

R Software for Regression with Inferential Statistics (2sls-is)

2sls-is/causality project 2sLS-isReOrg6aShort4.doc Aug 8, 2011
Sept 4 changed

Sortened Draft 4 SFI Not for general distribution

Doug White, Ren Feng, Giorgio Gosti, Tolga Oztan

R is open-source, cooperatively developed and free software with extensive capabilities. The advantage of working in the R statistical computing environment is the flexibility not only of executing existing programs but writing simple functions that operate on data or program output. Importing documented packages from the extensive R archives packages make it easy to extend these benefits. Successive improvements of software can be done by different authors, as is the case here. Potential collaborators or users and developers can communicate in further open-access program developments. The development of this two-stage least squares–inferential statistics (2sls-is) package builds on an existing R prototype by Eff and Dow (2009) aimed at improving the standard for regression models with autocorrelation controls and missing data imputation, and on James Dow’s initial SCCS.Rdata in which he edited the missing value codes to conform to conventions in R. The 2sls-is package provides further inferential statistics to evaluate regression models, and outputs results in a forms that are useful for networks of variables, causal inference, and structural equation modeling (SEM). Tests of significance can corrected, and standard output variables are extended with an “effect ratio” measure that provides a cross-check for evaluating independent variables to include in a structural model.

An R program is provided, 2sls-is, that replicates Eff and Dow’s (2009) prototype in controlling for autocorrelation in regression models and multiple-sample imputation of missing data using Rubin’s formulas (1987, Little and Rubin 2010) to combine estimates for adjusted significance tests, tests of endogeneity, and omitted variables. Aspects of the 2sls-is inferential statistics package are described in three parts. Part I shows how the new software generates statistical distributions to supplement significance tests in evaluating models: i.e., tests that use new results saved from fitting large random training subsamples (Does the model hold for random samples of the data?), holdout test data (Are the model coefficients from the training sample predictive of similar results such as R^2 , significance, and other “holdout test” subsample results?), and effect ratios computed to apply to networks of variables with multiple dependent variables coded for overlapping samples. These concern problems of regression on a single variable that develop out of the Eff and Dow (2009) software, design of the software package, and some of the more stringent statistics that may be used to apply a higher standard to regression models if they combine into networks of variables suitable for structural equation modeling (SEM). Part II validates the new package by replication of results that establish equivalence of results with Eff and Dow (2009) and Brown and Eff (2010) and includes evaluation of an improved extension of the Brown and Eff model. Part III goes beyond these extensions of Eff and Dow to address logical problems and solution concepts for networks-of-variables models in controls for bias and adjusting for common causes, and the problems involved in combining regressions for single dependent variables into networks of variables suitable for structural equation modeling (SEM) and causal graphs.

I. Regression models with controls for autocorrelation

In survey and comparative research the nonindependence of cases, designated long ago as Galton’s problem, “applies to all nonexperimental studies and to experimental design as well” (Wikipedia: Galton’s problem) in overcoming problems due to autocorrelation. Nonindependence of cases creates spurious correlations and meaningless tests of the null hypothesis (aka “significance tests”). Significance is the most misused of descriptive statistics (Henkel and Morrison 1970). Network effects were evaluated by Dow, White and Burton (1982), Dow, Burton, White, and Reitz (1984), and Dow, Burton and White (1982) as a major unsolved problem in regression analysis of survey data and cross-cultural research. The problem with 2sls as an improved solution to autocorrelation, however, and in all regression, is how to select and evaluate the variables (Greenland 2010), including how to justify the variables that are chosen, or how to avoid overfitting a model with spurious variables that happen to enhance significance or raise R^2 . To avoid these problems it is important to have prior theory, prior evidence, a deductive argument to connect theory to choices of variables, an acute sense of sources of bias (Dow 2007), inferential statistics that avoid reliance on significance tests, use of alternative samples, and replication of results. This makes any study that relies on correlations or regression models alone problematic. Once these complex problems are recognized and dealt with (Greenland 2010), the type of inferential statistics that are saved in the 2sls-is model provide an important level of overall model evaluation that is commonly ignored in regression and 2sls regression modeling.

The features of Eff and Dow (2009) and the earlier advances on which this new R package is built are encompassed within the new software, as summarized by Eff and Rionero (2011):

“We follow the methodology developed for cross-cultural data sets by Dow (2007), Dow and Eff (2009a, 2009b), Eff and Dow (2008, 2009), and Eff (2008). Multiple imputation addresses the problem of missing data. Weight matrices for geographical and linguistic proximity are used to model the influence of cultural transmission via borrowing and inheritance, respectively. The geographical distance weight matrix is created by calculating the great circle distance between centroids for each pair of countries, and the language phylogeny weight matrix is taken from Eff (2008). Since the two weight matrices are highly correlated, it is difficult to identify separate borrowing and ancestry effects. We accomplish this by employing the composite weight matrix method presented in Dow and Eff (2009a:142), and also used in Brown and Eff (2010): combining the two weight matrices, we select the combined matrix that results in highest model R2. Our model takes the form:

$$y = \rho W y + X \beta + \varepsilon \quad (1)$$

$$[\rho = \rho_0 + \rho_1 W_1 y + \rho_2 W_2 y + \dots + \rho_j W_j y + X \beta + \varepsilon \text{ (Dow 2007:347, Eff and Dow:13)}]$$

In equation 1, y is the dependent variable, and $W y$ is the composite weight matrix times the dependent variable, giving us a measure of cultural transmission. $W y$ is endogenous, which requires that the model be estimated using two-stage least squares (Dow 2007).”

Equation 1 measures autocorrelation in y by inclusion of $W y$ where the W matrices are row normalized: When multiplied by a column variable such as y (or those in X), the result is “neighborhood averaged” for each observation, within the range of column variable y , from the values of X in proximity-weighted societies. More broadly, through the linkages among the cases as measured by $W y$, “a loop of causality such as network effects that link the independent and dependent variables of a model leads to endogeneity” (Wikipedia 2011: Endogeneity). A variable in regression is *endogenous* if it is correlated with the error term ε , i.e., autocorrelated. If an **independent variable is endogenous its effect cannot be correctly estimated** because the ε prediction errors cannot be assumed to be independently and identically distributed (*iid*). Inclusion of autocorrelation measure $W y$ in the regression model has as its goal to make X and ε *conditionally* independent (in Dawid’s 1979 notation: $X \perp \varepsilon | W y$) so although endogeneity in equation 1 may remain for correlations between y and $W y$ or $W y$ and ε , *controlling* for $W y$ may eliminate the $X \sim \varepsilon$ correlations, satisfying $X \perp \varepsilon | W y$ as *conditional* exogeneity. This condition has to be tested *post hoc* using LaGrange Multiplier test for network dependence in the residuals (H0: no autocorrelation), i.e., correlation of ε with the W matrices, and Hausman tests (Wooldridge 2006:328-332) for independent variables. For ε in the regression results does $W \varepsilon \sim 0$ or depart significantly from 0? Regression results for equation 1, however, will split endogeneity with ε between X and $W y$. In a simpler model, $y = X \beta + \varepsilon$, autocorrelation due to missing the $W y$ term in equation 1 is displaced into loops of causality through network effects for some independent variables (X_1, \dots, X_m) that correlate with the ε . The solution for equation 1 thus cannot be achieved directly but requires two stages of regression and a test of conditional independence $W \varepsilon \sim 0 \Rightarrow X \perp \varepsilon | W y$ from results of the second stage.

Pg2Bib

A variable in regression is *exogenous* if it is uncorrelated or conditionally uncorrelated with the error term ε ; and only the effects of exogenous variables can be efficiently estimated in regression analysis. The solution for equation 1 is to test whether conditional control given $W y$ creates exogeneity for X . To achieve this goal ($X \perp \varepsilon | W$) for X and the error term requires a first-stage regression that predicts not the dependent variable y but makes an independent estimate of the control variable $W y$ in (1) from $W X$, i.e., from appropriate transformations of the independent variable as neighborhood effects, given in equation 2.

$$W y = \alpha_0 + \alpha W X + \varepsilon' \quad (\text{estimate } \alpha, \varphi, \varepsilon')$$

$$W y = \alpha_0 + \alpha_j \sum_{j=1}^m W_j X_j + \varepsilon' \quad (2)$$

$W y$ represents the local W -neighborhood (proximity-weighted) averages of y and similarly for each $W X_i$. Thus, equation 2 represents a neighborhood-level regression of X -neighborhoods on y -neighborhoods. The same W matrices are used for both y and X , so that if the network effects that produce autocorrelation in equation 1 are captured by the W networks, then $W X$ and the error term ε' are more likely to be independent as measured by W in regression equation 2, an assumption that can be tested, e.g., as to whether $W \varepsilon' \sim 0$. (Higher order effects like transitivity or clustering within W are already measured in W and do not require separate modeling or controls. Additional W matrices may be added to the model, but efficient estimation of X may be achieved with an initial set.)¹ Dropping the error term in equation 2 gives as a result an estimate of $W y$ that can be substituted into equation 1 (see equation 1' below).

¹ Among the features to be added to the 2-sls-is program are the equation-2 stage-one regression results for each independent variable and the Hausman test for whether $W_j \varepsilon' \sim 0$ for each W matrix. The idea here is that if the average neighborhood properties of y and X given W have no further autocorrelation, then ε may be random and X in equation 2 is exogenous; if so $W \varepsilon \sim 0$, subject to the Hausman test.

$$\hat{W}y = \alpha_0 + \alpha WX \quad (\text{drop } \varepsilon' \text{ from equation 2}) \quad (2')$$

Stage-two regression inserts the $\hat{W}y$ estimate into equation 1 to create 1'. (Because $\hat{W}y$ is an estimate, double primes ('') are used for new coefficients in model 1', as they vary from those in equation 1.) Here, whether X is exogenous (uncorrelated with ε'') is directly measured by a Hausman (1978) test of whether $W_j \varepsilon'' \sim 0$ for W_j in equation 1'. A Lagrange multiplier (LM) test measures whether the model is misspecified because y and ε'' are correlated. If these tests are satisfied, then the β'' coefficients and other measures for the effects of variables X on y can be efficiently estimated (2sls estimates of this sort have been shown to converge to maximal likelihood estimates, so this result can also be checked),² including accurate results for statistical significance.

$$\begin{aligned} (1, \text{ substituting } WX \text{ for } \hat{W}y \text{ from } 2') \quad & y = \hat{W}y + X\beta'' + \varepsilon'' \\ & y = \alpha_0 + \alpha WX + \beta''_0 + X\beta'' + \varepsilon'' \\ & y = \alpha_0 + \beta''_0 + \sum_{j=1}^r \alpha_j W_j X_j + \sum_{i=1}^r X_i \beta''_i + \varepsilon'' \end{aligned} \quad F \quad (1')$$

The ε'' in equation 1', however, like that in equation 1, is simply a prediction error with no causal content for prediction from X to y. Thus the question of which X, controlling for autocorrelation, have *causal* effects on y, as opposed to statistical predictions, has not been solved. This reflects the weakness of choosing independent variables for a regression model based on higher significance tests and R^2 results alone. Reliance on significance tests in the final regression model is a defect of the Eff and Dow (2009) methodology. Pg3Bib1

Why model? Advances in Econometrics and Causal Modeling

Economists use more theoretically motivated forms of regression to evaluate the effect of “interventions” as deliberate policy actions and incentives where the modeling problem is one of establishing the causes or drivers of change and effects on change. Hurwitz (1962) explained the concept of “structural models” in econometrics that identify how interventions create changes in parameters, equations, and observable or unobservable variables that are elements of the model. The model aims at an accurate characterization of the effect of an intervention on the dependent variable after the intervention. Some variables in the model are causal and others are not, holding fixed other variables that serve as controls. The Lucas (1976) critique is that, holding other equations fixed, changing only a policy equation to predict the effects of a policy change will fail because the other equations will change when the policy changes, i.e., due to response behavior to the policy and its effects. Structural equation models (SEM) date back to path analysis (Wright 1921, 1923, 1934) that uses the solution of simultaneous equations or complex computations of partial correlations. SEM has generated a variety of commercial packages for path analysis of networks of variables (LISREL, EQS, Amos, CALIS in SAS, SEPATH in Statistica, and Mplus), and *sem* freeware in R (Fox 2006, 2009, Grace 2009). Pg3bib2

Applied economics researchers use regression methods in ways that have benefited from the lessons of SEM and structural modeling (Pearl 2009:27) to distinguish and test claims (White and Lu 2010:4) “that an outcome or response of interest, Y, is structurally generated as

$$Y = r(D, Z, U) \quad (3)$$

where r is an unknown structural function, D represents observed causes of interest, and Z and U are other drivers of Y , where Z is observed and U is not. U represents not just ‘shocks,’ but all factors driving Y that are too costly or too difficult to observe precisely.” If the effects of the drivers are linear, a different form of regression (White and Lu 2010:4) will apply (where symbol \dagger is the transpose), “so

$$Y = D\beta + Z^\dagger \alpha + U \quad (4)$$

where β and α represent the effects of D and Z on Y , respectively.” D includes the primary causes of interest, and makes efficient estimation (“identification”) of the β parameters of interest. This requires only a conditional form of exogeneity where, given $X = (Z^\dagger, W^\dagger)^\dagger$, D is uncorrelated with U , i.e., $D \perp U | X$; and exogeneity need not apply to Z (Z often consists of control variables). Here (White and Lu 2010:31) the W^\dagger can include proxies for U (which include the W matrix controls for autocorrelation of such as those equation 1'), proxies for unobserved drivers of D , or unobserved drivers of D . What this means in practice is that the Hausman (1978) test that diagnoses misspecification, where an independent variable is endogenously correlated with the error term, such as ε'' in equation 1', need only be applied to the core D of observed causes of interest, and not to variables that are included in the model (e.g., as controls) but have no causal content. In statistical theory this means wider latitude to omit

² Ord (1975), for example, shows convergences between maximum likelihood estimation and various alternative models for autocorrelation. A two-stage model that gives additional autocorrelation controls for the error terms is given by Kelejian and Prucha (1998) and programmed using the IMSL program library. Maximum likelihood estimation is computationally challenging when the sample size is large.

tests that could disqualify some causal models because non-core variables are not exogenous. In substantive theory it means much more careful considerations about what variables are justified as causal core variables versus controls, and what variables can be included in a structural model identified by regression methods with the Hausman test of misspecification limited to the core variables. White and Lu (2010) include a straightforward new type of Hausman test of robustness for critical core regression model coefficients that apply to potentially causal variables. This marks another advance over Brown and Eff (2010), which might appear to suggest that their Hausman test (Wooldridge 2006:532-533) must be used uniformly for every independent variable in a regression model.

Pg 4Bib

The 2sls-is package: Building on Advances in Causal Modeling using Regression

The 2sls-is package retains the objectives of Eff and Dow (2009) as a prototype for controlling for autocorrelation (see White 2007 for spatial and linguistic autocorrelation in cross-cultural data), for imputing missing data, and further generalization for use of with survey data in which observations are almost always nonindependent and require instrumental variables or controls to measure endogeneity. The 2sls-is package is also aimed at (1) new inferential train-test statistics fixing coefficients with a large random subsample to test and record variations in significance tests in the remainder of the sample, (2) inclusion of regression models with temporal data, and (3) increasing the potential for generalization of structural equation models (SEM), possibly including dynamics. To that end, the 2sls-is package can be integrated with functions from *sem* freeware in R (Fox 2006), and models of networks of variables discussed in Part III.

The 2sls-is programs are sourced with the following R commands that re-factor the Eff and Dow R code into larger and embedded functional modules to make the new code easier to use and to modify. The user is expected to change only the 2nd source program, but may experiment with the 1st, and may choose a variant of the 4th source program indexed by a different “random subsample” index. The iterated loop for the 4th and 5th source program is optional, and saves inferential statistics that are not part of the Eff and Dow (2009) program.

```
setwd('/Users/drwhite/Documents/3sls/sccs/')           #setup working directory
source('Libraries2sls.R')                             #setup load R libraries
load('sccs.Rdata')                                   #read sccs or other data
source("examples/create/autocorr/test_depvar_wghts.R") #1st source program3 (optional)
source("examples/create/create_EduR_1/create_EduR_1.5Lang.R") #2nd source program (changed for the user's model)
source("R_3_s_ols/two_stage_ols.R")                  #3rd source program (functions defined)
for (i=1:6) {
  source("examples/create/R_run/run_model_70.R")     #4th source program, "random subsample" = 70
  #source("examples/stats/inferential_stats.R")       #5th source program, computes and saves results
}
source("examples/create/R_run/averageAll.R")         #6th source program, save imputed data for OLS
```

In this sequence, the commands set the working directory (`setwd`), which is illustrated for a Macbook but could be a shared directory in a classroom computer lab, install the program `library(sccs)` and read the `data(sccs)`. The user may edit an auxiliary (optional) source program (`test_depvar_wghts.R`) that regresses $y = \rho W y = \sum_j(\rho_j W_{jy})$ to test models of autocorrelation in the dependent variable alone, without consideration of independent variables. To change the two-stage model, including controls for autocorrelation, the user creates or modifies the contents of the 2nd source file (e.g., `create_EduR_1.5Lang.R`), which defines: 1) the dependent variable of a model or set of models and which of the autocorrelation variables (row-normalized W matrices with zeros in the diagonals, either in standard form of such as predefined language and spatial proximities, or including combinations fitted in the first source program) are to be included in the model, 2) giving a name and alias for the class of models for the dependent variable, 3) a frame of user-labeled variables from the `sccs` database that may be retrieved in R with its name, `my_sccs`, or a different name for a different database, 4) a frame of potential (and some excess) `indep_vars` for the user model; and 5) a frame of `restrict_vars` that will be adjusted by the user for goodness-of-fit and other criteria. A useful feature is that after the first (#1st) source program is run the data selected for analysis are in `my_sccs`, which can be further analyzed, e.g., by OLS (`results=lm(dep_var ~ {list of regression variables})`), and the multiply imputed variables are available in `mi_sccs` after the second and third (#2nd, 3rd) source programs are run.

Analytical steps are programmed in the 3rd-5th source programs. The 2nd source code has already specified which W matrices to use to measure nonindependence among cases in the sample as potential controls for autocorrelation. The third source code, `two_stage_ols.R` (2sls) is prepared to fit but not necessarily to optimize the model given inputs

³ This is an optional source code and may be expanded to include the stage-one regression results from the main program.

from how the model is created and how it is run. It is this program that creates dot-products of the independent variable row vectors (size n) by the chosen (or tested, each in turn) row-normalized n by n W matrices to define new row vectors (also size n) that are regressed against the original dependent variable (equation 2). The coefficients of nonindependence, e.g., for language and/or spatial proximity, estimate the “network effects” by which cases affect one another. Once these are fixed the correlation of \hat{w}_y with the dependent variable y is maximized by estimating the α_{ji} parameters, one for each WX matrix product in equation 1'. The fourth source code, `run_model...R`, does missing data estimation, and produces a set of statistical estimation results. These exceed the results of ordinary regression and 2sls, allow choice of a parameter for a 100% sample or 79% of a randomized subsample percentage for a model “train test”, e.g., `run_model_79.R` is identified in the name of the `run_model`. The FOR loop for multiple runs of each `run_model...R` (starting with 6 iterations for an early model but perhaps many more for a finished model), saves results in a new file. The fifth source code, `inferential_stats.R`, used for repeated “randomized subsamples” (if not 100%), evaluates these distributions of results for the independent variables in the train and holdover independent test sample, and saves further inferential statistics. The fourth source code provides inferential statistics on individual variables, while the fifth provides inferential statistics on the overall model. Practical advice for runs of the program is provided in Appendix 1. The sixth source code, `averageAll.R`, averages the multiple independent columns of imputed data, both numeric and ordinal for the full sample ($N=186$ for the SCCS). Another source code `Normalize.R`, is available to convert ordinal distributions to normalized, making the data available in a Probit analysis through the 2sls-is programs. The averaged data also becomes available for OLS regression (R source `Lm` package for linear models), without controls for autocorrelation, or can be analyzed again by the 2sls-is programs.

The purpose of the FOR loop for source codes four and five of the 2sls-is program (`run_model...R`, `inferential_stats.R`) is to save several kinds of inferential statistics for model evaluation, including the train-test inferential testing commonly used by computer scientists for model testing (e.g., split-halves) even if only for testing model R^2 . The key parameter of `run_model...R`, as in `run_model_50.R`, sets a ratio for choice of a subsample (in this example: 50%) that is randomly selected for estimating coefficients in both the first and second stage of the 2sls regressions. A minimum of six iterations of the `run_model` analysis (each taking about 40 seconds or less, depending on the W matrices) might be used, with each run (e.g., four minutes run-time for six iterations) selecting different random samples. These iterations generate distributions of four key parameters used for inferential statistics of the total model. These are: 1) effect ratios (see Pearl 2009: 368) and variable inflation factors (vifs), 2) raw and finished (2nd-stage) correlations between independent variables and the dependant variable and R^2 for the first and second stage regressions, and 3) p-values. Of these, correlations and effect ratios are dimensionless and potentially comparable across different studies (for effect sizes see Nakagawa and Cuthill 2007) while significance tests are not, as they depend on sample size and univariate distributions. Multiple measures of effect ratios and vifs avoid overreliance on significance tests. The following section describes our use of effect ratios and how they differ from conventional measures of effect sizes, which are also dimensionless.

Pg5Bib

Effect ratio heuristic and the variable inflation factor

Fox (2002:27) makes good use of percentages in his illustration of Duncan’s (1961) Occupational-Prestige regression, pointing out for example that “holding education constant, a 1 percent increase in high-income earners is associated on average with an increase of about 0.6 percent in high prestige ratings” (.5987 being the unstandardized regression coefficient for income). This is a “unit change” or *effect ratio* obtained by standardizing the ranges of the independent and dependent variables. The 2sls-is software gives *effect ratios* for each independent variable. It is computed by dividing each independent variable coefficient by its numeric range n_x and multiplying in each case by the numeric range n_y of the dependent variable. This can be used for suitable ordinal variables, logged or exponent transformed variables, and set to 1 for dichotomies: in any case the idea is to unitize the variables, as with percentages. Where necessary, outliers in each n_x and in n_y can be eliminated prior to calculating the ratio. Used in regression, the absolute value of these ratios, for a robust model, will often sum to one. An effect ratio may be high relative to other independent variables that are more significant, or vice versa. In a structural model of a network of variables, effect ratios may be used as comparable measures of effects. Pearl’s (2009:150,367-368) *do* operator, although purely conceptual, is intended to be a practical measure corresponding to an effect ratio.

Pg5Bib

Independent variables in a regression model need not be statistically independent of one another, and the extent of multi-collinearity is measured in regression results by the variable inflation factor (vif) for each variable. The square root of the vif measures how much larger is the standard error of the variable compared with what it would be if that variable were uncorrelated with other independent variables in the regression. Hence division of the effect ratio by the vif gives a *relative effect ratio* (*rer*) measure for which the sum of absolute values ought be less than 1, within error

bounds. The *rer* then, serves as an alternative measure of whether an independent variable should be included in a regression model. Table 1 shows for one out of 10 runs of the Brown-Eff (2010) moral gods model that there is little if any correlation between the *rer* and the significance for the independent variables. This marks a major step forward for bringing regression analysis closer to the language of causal graphs if the units of *rer* resemble those of Pearl's *do* operator. In Table 1 the percent contribution of animal husbandry to subsistence (*anim*) is the main effect according to *rer* but not as significant at external war, which is fifth in *rer*.

Table 1: Relative effect ratio (*rer*) measure compared to significance tests for the Brown-Eff moral gods regression model

1.3/create_EduR_1.5DistLangBrownEff.R			1.4/create_EduR_1.5DistB_Eff.NoPCAP&ANIMxBWEAL_PCsize2EcorichNoCaste.R		
Variable	rer=ratio/vif	p-value	Variable	rer=ratio/vif	p-value
+Anim	0.21660182	0.03525375	+AnimXbwealth	0.318931658	0.00542158
-PCsize	0.18941089	0.05958345	-PCsize2	0.257055683	0.05451395
-PCsize2	0.16787149	0.26711634	-PCsize	0.224945926	0.01437904
+Caststrat	0.15993666	0.13649458	-Ecorich	0.170754717	0.01685227
-Eeextwar	0.15509468	0.01935323	-Eeextwar	0.122015915	0.05574458
+Foodscarc	0.12633625	0.16709657	(later model developed in section II)		
Sum	1.01525178		Sum	1.09370390	

Links between Network Theory and Inferential Autocorrelation Statistics

Testing theories of network influence has been done extensively by Christakis and Fowler (2009, and other studies), their many collaborators, and many other social network researchers. One of the overall findings is that strong ties tend to affect the spread of behavior, through up to three degrees of all-strong ties, while weak ties (Granovetter 1978) tend to affect information flow, attitudes, affect and beliefs. Autocorrelation effects operate as endogenous network variables in a regression model. Comparative cross-cultural hypothesis for studies with network autocorrelation as an element of model estimation would do well to consider the theoretical assumption that societies connected through language proximity are more likely to transmit “strong tie” behavioral variables, while those connected through spatial proximity are more likely to transmit information, affective variables, attitudes, and beliefs through weaker or “weak ties”. Network effects may be considered as a substantive aspect of a model. Some researchers consider “Galton’s problem” of correcting for nonindependence of cases as a misnomer and that studies that include network effects provide “Galton’s Opportunity” (Witkowski 1974) or a “Galton’s Asset” (Korotayev and de Munck 2003) to improve the quality of understanding of interdependence in network-linked social processes. Pg6Bib

The first elements reported in the 2sls-is program results tested against existing cross-cultural models and data are the regression coefficients for the autocorrelation matrices. The first source code regresses language and distance on the user-chosen dependent variable, reports the ρ parameters for W matrices regressed on the dependent variable (equation 2), and provides the user with alternatives for the W matrices, such as including a multiplicative term of language times distance, with zero diagonals and renormalized. The user enters their choice of one or more W matrices in the second source code program. These choices may reflect theory-driven hypotheses influenced, for example, as to whether the dependent variable measures behavior or information and affect, or a multiplicative product of the two. In the second stage of regression, if none of the alternative Instruments (IVs) based on W matrices are found to have high effect ratios or high significance, then independent invention is a plausible alternative hypotheses and the W-matrix Instruments may be dropped from the model. Note however, that the autocorrelation coefficients in stage-one regression also need to be reported separately (from the first source program) because differ from the autocorrelation coefficients in the second stage (2nd-4th source program) of 2sls-is regression where independent variables are included.

One of the problems in interpreting “network effects” in regressions is that they may be confounded with homophily, that is, the tendency of similar people to form ties (rather than the tendency for people who have strong ties to become more similar). Even the formation of random ties within an arbitrarily bounded group will create greater network cohesion (all pairs having high numbers of mutual multiple independent paths), and cohesive interactions for a group bounded only by its level of cohesiveness may create similarities. The effects of cohesion \rightarrow {links, similarities} \rightarrow {similarities, links} \rightarrow cohesion are difficult to separate without temporal observations.

Setting the Random Subsample Training Ratio

Computer scientists typically use independent train-test samples to estimate a model from a large random subsample and test the model for replication on the remaining independent subsample. Parameters estimated from the training

sample are carried over to a holdover sample to observe extent of replication of other features of the model. Use of the FOR loop in our source code #4 acts to: 1) observe, given parameter estimates from in the larger train samples, the replicability of effects (e.g., significance tests) in the smaller (random) test subsample to see if the model still holds, 2) collect distributions of statistics for raw and second-stage correlations, p-values, and effect ratios used to gauge the robustness of the statistical model both for a larger random subsample and a smaller residual subsample where the regression coefficients have been fixed from estimation on the larger training subsample. Choice of a random sample percentage of the total sample sets the random training subsample (such as 70% or 79%, a current maximum). The subsample ratio can be varied from 50% to 79%, which determines the size of the holdover sample (50% to 21%). Reducing the size of either subsample lowers the strength of significance tests, which diminish with smaller sample size. The “run_model” in the fourth source program need not be edited by the user because the percentage parameter is chosen by name (as in `run_model_70.R` or `run_model_79.R`).

II. Replication tests of prototype models and identification of improved models

The examples in Part II illustrate: 1) the types of inferential statistics that can be evaluated, and 2) a test of whether the new 2sls-is program replicates the results of the Eff-Dow (2009) prototype on which the new software is partly based. Since the latter program was used in Brown and Eff’s regression model for Moral gods, two datasets provide a replication test of the software, and the two models can be tested for robustness by new features of 2sls-is.

Testing Robust and Non-Robust Model Replication: Moral Gods (Brown and Eff 2010) and Value of Children (Eff and Dow 2009)

Brown and Eff’s (2010) 2sls model of the predictors of moral gods (`sccs$v238`) from the Standard Cross-Cultural Codes (SCCS) coded variables (White et al. 2009) uses Eff and Dow’s (2009) software and thus provides a good test of our 2sls-is software, which at base simply re-factors the Eff and Dow program. The model is robust, with a substantial R^2 ($= 0.564$). We test whether our 2sls-is software replicates their results. Eff and Dow’s (2009) regression model, on the other hand, is not robust and has a very low R^2 ($= 0.107$). Their dependent variable, the value of children, is unreliable. It was constructed by averaging four high-inference variables (`sccs$v473-v476` for “how society values” boys and girls in early and late childhood; Barry et al. 1977:195-217) that are not often evident from the statements of ethnographers. The “degree to which children are desired or valued” is a very subjective comparative judgment for coders (and coder judgments are missing altogether for 15 societies). But although the Eff and Dow model has very low R^2 , the 2sls-is program using a 100% sample replicates their results. Our 2sls-is refactoring of their 2sls software, as intended, produces the same results, with the only differences due to estimating of missing data (with small random variations in results that do not substantially change the regression coefficients).

The 2sls-is train-test results for Eff and Dow’s model, however, illustrate the problem of overfitting by using significance tests. In ten training runs using random 79% subsamples of the data, the R^2 varies between 0.15 and 0.092. In the four cases where $R^2 < 0.126$ in the ten sets of results, the model R^2 ranges between 0.131 and 0.065, weak indicators for robustness. In the six training runs where $R^2 > 0.126$ (17% higher than the Eff-Dow $R^2 = .107$), however, the test result R^2 are all *negative*: they vary between -0.365 and -0.144 , which indicates replication failure. Four of these six cases of replication failure show a non-significant ($p > .0.10$) settype variable (`sccs$v234`, fixity of settlement). In the other two cases the femsubs variable (`sccs$v890`, percent female contribution to subsistence) is non-significant ($p > .0.10$). Thus, random variation in the sample that reduces significance for one of these two weaker variable is associated with inflated R^2 predictiveness for combinations of other variables. This illustrates the dangers in relying on significance tests alone to choose final variables for a regression model. These results show the utility of the train-test inferential statistics. Since in this case variable settype is the second lowest in its effect ratios but femsubs is the second highest, the conclusion would be to test their model after eliminating the settype variable. When this is done, however, other variables become nonsignificant. Thus the model is degenerate when tested by the additional inferential statistics of the 2sls-is software. To the credit of the 2sls-is train-test comparisons, the replicability of the Eff-Dow value of children regression model is put in doubt.

Pg7Bib

Replicating Brown and Eff’s (2010) model provides an even better test of how well our algorithm replicates Eff and Dow’s software results. Their dependent variable (`sccs$v238`, Moral gods), as to whether there are high gods and if so, whether the high gods are neutral or malevolent towards humans or supportive of human morality, was defined by an expert in comparative religion, Guy Swanson (1960), who did the initial coding. Murdock (1967) coded this variable for the Ethnographic Atlas of 1270 societies without any reported difficulty. The SCCS database includes Murdock’s coding from the Atlas for 168 of the 186 cases in the sample, with 18 cases having missing data. A priori, this is a very likely to be a high-concordance variable that allows reliable judgments by coders for a given society, given information

from the ethnographic sources. Results of a FOR loop of 4 runs of the Brown and Eff (2010) model with the 2sls-is source code, set at a 79% random subsample, are viewable from the model are posted at <http://bit.ly/jNZjSa>, short for the page on the intersciwiki where our regression and causal graph results are stored. The train sample results show that the model is robustly estimated, e.g., with average dependent variable $R^2 = .467$. Further, of the 28 significance tests for seven variables, 21 were significant at $p < .10$, although the PCAP variable (sccs\$v921, v928 1st principal component of PCA, measuring agricultural potential) mostly lacked significance ($p = .37, .22, .21$ and $< .10$). (In six extra runs $p = .40, .30, .17, .16, .11, .02$). Three other variables were nonsignificant at $p > .10$ in one out of four runs (foodscarc $p = .52$, anim $p = .44$, and PCsize2 $p = .29$). (For 6 extra tests PCsize2 $p = .28, .26$, and four significant values). Test results from the 21% holdout sample ($n = 39$ cases) tend to replicate at a somewhat lower level in significance tests, with 68% of the p-values $< .10$, 25% $< .15$, and 7% $> .15$. Brown and Eff (2010) is thus a robust model even for a 79% random subsample. (We have yet to reset our program to run a 100% sample to get a more exact comparison, where PCAP would be more significant on balance.)

Diagnostic statistics for regression results were part of the original Eff and Dow (2009) program and are included in the new 2sls-is software. They include whether: 1) the relationship of the independent and dependent variables are linear (RESET test), 2) the appropriate variables are dropped, i.e., whether those of the larger list of indep_vars that were not included in the restrict_vars list were significant predictors of the dependent variables (Wald test), 3) the error terms for the second-stage regression were not bunched (B-P test), 5) the residuals had no spatial lag for language autocorrelation (LM test). Test 4), however, fails for 2sls-is but passes for Brown and Eff, where $p = 0.370$ (N.S.) but three out of four of the 2sls-is tests have $p < .05$. Given that there are a total of 25 significance tests here, however, it is within the bounds of the null hypothesis that one to two of the p-values in a given row will depart from significance at $p > .10$. With multiple significance tests we expect to get some number of type I errors. One way to deal with this is Holm's (1979) simple sequentially rejective multiple test procedure (see Rice 1989) at some level of significance, e.g., $\alpha = .10$ for a family of k tests, where the p-value of the i'th test in order of significance is rejected if its p-value $\leq \alpha / (1+k-i)$. In the Brown and Eff and Eff and Dow models, each with six p-values $< .10$, only the top three values remain significant in the sequential (Bonferroni) test, and only two if we set $\alpha = .10$.

The lower part of Table 2 shows the output coefficients and p-values (deleting the foodscarc variable that was significant at $p > .50$), for which the top three variables in significance (eextwar, anim, and PCsize2) retain significance in the Bonferroni test at $p < .10$ and two at $p < .05$. The vifs for three variables indicate covariation between anim, PCsize, and (probably) residual neighborhood clusters not removed by the Instrumental Variable (IV) for autocorrelation. This would account for the high effect ratios sums for the independent variables. Dividing each by its vif, the sum is 1.14, reasonably close to one but indicative of a good model. Three variables (distance, anim, and eextwar) show a noticeably lower correlation due to the IV compared to the raw correlations without the IV control. Given the Bonferroni and train-test results, the PCAP and caststrat variables might best be considered spurious.

Table 2: Diagnostic tests for the Brown and Eff (2010) regression model for moral gods

Significance levels for Diagnostic tests		Brown and Eff (2010)		1.3/create_EduR_1.5DistLangBrownEff.R -----Four runs of 2sls-is (2011, run 79%)-----							
				a	b	c	d				
RESET test. H0: model has correct functional form		0.336		0.316	0.279	0.504	0.737 (all pvalues $p > .10$)				
Wald test. H0: appropriate variables dropped		0.402		0.343	0.141	0.286	0.152 (all pvalues $p > .10$)				
Breusch-Pagan test. H0: residuals homoskedastic		0.646		0.821	0.796	0.403	0.629 (all pvalues $p > .10$)				
Shapiro-Wilk test. H0: residuals normal		0.370		0.044	0.307	0.011	0.006 (some pvalues $< .10$)				
LM test. H0: Spatial lag (distance) not endogenous		0.298		0.446	0.845	0.393	0.524 (all pvalues $p > .10$)				
Variable (Intercept)	coef range	effect	ratio	Fstat	ddf	pvalue	VIF	abs.ratio	%coded	raw.cor	part.cor
distance	0.955 2.000	1.910	0.637	47.877	1282406.0	0.00000000	1.573	0.637	N.A.	0.608	>> 0.117
PCAP	0.041 19.920	0.808	0.269	3.633	3068761.3	0.05664798	1.150	0.269	1.00	0.090	0.119
anim	0.099 9.000	0.895	0.298	5.478	2021074.6	0.01925616	1.546	0.298	1.00	0.503	>> 0.278
PCsize	-0.093 8.061	-0.749	-0.250	4.430	9541935.1	0.03530595	1.442	0.250	1.00	-0.232	-0.123
PCsize2	-0.037 28.847	-1.072	-0.357	4.923	6601312.9	0.02649616	1.259	0.357	1.00	0.040	-0.089
caststrat	0.212 3.000	0.637	0.212	4.038	473449.9	0.04449298	1.122	0.212	0.97	0.279	0.211
eextwar	-0.011 88.000	-0.987	-0.329	6.453	4398725.5	0.01107611	1.061	0.329	1.00	0.000	> -0.136
Train R2:final model	Train R2:IV_distance										
	0.4389667	0.9790109									
	Fstat	df	pvalue								
RESET	3.439	114302.56	0.064	(some variables need to be logged or exponentiated).							
Wald.on.restrs	2.204	309.99	0.139								
NCV	0.021	538439.27	0.884								
SW.normal	8.289	1932682.43	0.004	(residuals not normally distributed)							
lag..distance	1.255	221418.47	0.263								
coef range	effect	ratio	Fstat	ddf	pvalue	VIF	abs.ratio	%coded	raw.cor	part.cor	
(Intercept)	0.014	NA	NA	NA	0.003	1190183.4	0.95602007	NA	NA	0	0.000
distance	0.955 2.000	1.909	0.636	46.644	622987.5	0.00000000	1.563	0.636	0	0.608	0.117
anim	0.129 9.000	1.158	0.386	9.543	2130162.8	0.00200720	1.440	0.386	1	0.503	0.279
PCsize	-0.081 8.061	-0.654	-0.218	3.700	4297605.8	0.05442355	1.277	0.218	1	-0.232	-0.122
PCsize2	-0.036 28.847	-1.025	-0.342	4.410	6139973.2	0.03573830	1.246	0.342	1	0.040	-0.090

```

eeextwar    -0.011 88.000 -0.939 -0.313  5.677 14612136.6 0.01718586 1.058    0.313    1    0.000   -0.137
Train  R2:final model  Train R2:IV_distance
          0.4140441          0.9786300
          Fstat          df pvalue
RESET          1.044 1.034383e+06  0.307
Wald.on.restrs 4.212 1.480818e+03  0.040
NCV           0.317 4.295854e+06  0.573
SW.normal     6.454 1.196561e+08  0.011
lag..distance 1.128 1.032930e+05  0.288

```

The model has very high significance for distance autocorrelation (in all runs, $p < .0000001$). Brown and Eff (2010:12) concluded that

“cultural transmission, Wy, turns out to be overwhelmingly the most important force conditioning the presence of moralizing gods, and that transmission is geographic, based on diffusion across space, rather than linguistic, based on transmission from a common ancestor.”

Language autocorrelation, however, is also significant at $p < .02$ in all of the 2sls-is runs, while distance multiplied by language has at least $p < .00001$ and often $p < .0000001$, similar to that of distance alone. In the alternative 2sls-is, however, it is only distance that is significant.

For the Brown-Eff model, the 2sls-is results show the correlations before and after the autocorrelation controls for distance are significantly different for anim (scs\$*v206*, percentage on animal husbandry for subsistence), becoming less positive in the second-stage correlation, with a small positive effect on moral gods, controlling for other independent variables. PCsize (scs\$*v63*, *v237* 1st principal component for community size) also shows a diminishing negative (second-stage) correlation, but remains a negative effect with autocorrelation controls. For an autocorrelation control of distance multiplied by language the difference between raw and second-stage correlations with the dependent is similar, but also occurs for the caststrat variables (less positive) and eextwar (more negative). As a behavioral variable, the inclusion of language multiplied by distance along with a distance effect indicates less war as a positive effect on more moral gods. Attention can be given as to how different or composite autocorrelation controls might affect regression results.

Although 2sls-is software replicates the Eff and Dow’s (2009) 2sls results for the Brown and Eff model diagnostics, the authors’ fitted model is not necessarily the best that could be achieved. Given that the Brown and Eff model of moral gods is robust, can we find improvements to the model that demonstrate what 2sls-is adds to 2sls alone? We found a slightly better model that has six rather than seven variables, simplifying the model by substituting ecorich⁴ for both PCAP (the weakest of the Brown-Eff model variables) and foodscarc, drops the caste stratification variable (the second weakest of the Brown-Eff model variables), and substitutes animXbwealth for anim, which has better explanatory value in measuring an aspect of pastoral exchange economies that might link moral gods to regulation of potentially unequal exchange where animals that produce valued fertility of stock in exchange for valued fertility in children. This model, even with the caste variable, results in 24 of 28 variables that test for significance at $p < .10$. The six-variable model⁵, with an average $R^2 = .494$, is 6.5% higher than the Eff-Brown seven-variable model. The five-variable model, for which some results are shown on the right half of Table 1, has $R^2 = .446$, and posits as predictors of high gods that also tend to be more concerned with human morality: a poor environment (negative coefficient for ecorich, p -value=.005), small communities (negative coefficient for PCsize and PCsize2, p -values=.05, .02), a pastoral economy with exchange of brides for animals (positive coefficient for animXbwealth, p -value=.02), and external war (negative coefficient for eextwar, p -value=.02). The right side of Table 1 shows the effect ratios divided by the vifs for each independent variable in the model, and shows increase over the Brown-Eff model results for: 1) the relative effect ratios (*rer*), which increase about 8% on average and 2) the significance tests, on average seven times lower (.016 versus .114). A further conceptual improvement of the model (four variables, average $R^2 = .431$), substitutes lopopdenmoney=(8-scs\$*v64*)*(1+(scs\$*v155*=1)*1) for the small community variables (first principal components of scs\$*v63*, scs\$*v237*) and

III: Networks of Variables (SEM)

Linked dependent variables, correcting for biased subsamples, and networks of variables

Regression models treat single dependent variables. To link regression results into a network of variables with both direct and indirect effects poses some new problems. These networks usually require 1) tests to check for spurious conclusions from the constituent regression models for dependent variables and 2) identification of and further

⁴ Ecorich groups scs\$*v857* categories to “harsher” codes 1&2&6, “temperate or rainforest” codes 3&4, and “savannah” code 5.

⁵ <http://bit.ly/jNZjSa> links to http://interse.ss.uci.edu/wiki/index.php/EduR-1.2_LangOnly/DistanceOnly_Brown_and_Eff_2010#79.25_train_21.25_test_Distance_June_6.2C_2011

corrections for spurious correlations in the larger network of variables, e.g., correcting for common causes and bias in subsamples. A major problem in linking networks of variables also occurs with missing data, and intensifies when the subsamples for dependent variables differ substantially. In this case, results must be tested for biased subsamples (Robins, Scharfstein and Rotnitzky 1999; Greenland 2010). An extreme example is provided from the SCCS database with variables from studies like Paige and Paige (1981), which contributed coded only the prestate societies for their variables. A dummy variable for state versus prestate societies would allow controls for relationships among many variables for cases in and out of the sample but for the Paige and Paige variables themselves lack sample comparisons with their variables, coded for state-level societies. To use these variables correctly, the researcher may have to select and extend their codes on these variables on a good-sized random subsample of the state-level societies.

Pg8bib

Where samples in the original studies from which data are drawn are equivalent except for missing data, or randomly drawn from the total sample, a similar but less serious problem is created by substantial differences in subsamples or missing data for different variables. Imputation of missing data does not correct for the possibility that these subsamples may be biased. The use of dummy variables for moderate to large missing-data subsamples for single variables can test and potentially control for such biases. This is especially effective when the author of an SCCS study coded variables for every other society (odd or even) in the list of societies, every third, or every fourth society. The more the imputation of missing data, however, the more the opportunities for bias. In principle, then, regression results that link dependent variables can be combined, either by constructing dummy variables for missing data bias, or by new coding on random subsamples of cases that were intentionally excluded from coding in a subsample focusing on particular types of cases.

Networks of variables open the possibility for SEM, or structural equation modeling. The long history of SEM and path analysis (Wright 1921, 1923, 1934) has generated a variety of commercial packages for path analysis of networks of variables (LISREL, EQS, Amos, CALIS in SAS, SEPATH in Statistica, and Mplus), and *sem* freeware in R (Fox 2006, 2009, Grace 2009). SEM uses the solution of simultaneous equations or complex computations of partial correlations. In purely structural graphical form, SEM model estimation is equivalent to causal graph analysis, as shown by Pearl (2009, 2011). As an example of the crossovers of these equivalencies, and crossovers of software, the Commentator program (Kyono 2010) couples SEM results of EQS software to the causal-separation concepts of causal graphs articulated by Pearl (2011).

Pg9bib

In the next section we show why it is virtually impossible to create a valid SEM or causal model beyond a limited scale in numbers of variables. That is, while it is possible to construct a network of variables linked by independent/dependent variables in regression analysis with 2sls adjustments for nonindependence of cases, it is impossible to focus on SEM or causal subgraphs for more than very limited clusters of variables. There is an advantage, however, of focusing on smaller clusters of variables where more intensive research can stimulate a gradual accretion of results in fields that possess adequate datasets. Results of regression models of relatively proximal common-causes among relatively small sets of dependent variables, along with variables needed to adjust for various types of bias, can contribute results on potentially causal relationships that bear further investigation.

Structural arrows and SEM: What can we learn about theory from Graphical language?

Regression analysis, even in the best case of complete coding of variables, does not give causal results for a network of variables. Even the solutions in SEM to multiple equations starting with correlations free of endogeneity do not offer a sure-fire method of testing causal models (Pearl 2009:150). For regression-based modeling the “unit change” interpretation of effect ratio coefficients included among our 2sls-is results and that appears in the SEM literature offers a structural interpretation to SEM in which Pearl (2009: 367) makes explicit the equivalence between SEM and causal graphs, an equivalence “operationalized through a precise mathematical definition.”

Pearl’s massive contribution to causality was to formally define concepts and measures as the basis for mathematizing and consistent proofs for a language of graphs that expresses precisely what can and cannot be derived from assumptions about causal relationships among variables. His definition of *causal* is neither deterministic nor probabilistic but refers to a set of structural equations that are functions of the form $x_i = f(pa, u_i)$, $i = 1, \dots, n_i$ (Pearl 2009: 27) where *pa* denotes the (causal) “parents” of *x*, and *u* denotes “disturbances” due to omitted factors (including noise, one would infer). Linear parametric models are the special case, “which have become a standard tool in economics and social science” (Pearl 2009:27; see chapter 5), with parameters α_{ik} that differ from those of ordinary regression coefficients, where

$$x_i = \sum \alpha_{ik} x_k + u_i, \quad i = 1, \dots, n_i \quad (3)$$

“In linear structural equation models, the hypothesized causal relationships between variables can be expressed in the form of a directed graph annotated with coefficients, some fixed *a priori* (usually to zero) and some free to vary” (2009:144). Pearl showed that the same elements of a graphical language for models of effects apply to experiments, observational data and hypothetical operations on changing a variable in a system of relationships while fixing the value of other variables (his *do* operator). He demonstrates why probability distributions alone are the wrong language for causality, and that a better approach is only partly Bayesian – what assumptions have what consequences? – but also partly a consistent mathematical extension of common sense, in the context of particular sets of assumptions about what might be inferred to affect what, what evidence is already given, and what can be taken as valid *a priori* Bayesian beliefs. Precursors for a language of graphs “originate mostly from Physics (Gibbs 1902), Genetics (Wright 1921, 1934) and Economics (Wold 1954)” (Lauritzen 2011). Pg10bib

One assumption of a language of graphs for a network of effects between independent and multiple dependent variables is the possibility of delimiting a system of interaction where there are no variables that have serious common outside causes, or where the chains of common outside causes are so long that their indirect effects can be small enough to be ignored, thus included in the u_i terms of equation (3), given that indirect of causal chains involve multiplication of effects. That is, chained positive or negative effects on a scale from -1 to 1, when multiplied, are decreasing in absolute value the longer the chain of multiplications for magnitude of indirect effects.

Delimitation of the boundaries of a network of effects is a key issue that allows for definition of *structural parameters*. While definitions differ, one variant is that, if all degrees of freedom of a delimited system are observable, a parameter in a theoretical model of the system is structural when it has a distinct structure for its effect. Informally, changing or removing a structural parameter alters system behavior. The delimitation of such systems (and external perturbations) borders on problems of complex systems and complexity sciences that recognize complex interactions among effects over time in bounded networks (which may also have external perturbations or changing boundary conditions). Often, understanding such interactions may involve multiple set of empirical data, simulations, and studies of longitudinal interactions.

For regression analysis, an important question is: Can a structural parameter α between $x \rightarrow y$ ever be equated with a regression coefficient? As Pearl (2009:150) shows: Yes, when all paths between X and Y (or x and y) in their graph are blocked by converging arrows. This is again an assumption about relative isolation of a system of variables in which $x \rightarrow y$ is embedded. A path p between x and y in a causal graph is *d-separated* or blocked (Pearl 2009:16-17) by a set of nodes Z when “ p contains a chain $x \rightarrow m \rightarrow y$ or a fork $x \leftarrow m \rightarrow y$ where m is in Z; or, p contains an inverted fork (collider) $x \rightarrow m \leftarrow y$ such that m is not in Z and such that no descendant of m is in Z.” (A set Z is said to *d-separate* X from Y “if and only if Z blocks every path from a node X to a node Y”; the definition does not apply reversibly to Y and Z). The overwhelming bulk of justification for conclusions about (*d-*) separability and connectivity in causal graphs must come from theory about a particular system of variables and not from an unlimited and thus unknowable set of interactions (Very nice!). Graphical language for structural arrows within such a delimited system can show us where data are relevant, and arrows or their absence in a causal graph show only the possibility of an effect, which may not be demonstrable in a given dataset. But only occasionally, where assumptions about structural arrows greatly limit the possibilities, will data be relevant to causal estimation because a majority of graphical structures of such arrows are irresolvably confounded, in which case data do not resolve questions of causality. Nevertheless a structural model is sometimes identifiable (i.e., values of structural parameters on the arrows could be inferred if valid treatments of empirical data were available), even if it cannot be fully solved (i.e., by unequivocally testing a model of the system with valid data).

Given sufficient advances in understanding the graphical language for identifying potential effects and how they may encapsulate in limited systems, some advances may be made by evaluating regression analysis in which *iid* assumptions are justified in conjunction with models that are strongly justified by past research and theoretical arguments. We can only begin to estimate structural parameters, even of limited theoretical models, if we bypass premature conclusions based on spurious correlations and significance tests and do estimations with 2sls and 2sls-is, experimental data, or opportunistic and carefully qualified quasi-experiments (like the event-based or event-sequence analyses discussed by contributors to Vayda and Walters (2011)). Given the *iid* property in 2sls regression, it is useful to raise the bar further to include the problem of whether the potential results would meet the higher standards of causal separability or interaction rather than fitting a statistical model for the sake of R^2 and statistical significance. (Unfortunately, that bar has rarely been raised in cross-cultural studies. In general, the learning gradient about sociocultural systems is not helped by using inadequate means of fitting models, as is frequently the case in cross-cultural studies.) Still, even with regression that produces *iid* residuals, “there can be no conclusion made regarding the existence or the direction of a cause and effect relationship only from the fact that A and B are correlated”; even “when the relationship between A

and B is statistically significant, a large effect size is observed, or a large part of the variance is explained.” In such cases determining “whether there is an actual cause and effect relationship requires further investigation” (Wikipedia: Correlation does not imply causation). But even in the weak form in incomplete models, identifying effects in models that are only partially tested can make use of the benefits of 2sls and 2sls-is. Pg11bib

Part IV: A Regression approach to networks of variables

Statistical 2sls models, now common in econometrics, genetics, epidemiology, physics and social science, create a context for eliminating poor estimation (e.g., misestimation in significance testing) and biases (e.g., unequal clusters of values in the sample of observed cases) that come from data samples in which the cases are nonindependent or variables are endogenous. Endogeneity can arise as a result of autoregression with autocorrelated errors, from measurement error, simultaneity, omitted variables, and sample selection errors. 2sls and 2sls-is offer ways to correct for endogeneities such as arise from endogenous “social effects” (e.g., spatial and language clustering of cases in samples of societies). In anthropology, for example, some societies may have features that are independently invented, but diffusion and common origin, alone or in combination, are recognized as sources of nonindependence of cases. Many sources of nonindependence need to be recognized and tested or corrected, based on thinking through how different kinds of variables are endogenous or autocorrelated, such as measured in our sample regressions here using regression terms like Wy (multiple network effects on autocorrelation in dependent variable) or WX (similarly, for columns of independent variables).

With the combination of exogeneity and multiple imputation of missing data, significance tests are more efficient and comparable with respect to the sample size for a dependent variable in the context of regression analysis. Significance test biases can be tested with respect to the proportion of imputed missing data cases in the sample for a given variable (the more imputation, the more randomness in the imputations, with measurable bias toward less significance expected in the significance test). More importantly, effect ratios and relative effect ratios (Part II) can be evaluated independently as against statistical significance, so that significance tests should not be the sole basis for rejecting a variable in a statistical model for which the effect ratio is larger than those of other variables in the model. Choosing variables in a regression model on the basis of the relative significance of independence variables in the second stage of 2sls regression is a dubious practice that may lead to failure of model replication when using multiple train-test statistics, for example, that have the advantage of testing independent random subsample replications. Use of other inferential statistical distributions in model testing may offer similar advantages. In addition to 2sls and 2sls-is models for single dependent variables an extension of findings about how different dependent variables are linked into networks of variables presents potential (but not always) soluble approaches to problems such as subsample sizes that differ (e.g., when the subsamples for which missing data are imputed vary with the cases coded for the dependent variable) or delimitation of systems that contain structural parameters that can be partially or completely estimated.

Conclusion

In sum, what is envisaged in the development of a preliminary 2sls-is R package that is useful in the social sciences and observational survey data analysis generally and that potentially connects to R *sem* software (Fox 2006, 2009) – and the mathematics that limit but in some cases enables causal inferences – is a gradient of research possibilities that utilize data, existing models and new simulations, building on theory and plausible Bayesian priors. Model building is not a sole or definitive end, but can open paths to new insights and research. The one example explored here simply aimed to show that the 2sls-is R package replicates in its base component a previous 2sls software package (Eff and Dow 2009) but also goes beyond that package in important ways. When we showed that adding a new train-test component to the base component provided distributions of new inferential statistics we were better able to show the sources of weakness in the model tested as an example by Eff and Dow. The addition of an effect ratio and a relative effect ratio in regression as auxiliaries to significance tests, with support from inferential statistics, helped to identify weaknesses in the Eff-Dow base software due to exclusive reliance on significance testing, even if endogeneities were largely eliminated by Instrumental Variables (such as Wy estimated in equation 2 first-stage regression and substituted as an estimate into equation 1' for a second-stage final regression model).

While 2sls-is software found the Eff-Dow model to lack replication of their model (but not their software), we showed that the 2sls-is software did replicate a more robust model estimated by Brown and Eff (2010) using Eff-Dow 2sls. A pathway to a better alternate model of effects of the dependent variable became evident, however, simply by placing one of the independent variables – dependence on animal husbandry (*anim*) – in the context of the type of socioeconomic exchange (*animXbridealth*) typical of societies that rely more extensively on pastoralism. As a socioeconomic model, this had the effect on the researchers of introducing questions about dynamics into the model

that can be investigated further. This example, opening new questions, illustrates the estimation of models not as an ending in themselves, aiming at conclusive final results, but as an opening to new questions. It is in these kinds of developments that the “higher bar” of asking questions about causality, and the mathematical language of causal separation and interdependencies in networks of variables that may help to push us out of an empirical wasteland of reliance on model-building software to the reopening of more fundamental concepts in social science research whose efficacy can be tested empirically.

Acknowledgements

The authors thank Halbert White, Nihat Ay, Sander Greenland and Judea Pearl for suggestions in the area of causal graphs; Anthon Eff, Malcolm Dow, and Christian Brown for their baseline software and model data; and Scott White for paid programming of the new `2sls-is` R code that refactors the Eff-Dow (2009) R software and for the train-test algorithm and other suggestions of the principal author. We hold none of these contributors responsible for any misstatements made in this draft. We thank Jürgen Jost, Director of the Max Planck Institute for Mathematics in the Sciences, for a generous invitation for our UC Irvine-based research group, including Ren Feng, Giorgio Gosti, and Tolga Oztan, to work at the MPI during the latter two weeks of June, 2011 and to make presentations to MPI students, faculty and researchers. We thank the Santa Fe Institute, and faculty director David Krakauer, for hosting of a similar working group seminar at SFI in August-September 2010, in which Scott White was also able to attend (but not Gosti). We thank the Institute of Mathematical Behavioral Sciences at UCI and its faculty research group in Social Dynamics and Complexity, for supporting the Human Sciences and Complexity videoconferences in which Hal White, Judea Pearl, Sander Greenland and members of Causality project at UCI gave presentations. Those talks can be seen in streaming video at <http://itunes.apple.com/us/itunes-u/human-sciences-complexity/id429669291>.

References - R Software for Regression with Inferential Statistics (`2sls-is`)

Pg1Bib

- Dow, Malcolm M. 2007. Galton's Problem as Multiple Network Autocorrelation Effects. *Cross-Cultural Research* 41:336-363. <http://ccr.sagepub.com/content/41/4/336.short>
- Dow, Malcolm M., Douglas R. White, and Michael L. Burton. 1982. Multivariate Modeling with Interdependent Network Data. *Cross-Cultural Research* 17(3-4): 216-245. <http://ccr.sagepub.com/cgi/content/abstract/17/3-4/216>
- Dow, Malcolm M., Michael L. Burton, Douglas R. White, Karl P. Reitz. 1984. Galton's Problem as Network Autocorrelation. *American Ethnologist* 11(4):754-770. <http://www.jstor.org/pss/644404>
- Dow, Malcolm M., Michael L. Burton, & Douglas R. White. 1982. Network Autocorrelation: A Simulation Study of a Foundational Problem in the Social Sciences. *Social Networks* 4(2): 169-200. http://eclectic.ss.uci.edu/~drwhite/pw/NetworkAutocorrelation_A_SimulationStudy.pdf
- Greenland, Sander. 2010. Overthrowing the tyranny of null hypotheses hidden in causal diagrams. Ch. 22 in: Dechter, R., Geffner, H., and Halpern, J.Y. (eds.). *Heuristics, Probabilities, and Causality: A Tribute to Judea Pearl*. London: College Press, 365-382. http://intersci.ss.uci.edu/wiki/pdf/Pearl/22_Greenland.pdf
- Henkel, Ramon, and Denton Morrison (Eds.). 1970. *The Significance Test Controversy: A Reader (Methodological perspectives)*. Chicago: Aldine.
- Wikipedia: Galton's problem. 2011.
- Pearl, Judea. 2009 (2nd edition). *Causality: Models, Reasoning, and Inference*. Cambridge: Cambridge University Press.

Pg2Bib

- Brown, Christian and Anthon Eff, 2010 pdf. The State and the Supernatural: Support for Prosocial Behavior, Structure and Dynamics: *eJournal of Anthropological and Related Sciences*, 4(1) art 1. <http://escholarship.org/uc/item/7cm1f10b>
<http://intersci.ss.uci.edu/wiki/pdf/eScholarshipBrown&Eff.pdf>
- Dow, Malcolm M, and E. Anthon Eff. 2009a. Cultural Trait Transmission and Missing Data as Sources of Bias in Cross-Cultural Survey Research: Explanations of Polygyny Re-examined. *Cross-Cultural Research*. 43(2):134-151.
- Dow, Malcolm M., and E. Anthon Eff. 2009b. Multiple Imputation of Missing Data in Cross-Cultural Samples. *Cross-Cultural Research*. 43(3):206 - 229.
- Eff, E. Anthon. 2008. Weight Matrices for Cultural Proximity: Deriving Weights from a Language Phylogeny. *Structure and Dynamics: eJournal of Anthropological and Related Sciences*. 3(2): Article 9. <http://repositories.cdlib.org/imbs/socdyn/sdeas/vol3/iss2/art9>
- Eff, E. Anthon, and Malcolm M. Dow. 2008. Do Markets Promote Prosocial Behavior? Evidence from the Standard Cross-Cultural Sample. MTSU Working Paper. <http://econpapers.repec.org/paper/mtswpaper/200803.htm>
- Eff, E. Anthon, and Malcolm M. Dow. 2009. How to deal with Missing Data and Galton's Problem in Cross-Cultural Survey Research: A Primer for R. *Structure and Dynamics: eJournal of Anthropological and Related Sciences*. 3(3): Article 1. <http://repositories.cdlib.org/imbs/socdyn/sdeas/vol3/iss3/art1>
<http://intersci.ss.uci.edu/wiki/pdf/eScholarshipEff&Dow2009.pdf>
- Eff, E. Anthon and Giuseppe Rionero. 2011. The Motor of Growth? Parental Investment and per capita GDP. *World Cultures eJournal*, 18(1). <http://escholarship.org/uc/item/5zh0t0q4>
- Little, Roderick, J. A., and Donald B. Rubin. 2010. *Statistical analysis with missing data*. Wiley Series in Probability and Mathematical Statistics.
- Rubin, Donald B. 1987. *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley & Sons.

- Wooldridge, J.M. 2006. *Introductory Econometrics: A Modern Approach*. NY, NY: Thomson South-Western.
- Pg3Bib1
- Kelejian, H., & Prucha, I. 1998. A generalized two-stage least squares procedure for estimating a spatial autoregressive model with autoregressive disturbances. *Journal of Real Estate Finance and Economics*, 17, 99-121.
<http://econweb.umd.edu/~kelejian/Research/P071897.PDF>
- Ord, Keith. 1975. Estimation Methods for Models of Spatial Interaction. *Journal of the American Statistical Association* 70:120-126.
- Pg3Bib2
- Dawid, A. P. 1979. Conditional Independence in Statistical Theory. *Journal of the Royal Statistical Society. Series B (Methodological)* 41:1-31.
- Fox, John. 2006. Structural Equation Modeling with the sem Package in R. *Structural Equation Modeling* 13(3): 465-486.
<http://socserv.socsci.mcmaster.ca/jfox/Misc/sem/SEM-paper.pdf>
- Grace, Jim. 2009. Modeling with Structural Equations/About SEM Software. <http://www.structuralequations.com/software.html>
- Pg4Bib
- Hausman, J. A. 1978. Specification Tests in Econometrics. *Econometrica* 46:1251-1271.
- White, Halbert, and Xun Lu. 2010. Robustness Checks and Robustness Tests in Applied Economics. <http://bit.ly/kK2M8g>
- Hurwicz, L. 1962. On the Structural Form of Interdependent Systems. *In*, Patrick Suppes, and Amos Tverski, eds., *Logic, Methodology and Philosophy of Science*, pp. 232-239. Stanford, CA: Stanford University Press.
- Lukas, Robert. 1978. Economic Policy Evaluation: A Critique. *In*, K. Brunner, A. Meltzer, eds., *The Phillips Curve and Labor Markets*. Carnegie-Rochester Conference Series on Public Policy. New York: Elsevier, pp. 19-46.
- Pg5Bib
- Nakagawa, Shinichi, and Innes C. Cuthill, 2007. Effect size, confidence interval and statistical significance: a practical guide for biologists. *Biological Reviews: Cambridge Philosophical Society* 82 (4): 591-605. doi:10.1111/j.1469-185X.2007.00027.x. PMID 17944619.
- Duncan, Otis. D. 1961. A Sociometric Index for All Occupations. Pp. 109-138 in, A. J. Reiss, Jr., ed.), *Occupations and Social Status*. New York:Free Press
- Fox, John. 2002. *An R and S-Plus Companion to Applied Regression*.
- Pg6Bib
- Christakis, Nicholas A. and James H. Fowler. 2009. *Connected: The Surprising Power of Our Social Networks and How They Shape Our Lives*. New York, NY: Little, Brown and Company.
- Granovetter, Marc S. 1973. The Strength of Weak Ties. *American Journal of Sociology* 78(6): 1360-1380.
- Korotayev, Andrey, and Victor de Munck. 2003. Galton's Asset and Flower's Problem: Cultural Networks and Cultural Units in Cross-Cultural Research. *American Anthropologist* 105 (2): 353-358.
- Witkowski, Stanley. 1974. Galton's Opportunity-Hologeistic Study of Historical Processes. *Behavior Science Research* 9 (1): 11-15.
- Pg7Bib
- Barry, Herbert, III, Lili Josephson, Edith Lauer, and Catherine Marshall. 1977. Agents and Techniques for Child Training. *Ethnology* 16:191-230. Reprinted in Herbert Barry, III, and Alice Schlegel. 1980. *Cross-Cultural Samples and Codes*. Pittsburgh: University of Pittsburgh Press. <http://www.jstor.org/stable/3773387>
- White, Douglas R., Michael Burton, William Divale, Patrick Gray, Andrey Korotayev, Daria Khalturina. 2009. *Standard Cross-Cultural Codes*. <http://eclectic.ss.uci.edu/~drwhite/courses/SCCCodes.htm> (Last updated May 2011. See variable 238.
- Murdock, George Peter. 1967. *Ethnographic Atlas*, Pittsburgh: University of Pittsburgh Press.
- Swanson, Guy E. 1960. *Birth of the Gods*. Ann Arbor, MI: University of Michigan Press.
- Pg8bib
- Holm, S. 1979. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics* 6:65-70.
- Rice, William R. 1989. Analyzing Tables of Statistical Tests. *Evolution* 43(1): 223-225.
- Paige, Karen and Jeffrey Paige. 1981. *The Politics of Reproductive Rituals*. University of California Press.
- Robins, James M., Daniel O. Scharfstein, and Andrea Rotnitzky. 1999. Sensitivity Analysis for Selection Bias and Unmeasured Confounding in Missing Data and Causal Inference Models. *Statistical Models in Epidemiology: The Environment and Clinical Trials*. Halloran, M.E. and Berry, D., eds. NY: Springer-Verlag, pp. 1-94. <http://biosun1.harvard.edu/~robins/sasbcifnl.pdf>
- Pg9bib
- Kyono, Trent Mamoru. 2010. Commentator: A Front-End User-Interface Module for Graphical and Structural Equation Modeling. Technical Report R-364. Cognitive Systems Laboratory. Department of Computer Science. UCLA.
http://ftp.cs.ucla.edu/pub/stat_ser/r364.pdf
- Pearl, Judea. 2011. *The Causal Foundations of Structural Equation Modeling*. Chapter for R. H. Hoyle (Ed.), *Handbook of Structural Equation Modeling*. New York: Guilford Press. Dept. of Statistics Papers <http://www.escholarship.org/uc/item/490131xj>
- Revelle, William. 2007. Using R for psychological research: A simple guide to an elegant package Structural Equation Modeling in R. [http://www.personality-project.org/r/\(r.sem.html\)](http://www.personality-project.org/r/(r.sem.html))
- Wright, Sewall. 1921. Correlation and causation. *Journal of Agricultural Research* 20: 557-585.
- Wright, Sewall. 1923. The theory of path coefficients: A reply to Niles' criticism. *Genetics* 8: 239-255.
- Wright, Sewall. 1934. The method of path coefficients. *Annals of Mathematical Statistics* 5: 161-215.
- Pg10bib
- Gibbs, J. Willard. 1902. *Elementary Principles of Statistical Mechanics*. New Haven, Connecticut: Yale University Press.
- Lauritzen, Steffen. 2011 (January). <http://www.stats.ox.ac.uk/~steffen/teaching/grad/graphicalmodels.pdf> Graphical Models. PPT in PDF. Graduate Lectures, Oxford.
- Wold, Herman O. A. 1954. Causality and Econometrics. *Econometrica* 22: 162-177.

Pg11bib

Vayda, Andrew P. and Bradley B. Walters. 2011. Causal Explanation for Social Scientists. Lanham, MD: Rowman and Little.

Appendix 1: Practical advice for runs of the program

To run the program the first time, use one of existing examples that come with the package. Run the three `#setup` lines and source code `#1` to check if the `setwd`, package and data are working along with code `#2`. Once this is verified you can run the source codes `#2-4`.

Remote sourcing can be done, e.g., for replication of the Brown and Eff (2010) model (at <http://bit.ly/jNZjSa>), provided that R, the library, and the data are installed:

```
setwd('/Users/drwhite/Documents/3sls/scs/') #setup (output will be stored here)
library(scs) #setup
data(scs) #setup
source("http://intersci.ss.uci.edu/2sls/ccs/create_EduR_1.5DistBrownEff.R") #2nd source program (model)
source("http://intersci.ss.uci.edu/2sls/ccs/two_stage_ols.R") #3rd source program (2sls)
source("http://intersci.ss.uci.edu/2sls/ccs/run_model_79.R ") #4th source program (run)
```

For the Eff and Dow (2009) model,

```
source("http://intersci.ss.uci.edu/2sls/ccs/create_model_value_childrenLangOnly.R") #2nd source program (model)
source("http://intersci.ss.uci.edu/2sls/ccs/two_stage_ols.R") #3rd source program (2sls)
source("http://intersci.ss.uci.edu/2sls/ccs/run_model_79.R") #4th source program (run)
```

After a successful run, the ratio parameter in source code `#4` can be changed (e.g., from 70 to 79) provided that the parameter within the program is changed and not just the name. If the `#2nd` (model) program is changed, that line should be run by itself to check for errors, and only then should the `#3rd` and `#4th` programs be run.

The `run_model_...R` program has multiple imputation parameters set to:

```
nimp=8,
niterations=8,
```

These can be raised for greater convergence in multiple imputation of missing data, say to 10, 10, but then the time for runs may increase dramatically. Lower settings are not recommended.