Inferential Statistics for finding networks of causality: Designing an Open eRepository for Anthropological Knowledge (ORAK)

Open access to data, software for inferential statistics, and findings of these new methods expand the benefits of broad interdisciplinary collaborations analyzing and expanding on world ethnographic data coded for the Standard Cross-Cultural Sample (SCCS: 1969) and other databases.

This talk explains how these benefits can contribute to a proposed Open eRepository for Anthropological Knowledge that can serve researchers and classes worldwide:
An Open eRepository for Anthropological Knowledge: What is it?

Anthropology is the source of ethnographic studies all over the world, and these ethnographies are compared and coded for distinctive features and variables in what are called cross-cultural studies. One database (SCCS) deals with the best early descriptions of societies in 186 distinctive cultural areas, others with, e.g.: samples of foragers, archaeological and historical studies of chiefdoms and early states, others with contemporary societies. Typical problems: how to analyze samples of cases that are interrelated, and whether proper causal inferences can be drawn.
1. Benefits: to analysis
   A. Solving Galton’s problem with 2-stage least squares
   B. Causality   C. MI   D. GIS
2. Benefits: to data
3. Benefits: findings and models
4. Benefits: networks of findings
5. Examples
Project team

Duran Bell, Economics, UCI
Karim Chalak, Economics, Boston College
Scott White, ABD, I&CS, UCI, and Senior Systems Architect
Tony Soeller, Network & Academic Computing Services, UCI
Advice
Halbert White, Economics, UCSD
Anthon Eff, Economics, Tennessee State University
Judea Pearl, Computer Science, UCLA (advice and inspiration)
Participants today
Skye McDonald – from the Anthro 174AW class
Reng Feng –Anthro 174AW tutor, visiting from Xi’an Xiaotung U
1. Project benefits: to analysis

• **A. Solving Galton’s problem** of nonindependence using Instrumental Variables for language and spatial clustering (Eff and Dow 2009 2-stage least squares) to significantly improve the quality of regression analyses that set the stage for Bayesian causal analyses.

• **B. Detecting causality** from observation and categorical data (computer science and econometric use of counterfactuals and Bayesian statistics: Pearl 1987, 1988, 2000, 2009a, 2009b; Chalak and Hal White 2009).

• **C. Imputing missing data** (Eff and Dow 2009, and derived from National survey research methods: Rubin 1987).

• **D. Analyzing spatiotemporal processes and patterns** that allow open access interactive geodatabase GIS visualizations.
1.A. Solving Galton’s problem with 2-stage least squares

- I call this the “network problem”: societies are related by historical network processes by which they influence one another.
- Solving Galton’s problem of nonindependence using Instrumental Variables for language and spatial clustering (Eff and Dow 2009 2-stage least squares) to significantly improve the quality of regression analysis estimates that set the stage for Bayesian causal analyses.
Instrumental Variables

- Regression models are not well estimated when their error terms are not independent but contain the residuals of unspecified variables or systematic network interactions.

- Instrumental Variables (IVs) are observed variables in structural models (SEMs), other than the cause or treatment of interest, that can play an instrumental role in identifying and estimating causal effects.

- They require two stages of OLS analysis (2SLS): first to identify the instrumental variables, including network variables; second to solve for the cause or treatment of interest plus the instrumental variables.
2SLS

- Nonindependence of cases is common in survey research due to network influences. In cross-cultural research the branching of common linguistic families is one source of “background similarity” and the other is spatially proximate interaction reflected in “spatially clustered similarity”
- The instrumental variables (I-Vs) in 2SLS tend to reduce significance when these are operative, representing a downgrading of “effective sample size” due to the reduction of sources of variance: “Galton’s problem.” We begin with network variables for the I-Vs.
Network W matrices as I-Vs

• Let $W_L$ be a row-normalized matrix (summing to 1 in each row) for a tree of linguistic families. When $W_L$ is multiplied by a variable $y$, $W_L y$ is a new variable with the same range of values as $y$. If $y$ and $W_L y$ are significantly correlated, then $y$ is influenced by the “background similarity” of common linguistic origin. Similarly, correlation between $y$ and a $W_D y$ product for geographic closeness reflects the degree of distance or “spatially clustered similarity” in $y$.

• $W_L y$ and $W_D y$, then, and other network variables like them, become the instrumental variables (I-Vs) for reduction of sources of variance that stem from “Galton’s problem.” We begin with these network variables for the I-Vs of cross-cultural studies, such as the linguistic similarity matrix.
Here’s the idea of I-Vs

The network effects regression model (Dow 2007; Eff and Dow 2009) is:

\[ y = \rho_1 W_{1y} + \rho_2 W_{2y} + \ldots + \rho_t W_{ty} + X\beta + \epsilon \]

Since \( y \) is a function of the error term \( \epsilon \) (Eff and Dow 2009:15) each of the \( W_iy \) variables is also a function of \( \epsilon \), is thus endogenous and must be replaced by an instrumental variable. In the program we create two instrumental variables for network-lagged dependent variables: one using the language matrix (fyll); the other using the distance matrix (fydd).

Since \( W \) is \( nxn \), and \( y \) is \( nx1 \), \( Wy \) is \( nx1 \). One thus regresses \( Wy \) on a \( nxj \) matrix of exogenous variables \( Z \):

\[ (Wy)_{i} = \alpha_0 + \sum_j \alpha_j z_{ji} + \mu_i \] (1)

which gives estimated coefficients \( \hat{\alpha}_j \) that can be used …
… to create a fitted value for $W_y$:

$$ (\hat{W}_y)_i = \hat{\alpha}_0 + \sum_j \hat{\alpha}_j z_{ji} $$

It is this fitted value that is our instrumental variable. Now the instrumental variable from this stage-1 regression, (one such term for each $W_i$),

is simply added into the original effects model

$$ y = \beta_1 W_{1y} + \beta_2 W_{2y} + \ldots + \beta_t W_{ty} + X\beta + \varepsilon \quad (2) $$

Since the error terms in (1) and (2) are uncorrelated with other terms, then OLS regression in 2 stages gives our result.
Visualizing GIS findings from 2SLS

2SLS can tell us how much influence on the y dependent variable for each society comes from proximal societies or language mates. The log(pi/si) for pi=proximity prediction/si=society i’s original variable value tells us the magnitudes of proximal influence effects.

This map for language is a notional image of places where the language similarity effect is greatest for Bantu and Malayo-Polynesian societies. The black ring around a high effect area indicates a dissimilarity effect.

Phylogenetic reconstruction algorithms for each language can show the magnitude of ANCESTRAL effects for each language.

1/15/2010
oddities

Language autocorrelation effects are sometimes *negative*, indicating that societies tend to *differentiate* from their language mates as to behaviors like *warfare*, which is more likely with proximal societies that are *not* language mates. Here the distance effect is positive.
1.B. Detecting Causality with Causal graphs, e.g. Pearl 1987

Figure 10.1. An inheritance network depicting ambiguity. Heavy arcs represent factual information about individuals, thin arcs represent default statements among properties, slashed arcs represent denials.
Causality and causal graphs

• Here are some core ideas of Pearl’s partially Bayesian causal graphs.

• Causality in an indep $X \rightarrow_c$ dep $Y$ variable pair with a coefficient c means that a unit change in X will bring about c units of change in E(Y), the expected value of Y. "Change" means that if the physical means exists of fixing Y at some constant Y1, and of changing that constant from Y1 to Y2, then the observed change in the expected value E(Z) will be c(Y2-Y1).
Counterfactuals, d-seperability

- The relation of an independent to the dependent variable cannot involve a counterfactual such as “mud predicts rain” that cannot be causal.

- $X \rightarrow W \rightarrow Y$ causal chains are d-separated if there is no confounding $X \rightarrow Y$ causality. If $X$ and $Y$ are d-separated and $X$ is not an ancestor of $Y$ (written $X^\circ \rightarrow Y$), then for a no-counterfactuals regression relation indep $X \rightarrow_b$ dep $Y$ the coefficient $b$ is an unbiased estimate of coefficient $c$ in a causal relation $X \rightarrow_c Y$. 
1.C. Multiple imputed (MI) missing data

- National survey research methods developed by Rubin (1987), summarized by Dow and Eff (2009), provide the Rubin’s formulas for combining estimates from multiple copies of the same data with independently replaced missing values predicted by Aux Vars.
- (Eff and Dow 2009) incorporate MI into 2SLS.
- This solves two big problems in learning about social logics from cross-cultural studies.
Inferential Stats: the beauty of Don Rubin’s formulas (1987)

• e.g. You have a statistical result and you convert it to probabilities, e.g. missing data value probabilities estimated from a fixed and completely coded auxiliary databases

• i.e., you have missing data, and general demographic or environmental variables to predict missing values, then stated as probabilities from those predictors to missing values (i.e., missing data imputation)

• So you have j=2-10 simulations of these randomly generated distributions, and j=2-10 statistics $Q^j$ means and $U^j$ variances for each
MI for statistic Q

Then Rubin’s formulas combine variances in multiple imputation (MI) tend to reduce significance.

Imputed Data for a variable

\[ j = 1 \quad 2 \quad 3 \quad \ldots \quad m \]

\[ \begin{array}{cccc}
1 &  &  & \\
2 &  &  & \\
\cdot &  &  & \\
\cdot &  &  & \\
186 &  &  & \\
\end{array} \]

1 cases

\[ \bar{Q} = \frac{\sum_{j=1}^{m} \hat{Q}^{(j)}}{m} \]

\[ W = \frac{\sum_{j=1}^{m} U^{(j)}}{m} \]

\[ Q_j = \_ \_ \_ \_ \_ \_ \_ \text{ means in } \]

\[ U_j = \_ \_ \_ \_ \_ \_ \_ \text{ variances in } W \text{ for columns 1 to } m \]
Mean
\[ Q = \sum_{j=1}^{m} \hat{Q}^{(j)} / m \]  
(1)

Mean of within variances
\[ W = \sum_{j=1}^{m} U^{(j)} / m \]  
(2)

Between variance of Means
\[ B = \sum_{j=1}^{m} \left( \hat{Q}^{(j)} - \bar{Q} \right)^2 / (m - 1) \]  
(3)

Combined W&B variances
\[ T = W + \frac{(m+1)}{m} B \]  
(4)

So while the mean of Q is given in eqn. (1), the variances come from two sources: (2) within the rows and (3) between the columns, and (4) this variance is larger than either W or B. (Rubin 1987)
Effects on significance

• With better inferential statistics, MI and 2SLS both tend to *thin out the multitude of spurious results that characterized earlier cross-cultural studies* by inflated significance tests.

• The network of indep $X \rightarrow_b$ dep $Y$ relations from 2SLS and potentially causal graphs is potentially very sparse compared to matrices of significant correlations by descriptive stats.

• In our pretests it is also much flatter.
Network of 2SLS findings from pretests

Colored nodes are indep vars sharing depvars that are circled for different types
1.D. Analyzing spatiotemporal processes with GIS mapping

Boats or domesticated animals for travel ->

Low warfare and fighting (Peggy Sanday)

Trade and warfare

High warfare and fighting (Sanday)

large canoes in native America used in trade

Large pack animals in the Old World involved in major E-W trade

Outliers of trade
Indigenous money: currency or coin (Murdock)  

Showing pack animal trade routes

Semi-marketized farming (orange dots : Pryor) also on those trade routes
2. **ORAK** Benefits: data

The eRepository proposed to NSF along with these methods is open access but peer reviewed: files can be uploaded and downloaded, and used with these analytical tools, but peers in the field of study do contribute to cleaning and documenting data, as with the 186 society database for the 2,000+ variables and codebook for the Standard Cross-Cultural sample.
3. Benefits: findings and models

With online open access software eRespository proposed to NSF for ORAK and the use of interactive software and codebooks to analyze data like the 186 society 2,000+ variables and coding subcategories of the Standard Cross-Cultural sample, results of analysis can be saved on-line, viewed, and replicated (or the analysis modified) by other researchers. **This can be done in classrooms with internet access outside of computer labs, hence world classroom access.**
3. Benefits: findings and models

The existing eRespository on its own wikimedia server (InterSci wiki) already allows access to interactive software and codebooks to analyze Standard Cross-Cultural sample data, and with 970 published searchable by Google Scholar, this can be done in classrooms with internet access outside of computer labs, hence worldwide classroom access.
4. Benefits: networks of findings

The eRespository software will inventory the 3-way network of researcher-independent-dependant variable links and the regression coefficients on the arc in network graphs like that on slide 20. The viewer can then pick a connected subnetwork and run statistical programs to view the resultant causal graph analyses. The software can then run higher order analysis through the Settable Systems causal modeling programs of Chalak and H. White (2009), to be implemented under the NSF proposal, if funded. **This too can be done in classrooms with internet access outside of computer labs, hence worldwide classroom access.**
Examples

Now, to see examples of student projects from the experimental class from fall, 2009, we might present a powerpoint by Skye McDonald representative of a larger set of 23 powerpoints

depvar v678 warfare and fighting

EduMod-6: Imputation and Regression depvar 679 warfight
Kat

2SLS model for warfare and fighting = SCCS$v678

<table>
<thead>
<tr>
<th></th>
<th>coef</th>
<th>Fstat</th>
<th>ddf</th>
<th>pvalue</th>
<th>VIF</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>1.072</td>
<td>54.266</td>
<td>3038.376</td>
<td>0.000</td>
<td>NA</td>
</tr>
<tr>
<td>pigs</td>
<td>0.157</td>
<td>2.028</td>
<td>21877.912</td>
<td>0.154</td>
<td>1.338 v244==2</td>
</tr>
<tr>
<td>milk</td>
<td>0.099</td>
<td>1.060</td>
<td>6924.022</td>
<td>0.303</td>
<td>1.325 v245&gt;1</td>
</tr>
<tr>
<td>popdens</td>
<td>0.010</td>
<td>0.148</td>
<td>30242.760</td>
<td>0.700</td>
<td>1.293 v156</td>
</tr>
<tr>
<td>agrlateboy</td>
<td>0.054</td>
<td>6.140</td>
<td>613.970</td>
<td>0.013</td>
<td>1.160 v300</td>
</tr>
<tr>
<td>war</td>
<td>0.021</td>
<td>11.647</td>
<td>140.761</td>
<td>0.001</td>
<td>1.122 v1648</td>
</tr>
</tbody>
</table>

r2

R2: final model

<table>
<thead>
<tr>
<th>IV(distance)</th>
<th>IV(language)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.2086215</td>
<td>0.9295857</td>
</tr>
<tr>
<td>0.9397868</td>
<td></td>
</tr>
</tbody>
</table>
Amanda: depvar 667 rape

EduMod-11: Imputation and Regression depvar 667 rape
Amanda

2SLS model for rape = SCCS$v668

<table>
<thead>
<tr>
<th>coef</th>
<th>Fstat</th>
<th>ddf</th>
<th>pvalue</th>
<th>VIF</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>1.204</td>
<td>3.717</td>
<td>3743.789</td>
<td>0.054</td>
</tr>
<tr>
<td>fydd</td>
<td>-0.514</td>
<td>1.383</td>
<td>1078.445</td>
<td>0.240</td>
</tr>
<tr>
<td>freintovio</td>
<td>0.325</td>
<td>7.306</td>
<td>212.412</td>
<td>0.007</td>
</tr>
<tr>
<td>ideomaltough</td>
<td>0.317</td>
<td>5.144</td>
<td>241.921</td>
<td>0.024</td>
</tr>
</tbody>
</table>

> r2

**R2:final model**

<table>
<thead>
<tr>
<th>IV(distance)</th>
<th>IV(language)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.9461911</td>
<td>0.9609407</td>
</tr>
<tr>
<td>0.2258225</td>
<td></td>
</tr>
</tbody>
</table>
Villy: depvar Socially Organized Homicide

EduMod-12: Imputation and Regression depvar 1675 Socially Organized Homicide

Villy

2SLS model for Socially Organized Homicide = SCCS$v1675

<table>
<thead>
<tr>
<th>coef</th>
<th>Fstat</th>
<th>ddf</th>
<th>pvalue</th>
<th>VIF</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>-6.191</td>
<td>0.553</td>
<td>19191.517</td>
<td>0.457</td>
</tr>
<tr>
<td>fyll</td>
<td>-1.573</td>
<td>2.972</td>
<td>8780.817</td>
<td>0.085</td>
</tr>
<tr>
<td>money</td>
<td>0.288</td>
<td>1.750</td>
<td>134319.883</td>
<td>0.186</td>
</tr>
<tr>
<td>dateobs</td>
<td>0.007</td>
<td>3.496</td>
<td>249175.045</td>
<td>0.062</td>
</tr>
<tr>
<td>war</td>
<td>0.288</td>
<td>32.758</td>
<td>14518.659</td>
<td>0.000</td>
</tr>
</tbody>
</table>

> r2

R2:final model R2:IV(distance) R2:IV(language)

0.4279652  0.9724051  0.9459899
Nathan: depvar 122 games of strategy

EduMod-17: Imputation and Regression depvar 122 games of strategy

Nathan

2SLS model for games of strategy = SCCS$v122

<table>
<thead>
<tr>
<th>coef</th>
<th>Fstat</th>
<th>ddf</th>
<th>pvalue</th>
<th>VIF</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>-0.1963683</td>
<td>10.697370</td>
<td>3149943.9</td>
<td>0.01107289</td>
</tr>
<tr>
<td>fyll</td>
<td>-0.3573787</td>
<td>1.925958</td>
<td>291090.7</td>
<td>0.1652024</td>
</tr>
<tr>
<td>fydd</td>
<td>0.9878472</td>
<td>32.366629</td>
<td>281700.2</td>
<td>0.000000013</td>
</tr>
<tr>
<td>stratif</td>
<td>0.1206975</td>
<td>36.494753</td>
<td>2090781.4</td>
<td>0.000000002</td>
</tr>
</tbody>
</table>

> r2

R2: final model      IV(distance)      IV(language)
0.4863288             0.9871563             0.9945682
Hiu Kwan: depvar 754 wifebeating

EduMod-18: Imputation and Regression  depvar 754 wifebeating
Hiu Kwan

2SLS model for wifebeating = SCCS$v754

<table>
<thead>
<tr>
<th></th>
<th>coef</th>
<th>Fstat</th>
<th>ddf</th>
<th>pvalue</th>
<th>VIF</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>7.19267</td>
<td>23.44421</td>
<td>45407.72</td>
<td>0.00000</td>
<td></td>
</tr>
<tr>
<td>fyll</td>
<td>-3.62064</td>
<td>17.63661</td>
<td>46486.63</td>
<td>0.00003</td>
<td></td>
</tr>
<tr>
<td>language clus</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>fydd</td>
<td>0.82320</td>
<td>11.81269</td>
<td>155871.13</td>
<td>0.00059</td>
<td></td>
</tr>
<tr>
<td>spatial clus</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>roots</td>
<td>-0.31280</td>
<td>6.76358</td>
<td>818899.57</td>
<td>0.00930</td>
<td></td>
</tr>
<tr>
<td>cereals</td>
<td>-0.33612</td>
<td>10.67803</td>
<td>1164233.14</td>
<td>0.00108</td>
<td></td>
</tr>
<tr>
<td>techspe</td>
<td>0.28669</td>
<td>3.08090</td>
<td>1495695.42</td>
<td>0.07922</td>
<td></td>
</tr>
<tr>
<td>v233==5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>hunt</td>
<td>-0.09536</td>
<td>10.98199</td>
<td>461251.41</td>
<td>0.00092</td>
<td></td>
</tr>
<tr>
<td>v233==6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&gt; r2</td>
<td>0.5259626</td>
<td>0.9819218</td>
<td>0.9482473</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
EduMod-22: Imputation and Regression depvar 17 money
Peyman

2SLS model for money = SCCS$v17

<table>
<thead>
<tr>
<th>coef</th>
<th>Fstat</th>
<th>ddf</th>
<th>pvalue</th>
<th>VIF</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>1.258</td>
<td>5.271</td>
<td>14255257</td>
<td>0.022</td>
</tr>
<tr>
<td>fyll</td>
<td>-0.722</td>
<td>5.412</td>
<td>33859578</td>
<td>0.020</td>
</tr>
<tr>
<td>fydd</td>
<td>0.727</td>
<td>10.400</td>
<td>73121870</td>
<td>0.001</td>
</tr>
<tr>
<td>foodtrade</td>
<td>0.018</td>
<td>3.878</td>
<td>229635082</td>
<td>0.049</td>
</tr>
<tr>
<td>ecorich</td>
<td>-0.176</td>
<td>5.025</td>
<td>317122257</td>
<td>0.025</td>
</tr>
</tbody>
</table>

> r2

R2: final model

<table>
<thead>
<tr>
<th>IV(distance)</th>
<th>IV(language)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.4527387</td>
<td>0.9509774</td>
</tr>
<tr>
<td>0.9879003</td>
<td></td>
</tr>
</tbody>
</table>
oddities

• Note: I coded *negative* fyll in *blue* but 8 out of 25 were positive. Distance usually dominates language as an autocorrelated predictor, often leaving language a negative residual for items that are linguistically rare controlling for distance, for example, attacking a neighbor or making external war.
end