the Editors' Choice ranking by this publication. Apple's machines deliver good performance along with solid reliability and handsome, functional design. Somewhat remarkably, this is true of all their hardware– from their desktop machines, laptop machines, and displays to peripherals such as the iSight and Airport Base Station.

Apple currently offers three lines of desktop machines– the Power Mac G5, the iMac G5, and the eMac. The Power Mac G5 is the high end of the Apple desktop line. The Power Mac G5 family ranges from a single processor 1.8GHz machine to a dual processor 2.5GHz machine. For users who do not do a great deal of time intensive computing, the iMac G5 provides a nice mix of value and performance. Finally, the eMac series, which sells for under $1000, is geared towards those on a tight budget. While none of these machines (even the high end Power Mac G5s) are as fast as comparably priced machines with AMD or Intel processors running Linux, they are reasonably fast; i.e, certainly fast enough for anything you would want to do on a true desktop machine rather than a dedicated workstation.

Apple laptops run from the enormous 17-inch PowerBook G4 to the very compact and light 12-inch PowerBook G4 and 12-inch iBook G4. In my opinion, this is where Apple really shines. My 12-inch PowerBook G4 is rugged, compact, comfortable to use, and has great battery life. OS X also provides one of the easiest ways to get Unix functionality on a laptop.

A couple of other pieces of hardware that are worth mentioning are Apple's iSight and Airport Extreme. iSight is a small video camera that, for the price of a .Mac membership (more about this below) allows users to video conference with colleagues. My experience with iSight has been quite good. It produces clear high resolution images with good color reproduction under a wide range of lighting. Airport Extreme is Apple's wireless technology. In my experience, Airport Extreme works well and is quite easy to setup.

From a user's perspective, one of the great things about Apple machines running OS X is that the hardware

and the OS have been built with an eye toward each other. From a practical standpoint, what this means is that peripheral devices such as displays, external hard drives, video cameras, PDAs, etc. work directly out of the box. You do not have to recompile the kernel every time you bring a new peripheral device home.

Getting Started with Mac OS X

As noted above, Mac OS X is built on FreeBSD. Thus, straight out of the box, Mac OS X has most of the Unix tools that one would expect. For instance, Mac OS X comes with a full-color terminal application, Emacs, vim, gnutar, gzip, Secure Shell, the Apache web server, and many other commonly used tools. However, the experienced Unix user will find some common tools and applications missing, and, in some cases, will want improved versions of the tools that are automatically available. Thankfully most, if not all, of these additional packages can be very easily downloaded and installed from either Apple's website or third-party websites.

One of the most noticeable omissions from the standard releases of Mac OS X is an implementation of the X Window System (X11)– the windowing system upon which most graphical applications for Unix-based operating systems are built. Without an implementation of X11 it would be impossible to run Unix applications (such as XEmacs) that rely on X11 on a Mac. The good news is that Apple produces an implementation of X11 that is based on the open source XFree86 project. Apple X11 is easily downloaded from the Apple website and can be installed with just a few simple clicks of a mouse button.

While the earliest releases of Apple X11 were somewhat buggy, my experience with recent releases has been very good. I have not had any compatibility problems with X11 applications, and the performance of Apple X11 is not noticeably different from that of XFree86 on a roughly comparable Linux machine. Apple X11 allows for both rootless operation and full screen operation. Under rootless operation the user never leaves the Aqua environment. X11 windows are displayed directly on the Aqua desktop and behave like other Aqua windows. This is quite nice in that it allows a user to work with X11 and Aqua windows at the same time. Full screen mode puts all of a user's X11 windows on a separate screen. A hot key is used to move back and forth between the Aqua screen and the X11 screen. The default window manager is Apple's Quartz window manager. This option works fine for me. Users who don't like the Quartz window manager can easily install a different Linux window manager.

One of the real advantages of having X11 on a local Mac is that it allows one to run X11 applications on remote Unix, Linux, and OS X machines as easily as running the same application locally. To do this one simply connects to the remote machine using Secure Shell with the `-X` flag (this flag enables X11 forwarding) and then starts the X11 application. For instance, at the OS X shell prompt one might enter:

```
ssh -X username@somemachine.somewhere.edu
```

Then at the shell prompt on `somemachine.somewhere.edu` one could start an X11 application in the usual way. The resulting application windows will automatically be forwarded to the display on the local Mac OS X machine. In my experience, this a much easier way to use X11 applications on remote machines than using software such as VNC.

Another noteworthy omission from the standard Mac OS X distribution is the GNU Compiler Collection– GCC. Again, the good news is that Apple bundles a recent version of GCC (currently GCC 3.3) with their Xcode IDE as their Xcode Tools. Recent purchasers of a Mac can freely install the Xcode Tools by clicking on the self-installer in their Applications > Installers > Xcode Tools folder. The Xcode Tools can also be downloaded directly from the Apple website.

A resource that is very useful for all Mac OS X users, but especially for Unix users, is Fink (`http://fink.sourceforge.net/`). Fink is a repository of open source software originally written for Unix that has been modified to work correctly under Mac OS X. Fink uses the Debian packaging tools to allow users to easily download and install software from the Fink servers. For example, after you have downloaded and installed the Fink client you could issue the following command at the shell prompt:

```
sudo fink install emacs21
```

to install Emacs version 21.3 with X11 support. Fink provides a very easy way to install useful packages such as: CVS, GNU Fortran compilers, teTeX, Subversion, and many others. As of January 7, 2005, there are 4638 packages available on the Fink servers. A list of packages broken down by application area can be found at: `http://fink.sourceforge.net/pdb/index.php`.

Users who work primarily at the shell prompt will probably want to install a terminal application that has more features than the standard Apple terminal. My personal bias is for an application called iTerm. iTerm is written explicitly for Mac OS X using the Cocoa framework (a class library that allows for close integration with the Aqua GUI), and thus can take full advantage of the Aqua user interface. iTerm has a number of nice features including: multi-tab displays; support for VT100, xterm, and ANSI emulation; custom key-mapping; support for non-Latin alphabets; and many other useful features. I find the multi-tab displays to be indispensable. Rather than have a large number of terminal windows open (and taking up most of

the room) on one's desktop, one can stack all of these terminals in the space it would take to display a single terminal; and then, because the terminals have tabs attached, one can easily move back and forth from one terminal to another by simply clicking on the appropriate tab. An iTerm window is displayed in Figure ??. iTerm can be freely downloaded from `http://iterm.sourceforge.net/`.

Readers of this essay are undoubtedly interested in what statistical packages are supported under Mac OS X. Readers can rest assured that packages such as R, RATS, SPSS, and Stata are all available for Mac OS X and that they run smoothly. A very useful website for those wishing to build R from source on their OS X machine is maintained by Stefano Iacus and is located at: `http://www.economia.unimi.it/R/`. Further, mathematics packages such as Maple, Mathematica, and MATLAB run under Mac OS X as well.

It is also worthwhile to note that Microsoft Word, Excel, PowerPoint, and Entourage work under Mac OS X. Further, filesharing between Windows machines and Mac OS X machines is possible.

Finally, there are some services that, while not free, are extremely useful. Apple's .Mac is clearly one such service. For the price of $99.95 a year one gets 125 MB of storage on what Apple calls an iDisk. One's iDisk is located on an Apple server and can be accessed from anywhere regardless of whether you have your Mac with you. It is also possible to keep a local copy of one's iDisk on all of one's Macs. The real advantage of the iDisk is that it can be used with Apple's iSync utility which automatically synchronizes files across all of a user's Macs. For instance, I tend to keep the files related to the courses I'm teaching in a particular semester on my iDisk. Using iDisk and iSync I can revise the lecture notes for a class on the Mac on the first floor of my home, click the sync button, and by the time I walk upstairs to get my laptop I can view the new revisions on the laptop. iDisk together with iSync makes working on multiple machines very easy– one never has to worry about manually copying files or overwriting a more current version of a file with a less current version of the same file. A .Mac subscription also provides some other benefits such as a .Mac email account, anti-virus software, and backup software. Even if one does not subscribe to .Mac one can still use iSync to synchronize one's calendar, addressbook, mail, etc. at no charge.

Conclusion

In short, Mac OS X is a great operating system for those who want to spend more time using their computer than administering their computer. Mac OS X is built on FreeBSD and thus has many of the tools Unix users expect. Unix tools that are not part of the standard OS X release can be easily downloaded and installed. While Mac OS X has much of the power of a traditional Unix machine it also has an extremely well-designed user interface that makes everyday tasks a breeze.

An Introduction to Sweave

Peter Rudloff
University of Illinois at Urbana-Champaign
rudloff@uiuc.edu

Introduction

Statistical analysis with R and typesetting with LaTeX are increasingly popular among political scientists. However, for users who are relatively new to these programs, learning the syntax and tricks can be a considerable time investment. This article introduces a package that allows users to combine both R and LaTeX in time saving way. Sweave is an R package written by Friedrich Leisch which allows a researcher to "weave" R code within a LaTeX document through the noweb literate programming tool (Ramsey 1998). The noweb literate programming tool allows for the inclusion of text alongside programming code.[1] Instead of two different work flows (one for R and one for LaTeX), researchers can combine both statistical analysis and typesetting within a single framework. This paper illustrates how Sweave allow for the inclusion of R code "chunks" into what would otherwise be a LaTeX document. All of this information is saved as a noweb (`.Rnw`) document. Sweave processes the noweb file, replaces the code chunks with any statistical analysis the researcher requests (as LaTeX code), and then produces a `.tex` file. Figure 1 outlines this process.

Why use Sweave if it is only going to add another step to processing your manuscript? Since `.Rnw` documents combine R and LaTeX code, there is no need to write, revise

[1] http://www.eecs.harvard.edu/∼nr/noweb/.

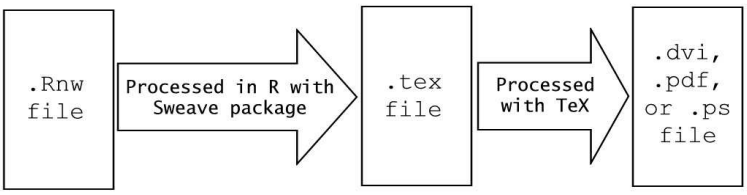

Figure 1: Outline of Sweave work flow

and maintain an R source file separate from a LaTeX file. It also means statistical analysis can be updated dynamically alongside a narrative. Sweave can be used in conjunction with R packages such as xtable, which produces statistical results in LaTeX code rather than the verbatim output produced by R. Sweave produces graphs and plots in both Encapsulated PostScript and Portable Document Format as requested by the researcher, and then includes the appropriate `\includegraphics{}` lines of LaTeX code in a `.tex` file. In addition to saving time, Sweave also increases accuracy by eliminating the need to manually transfer statistical results from R to a LaTeX file.

For those who already use LaTeX and R, the benefits of learning Sweave greatly outweigh the initial learning costs. Plus, Sweave is easy to obtain because it is already part of R's base installation. Either a text editor with the ability to process R code or R itself can Sweave a `.Rnw` noweb file into a `.tex` file. Thus, utilizing Sweave requires no additional software for those already familiar with R and LaTeX.

The Basics of Sweave

In most ways, a Sweave file looks exactly like a typical LaTeX file, with a few notable additions.[2] These additions are lines of R code referred to as "code chunks," which are delineated by `<<>>=` and `@`. The first bit of syntax ,`<<>>=`, demarcates the beginning of the code chunk and the second, @, indicates the end. In between these two symbols lies R code much as you would find in an R file. When Sweave processes a `.Rnw` file, it ignores anything not contained within `<<>>=` and `@`. Sweave processes these code chunks using R and replaces these code chunks with output that makes sense to LaTeX. If any of the code chunks create images such as graphs, Sweave creates these images in both `.eps` and `.pdf` format. After Sweave replaces the code chunks with LaTeX code and creates the necessary images, it then creates a `.tex` file which can be processed by any of the standard TeX distributions (e.g. fpTeX, MikTeX, teTeX, etc.).

Figure 2 illustrates a sample `.Rnw` file. Notice there are three different code chunks within this file, each delineated by `<<>>=` at the beginning and `@` at the end. The first code chunk loads McNeil's (1977) data on urban populations and crime rates by state, and attaches the column names for easier reference.[3] The second code chunk produces scatter plots of each pair of variables in the `USArrests` data set. Notice this code chunk is contained within a `figure` environment, so that the plot can be manipulated using the various options of the environment, such as the `h` (`h` stands for "here") to indicate to LaTeX that you wish the figure to appear in this position within the final text output. The final code chunk produces the results of the two regressions. The text contained within `<< ... >>=` notifies Sweave of changes in any of the default options for the subsequent code chunk. These options are quite important in producing the desired final text.

Let us ignore these options for the moment and simply process the file using Sweave. The above text should be saved as a noweb source file with the `.Rnw` extension. To process the file, simply begin a session of R and enter `Sweave(`*`filename`*`)` at the prompt, where *filename* represents the location and name of the source file.[4] Sweave processes each of the code chunks and replaces them with LaTeX code. After this is complete, Sweave saves the final output as a `.tex` file and produces any graphics requested. Figure 3 presents the results of processing Figure 2's `.Rnw` file in Sweave.

The resulting `.tex` file contains several unusual lines of code for the scholar unfamiliar with Sweave. Specifically, the `Schunk` and `Soutput` are not found in typical TeX files. The `Sweave.sty` inserted in the preamble of the resulting TeX file allows LaTeX to process these lines as a special environment for R results. The `Sweave.sty` style package is

[2] Sweave files use the file extensions `.Rnw` or `.Snw` to indicate whether the file is processed using S or R as well as that it is a noweb file type.

[3] According to the R documentation, the original sources of this data are *World Almanac and Book of Facts 1975* for the crime rates and *Statistical Abstracts of the United States 1975* for the urban populations statistics.

[4] For instance, if you saved this text as `crime.Rnw` in the `c:/` directory, the correct input would be `Sweave("C:/crime.Rnw")`

```
\documentclass[letterpaper,12pt]{article}

\begin{document}

<<echo=false,results=hide>>=
data(USArrests)
attach(USArrests)
@

Let's examine whether there is a relationship between the percentage
of a US state's urban population and crime rates.  First, we'll try
some simple plots to check for obvious patterns in the data.

\begin{figure}[h]
\centering
<<echo=false,fig=true,width=8,height=8>>=pairs(USArrests) @
\end{figure}

It would seem that there is at least a hint of a linear relationship
between the rate of assault and percent urban population, while the
relationship between murder rates and urban population appears
nonexistent.  Let's run a simple regression to examine whether our
hunch about murder rates is true.

<<echo=false>>=
murder<-lm(Murder~UrbanPop)
print(summary(murder))
@

\end{document}
```

Figure 2: A sample Sweave file

```
\documentclass[letterpaper,12pt]{article}
\usepackage{Sweave}
\begin{document}

Let's examine whether there is a relationship between the percentage
of a US state's urban population and crime rates.  First, we'll try
some simple plots to check for obvious patterns in the data.

\begin{figure}[h]
\centering
\includegraphics{illustration1-002}
\end{figure}

It would seem that there is at least a hint of a linear relationship
between the rate of assault and percent urban population, while the
relationship between murder rates and urban population appears
nonexistent.  Let's run a simple regression to examine whether our
hunch about murder rates is true.

\begin{Schunk}
\begin{Soutput}
Call: lm(formula = Murder ~ UrbanPop)

Residuals:
   Min     1Q Median     3Q    Max
-6.537 -3.736 -0.779  3.332  9.728

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.41594    2.90669   2.207   0.0321 *
 UrbanPop    0.02093    0.04333   0.483   0.6312
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.39 on 48 degrees of freedom Multiple
R-Squared: 0.00484,    Adjusted R-squared: -0.01589 F-statistic:
0.2335 on 1 and 48 DF,  p-value: 0.6312
\end{Soutput}
\end{Schunk}
\end{document}
```

Figure 3: The TEX output of the Sweave file in Figure 2

found in the /share/ folder of R's directory.[5,6]

Manipulating Code Chunks

Figure 2 includes a variety of code chunk options which determine the resulting LaTeX code. Inclusion of the correct options is vital in determining the appearance of the final document. For instance, notice the first code chunk in Figure 2, which loads and attaches USArrests. As Figure 3 indicates, this R code chunk does not result in any LaTeX code in the final .tex file. Since these lines only initialize data, there is no reason for their inclusion in the final .tex document, even though they are vital for the subsequent statistical analysis in the original .Rnw file. The option echo=false informs Sweave to include the lines of R code in the subsequent code chunk in the .tex file. The default option is echo=true, meaning that Sweave produces every command line in the final .tex document unless otherwise indicated. The results=hide option suppresses any analysis produced by these lines of R code.[7]

The second code chunk in Figure 2 determines the graphics produced by Sweave in addition to how these graphics are included in the resulting .tex file. Again, I do not wish the command line pairs(USArrests) to appear in the final document, so I use the echo=false option. However, I do wish Sweave to produce the pairs graphic in the final document, which is why I include the line fig=true. The width and height options indicate to Sweave the dimensions of the resulting graphic. The default unit of measurement in this case is inches.

The final code chunk produces a summary of the regression analysis. Because I want the results to be produced in the final .tex document, I do not include the results=hide option as I did in the first code chunk. The default setting of the results option is verbatim, which orders Sweave to place results in the verbatim-like Soutput environment. I still wish to hide the command lines, so I include the echo=false option. As a result, Sweave does not include the lines of R code in the final .tex document, though the results of the statistical analysis are produced.

Space constraints do not allow for a full accounting of the options available in Sweave. There are options that allow the user to suppress the creation of either an .eps or .pdf graphics file when creating figures. The prefix option will ensure any figures created will share a common prefix. To learn more about these and other useful options, consult the Sweave manual (Leisch 2004, 11 - 12).

If R's default output does not appeal to the eye, R has contributed packages which produce results in more attractive LaTeX code. For example, Leisch (2002) produces several graphics using some of R's graphics options as well as the xtable contributed package available on CRAN. The xtable package is useful for anyone using LaTeX in conjunction with R, even in the absence of Sweave, because it produces the LaTeX code necessary to present analysis in a format more pleasing to the eye. For instance, by including two lines, library(xtable) after data(USArrests) and xtable(murder) instead of print(summary(murder)), the unattractive R output in Figure 2 is replaced by the following LaTeX code:

```
\begin{table}[ht]
\begin{center}
\begin{tabular}{rrrrr}
\hline
 & Estimate & Std. Error & t value & Pr($>$$|$t$|$) \\
\hline
(Intercept) & 6.4159 & 2.9067 & 2.21 & 0.0321 \\
UrbanPop & 0.0209 & 0.0433 & 0.48 & 0.6312 \\
\hline
\end{tabular}
\end{center}
\end{table}
```

This code in turn produces the following table, which is much more likely to appear in a paper manuscript.

Further Reading

I hope this brief article has made the value of Sweave apparent. Those already familiar with R and LaTeX will find that little further investment is required to reap the greater efficiency and accuracy that Sweave allows. For further reading regarding Sweave, Friedrich Leisch's website is an excellent starting point.[8] The website contains the Sweave manual as well as several articles written on the subject.

References

Friedrich Leisch. 2002. "Sweave, Part I: Mixing R and LaTeX." *R Newslet-*

[5] One common problem is while the style file sweave.sty is included in the base distribution of R, the file is often inaccessible to LaTeX. This is because R is often installed in a directory where the path contains empty spaces. For instance, if R is installed in your My Documents folder in Windows, then LaTeX will be unable to access the style file because the path to that file contains a blank space (i.e. the space between My and Documents). Friedrich Leisch suggests two solutions to this problem: 1. Install R with no spaces in the path (such as directly onto your c:/ drive), or 2. copy sweave.sty into your local TeX tree (Leisch 2004, 16). However, keep in mind that if you decide on option two, you may have to update the file name database of your local TeX tree as well as include the line Sweave manually in the preamble of your .tex file.

[6] If you would like to examine the final product in .pdf format, it is available at https://netfiles.uiuc.edu/rudloff/www/illustration1.pdf.

[7] I should note there is actually no reason to include results=hide option since I never ask R to conduct any statistical analysis in this code chunk. I include this option only for illustration purposes.

[8] URL http://www.ci.tuwien.ac.at/~leisch/.

ter: The Newsletter of the R Project 2:3, 28 - 31. URL `http://cran.wustl.edu/doc/Rnews/Rnews_2002-3.pdf`.

Friedrich Leisch. 2003. "Sweave, Part II: Package Vignettes." *R Newsletter: The Newsletter of the R Project* 3:2, 21 - 24. URL `http://cran.wustl.edu/doc/Rnews/Rnews_2003-2.pdf`.

Friedrich Leisch. 2004. *Sweave User Manual.* Vienna University of Technology, Autria. URL `http://www.ci.tuwien.ac.at/~leisch/Sweave/`. R version 2.0.0.

Norman Ramsey. 2004. *Noweb home page.* University of Virginia, USA. URL `http://www.eecs.virginia.edu/~nr/noweb`.

Book Review

Undergraduate Research Methods Texts

Richard A. Almeida
Southeast Missouri State University
ralmeida@semo.edu

Political Science Research Methods, 5th edition. Janet Buttolph Johnson & H. T. Reynolds. Washington, DC: CQ Press, 2005. Pp. xxii, 515.

The Practice of Social Research, 10th edition. Earl J. Babbie. Belmont, CA: Wadsworth/Thomson Learning, 2004. Pp. xxiv, 493.

It's difficult to say which task is more formidable: teaching research methods to undergraduates or selecting an appropriate text or texts to do the same. Textbook offerings range widely in terms of rigor and comprehensibility, as do preferences and levels of enthusiasm and preparation in students, faculty, and departments alike. Some courses focus solely or primarily on gathering and analyzing information, others, like that at my own institution, also comprise a healthy portion of "scope," or "Introduction to Political Science," as the course catalog terms it. These texts, *Political Science Research Methods* by Janet Buttolph Johnson & H. T. Reynolds (hereinafter PSRM and J&R, respectively) and *The Practice of Social Research* (hereinafter TPSR) by Earl Babbie reflect some of this diversity. The works by J&R and Babbie appear to be intended for use as the central text in a class primarily geared toward political science applied research, though J&R offer some leverage on the "political" end of "political science;" what questions do practitioners find interesting, why, and what sorts of answers do they find, though any discussion of the real breadth of the discipline of political science would have to lean heavily on materials drawn from outside the text.

Babbie and J&R's books are both set up as fairly comprehensive introductory research methods textbooks, and either would succeed, with caveats, as central textbooks in a one-semester, lower division course. PSRM has the added benefit of being considerably more useful for a more rigorous single semester class or a two-course "scope & methods" series, as its explicit political science focus makes it relevant to discussions of the "state of the discipline," though examples drawn from outside the fields of American government and public opinion are fairly limited. PSRM leans heavily on relevant, accessible research on such appealing questions as voter turnout, the impact of negative advertising, and judicial decision-making. Babbie's time-tested work, on the other hand, is more engaging and intuitively appealing, particularly for a more general, lower-division course where students' technical preparation can be lacking. However, the focus of TPSR is on social science quite generally, as evidenced by the title, and so would most likely require explicitly political science supplemental materials if adopted as a primary textbook, though Babbie's insights, logic, and explication are thoroughly sound.

The two methods texts are structured similarly, starting with some discussion of philosophy of inquiry and proceeding through what Babbie terms, "The Structure of Inquiry," through operationalization, measurement, validity, data gathering, analysis, and, finally, statistical analysis (both texts do a commendable job of demonstrating that analyzing data goes beyond "crunching numbers" to extract meaning from information). Both texts provide a great deal of supplemental material; TPSR excels here. Babbie's book contains a particular variety of supplemental materials, from a content-rich website and instructor's manual/test bank to a highly interactive (albeit video-intensive) included student CD-ROM featuring a well-executed guide to writing research papers and detailed chapter outlines.

PSRM also features a student workbook with CD-ROM (provided) and website. While the J&R package lacks Babbie's multimedia pizzazz, this is not necessarily a detraction. The PSRM website (`http://psrm.cqpress.com`), however, lacks many desirable features, particularly quizzes and chapter outlines. The PSRM website does, however, provide data resources and an accessible set of supplemental materials. The supplemental CD-ROM included with the (optional) workbook seems to be of fairly limited use; consisting of topical data files in SPSS and other formats (Excel and SPSS portable) and texts of speeches and reports by contemporary political figures.

While Babbie's textbook is superior to PSRM in terms of accessibility and a wealth of supplemental information, it is nevertheless the case that J&R's explicit focus on political science trumps Babbie's polish and breadth. PSRM consistently highlights contemporary political science research that is relevant, informative, and accessible. Having a textbook that is centered around contemporary political science research is clearly a benefit, and instructors who choose to use Babbie's text would necessarily have to supplement it heavily with political science-specific additional material.

But perhaps the greatest drawback to TPSR as a methods text is that it concludes by teaching students *about* statistical research instead of *teaching* statistical research. There is a significant difference. For example, in Babbie's chapter 14 (p. 405-6), he introduces the concept of standard deviation. However, a *formula* for calculating standard deviation is nowhere presented in the textbook. This is a real shortcoming. It is of course vital for new methods students to understand the *concept* of dispersion, but it is equally desirable for students to be able to *compute* measures of dispersion, particularly if a goal of such a course is to inculcate in students a belief that extracting meaning from data is not beyond their grasp. In similar vein, his chapter on "social statistics" (ch. 15), introduces concepts like proportional reduction of error, linear regression, and curvilinear relationships without granting the student any appreciation for where these techniques come from or how they are derived. In short, TPSR does students and instructors a real disservice by introducing complicated concepts and ideas without providing the tools for the instructor to impart much actual comprehension.

PSRM, on the other hand, functions well as an introduction to statistical reasoning. Johnson and Reynolds proceed smoothly from techniques of data collection to principles of data analysis, from univariate through bivariate and ultimately multivariate data analysis. The authors present complicated ideas effectively, achieving a well-struck balance between statistical theory, substantive interpretation, and the needs and limitations of the scientific method.

Organizationally, the quantitative methods portion of PSRM could be broken apart slightly for greater comprehension. The chapter on bivariate analysis covers crosstabs, statistical independence, three measures of association, statistical hypothesis testing, ANOVA, and linear regression. The latter two topics would seem to fit more directly into the next chapter, on multivariate analysis, which provides a very effective link between multiple regression and nonlinear models, specifically logistic regression. In teaching methods, I believe it is more feasible for an instructor to remove rigor or detail from a text than it is to impart the same; and the design and execution of PSRM makes this easy, in contrast to Babbie's text.

It is unlikely that any single text could provide a one size fits all solution to the multiple needs of undergraduate methods students and faculty. That being said, *Political Science Research Methods* comes closest, at least among the works under review here. Johnson & Reynolds have put together a rigorous methods textbook that couples accessible contemporary political science research with a strong focus on uncovering substantive meaning from the empirical political world. Babbie's *The Practice of Social Research*, while strong in several respects, falls short on at least two counts, as its interdisciplinary nature and lack of statistical rigor would force the instructor to incorporate or adopt too many ancillary materials.

In Memoriam

John T. Williams

John T. Williams died Monday, September 13, 2004, at his home in Riverside. John was born April 14, 1958, in Odessa, Texas. John is survived by his wife of sixteen years, Ilona M. Hajdu, and one daughter, Miriam Claire Williams. John is also survived by his mother, Joyce Elam, of Denton Texas, and two sisters, Celeste Williams, of Arlington, Texas, and Melanie Williams, of Richardson, Texas. He was preceded in death by his father, John T. Williams.

At the time of his death, Williams was Professor and Chair in the Department of Political Science, University of

California Riverside. He received three degrees in Political Science, a Bachelor of Arts (1979) and Master of Arts (1981) from North Texas State University (later renamed the University of North Texas) and a Ph.D. (1987) from the University of Minnesota. Before moving to Riverside in 2001, Professor Williams held academic positions at the University of Illinois Chicago (1985-1990) and at Indiana University in Bloomington (1990-2001). At Indiana, he also served as the department's Director of Graduate Studies (1996-2001).

Prof. Williams was a nationally recognized scholar in the use of statistical methods in the study of political economy and public policy. He co-authored two books: Compound Dilemmas: Democracy, Collective Action, and Superpower Rivalry (University of Michigan Press, 2001) and Public Policy Analysis: A Political Economy Approach (Houghton Mifflin, 2000). He published over twenty journal articles and book chapters on a wide range of topics, ranging from macroeconomic policy to defense spending to forest resource management. He was a leader in the application of new methods of statistical analysis to political science, especially the use of vector autoregression (VAR), Bayesian, and event count time series models.

In the undergraduate classroom, Williams was known for his innovative courses on such topics as law and economics, water resources, and the political economy of sports policy. Throughout his tragically-short career, Prof. Williams demonstrated an intense commitment to the training of graduate students in the latest advances in statistical methodology. He taught graduate seminars in time series, maximum likelihood, and other methods of statistical analysis. He taught time series analysis at the ICPSR (Interuniversity Consortium for Political and Social Research) Summer Training Program virtually every year from 1989 to the end of his career. Williams was especially active in the Political Methodology section of the American Political Science Association. He was a regular participant in the yearly conferences of this section, and hosted the 1995 meeting in Bloomington, Indiana.

Known as Johnny to family and friends in the Denton area and as John to those who met him in graduate school and in the early stages of his career, all his professional colleagues later came to know him best as jotwilli. Although originally selected by an impersonal address assignment algorithm at Indiana University, this e-mail moniker fit him like a glove. To anyone who knew him, jotwilli conveys the vibrant combination of energy, enthusiasm, expertise, and creative playfulness that characterized both his professional work and his personality.

In recognition of his contribution to graduate training in the field of political science, the Political Methodology Section of the American Political Science Review has established the Jotwilli (John T. Williams) Travel Fellowship to support graduate students presenting papers at professional conferences or participating in specialized training programs. Each year, recipients of this award will be selected by a committee of his colleagues from that section. Contributions towards the establishment of this fellowship can be sent to Professor John Aldrich, Department of Political Science, Box 90204, Duke University, Durham, NC 27708-0204. Please make your checks or money orders out to the Society for Political Methodology.

Patrick Brandt
Michael McGinnis
Burt Monroe
John Aldrich

The Political Methodologist
Department of Political Science
University of California at Los Angeles
Los Angeles, CA 90095

Nonprofit Org.
U.S. Postage
Paid
UCLA

The Political Methodologist is the newsletter of the Political Methodology Section of the American Political Science Association. I gratefully acknowledge the support of the Department of Political Science of the University of California at Los Angeles in helping to defray the editorial and production costs of the newsletter.

Subscriptions to *TPM* are free to members of the APSA's Methodology Section. Please contact APSA (202 483-2512, `http://www.apsanet.org/about/membership-form-1.cfm`) to join the section. Dues are $25.00 per year and include a free subscription to *Political Analysis*, the quarterly journal of the section.

Submissions to *TPM* are always welcome. Articles should be sent to the editor by e-mail (jblewis@ucla.edu) if possible. Alternatively, submissions can be made on diskette as plain ascii files sent to Jeffrey B. Lewis, Department of Political Science, 4289 Bunche Hall, University of California at Los Angeles, Los Angeles, CA 90095-1475. LaTeX format files are especially encouraged. See the *TPM* web-site, `http://polmeth.wustl.edu/tpm.html`, for the latest information and for downloadable versions of previous issues of *The Political Methodologist*.

TPM was produced using LaTeX on a PC running MikTex and WinEdt and a Mac running teTex and TextEdit.