Methods for the comparative quantitative analysis of the intergenerational transmission of wealth in pre-modern societies

Adrian Bell, Suresh Naidu, Monique Borgerhoff Mulder, and Samuel Bowles[1]

20 November, 2007

This memo provides guidelines to Phase I project participants on how to prepare and analyze data sets describing the transmission of wealth across one generation. It is part of a Santa Fe Institute project on the transmission of wealth and the dynamics of inequality in small-scale societies coordinated by Monique Borgerhoff Mulder and Samuel Bowles (http://www.santafe.edu/events/workshops/index.php/Inter-generational-_Transmission_of_wealth_in_Premodern_Societies).

Because the goal is comparability across different societies and different kinds of wealth we need to adopt common methods for all of the data sets, even where there might be modifications of our approach that might be more adequate to the idiosyncrasies of some cases. Another implication is that these guidelines should be followed as strictly as possible.

The first step is to verify the data are sufficient and prepared for analysis. This is described in section 1. Section 2 describes the process of estimating the model's main parameter of interest, the intergenerational transmission coefficient β including a discussion of pertinent corrections to the estimation procedure.[2] Section 3 extends the discussion of necessary corrections to the parameters estimated in section 2. Section 4 describes what to do so that our statistical team may double-check the estimates. If you are looking for a good reference work that addresses the problems raised here, we recommend Johnstone and DiNardo (1997). If you have problems with the estimation contact Adrian Bell at avbell@ucdavis.edu. This memo will be updated as we gain experience and confront as yet unforeseen problems. The current version will be available on the team webpage.

We welcome comments and suggested improvements on this memo (or on other aspects of the project, of course) at any time.

---

[1] Tom Hertz contributed indispensably to this document.

[2] We will discuss estimation and correction routines in the statistical programs STATA and R.

1 Data preparation

1.1 Data content

We define "wealth" as any form of value or capital, either material, somatic, social (e.g. network-based) or knowledge-based. Minimally these forms of capital can be stored in objects, in bodies, in networks of people, or in institutional agreements, and no particular form of wealth need be stored exclusively in any single form (e.g., land is only "wealth" if one has secure access to its produce (not necessarily private property in the classic sense, but through some institution linking the land with the income of the person)). Contributors to the project are asked to identify forms of wealth that, from their ethnographic sense, are important to success in that particular culture, population or community, for example indicators of "cultural success" as defined by Irons (1979). Note that forms of "wealth" that are indicative of success in one population, e.g. income in Tsimane, may be indicative of relative failure in others (e.g. wage income among women in Bangalore). Accordingly in the latter case income would not be considered as "wealth".

1.2 Variables and dataset structure

Wealth is measured by its natural logarithm. If there are zeros in your data set please follow the correction procedures described in the Appendix (A.1) before estimation. We measure wealth in natural logs because this makes our estimate of β a "unit free elasticity". Irrespectively of whether we are dealing with cows, florins or kilos, β gives the percentage difference in the offspring's wealth associated with a percent difference in the parents' wealth. Remember that in estimating β we are not attempting to identify the effect of any particular form of inheritance (bequest, genetic inheritance, etc.) nor any other specific causal mechanism. We are interested in a more coarse-grained concept: the statistical association between the wealth of two generations, however this may come about. (We anticipate a Phase 2 project that will explore similarities and differences among causal mechanisms for intergenerational transmission.) A useful relationship between the simple correlations and the elasticity to keep in mind is that (ignoring age corrections) where wealth is measured in natural logs, and ρ is the correlation of parental and offspring wealth, and σ and σ_p respectively are the standard deviations of offspring and parental wealth, then $\beta = \rho\sigma/\sigma_p$. While the correlation obviously cannot exceed one, β can, but only if inequality of wealth (measured by its standard deviation) is greater in the offspring generation than among parents.

Data sets should be structured on parent-offspring wealth pairings. For comparability, we are specifying the following pairings: (1) for father and son (denoted f-s), (2) for father and daughter (denoted f-d), (3) for mother and son (m-s), and (4) for mother and daughter (m-d). From here on, when we refer to "parent" and "child" we refer to any or each of the pairings above. In some cases you may additionally want to use an averaged measure of parental wealth (par-s, or par-d); refer to Appendix (A.2) for guidelines. Of course not all of these pairings will be substantively relevant and practically possible, but

we urge everyone to try to produce as many as is possible and sensible. i.e (f-s, f-d, m-s, m-d).

1.3 Reporting the results

When you have finished analyzing your data, please fill out the worksheets in the file *IntergenSummary.xls* that is distributed with this memo or may be downloaded from the webpage. You will note that we focus on "wealth-type pairings". This is because several contributors are providing β estimates for different wealth types, and sometimes these are based on different pairings and different samples. The first worksheet provides the summary β estimates, the second the descriptive statistics, the third a (series of) worksheet(s) with lowess-smoothed scatterplots for the different wealth-types, and the fourth (optional, see below) the data for each particular wealth-type pairing; if you have multiple wealth types, you will have multiple sheets with scatterplots and data. On the webpage you will also find a sample file that reports the results for the Kipsigis (*Kips-IntergenSummary.xls*).

Should you wish to have your analyses checked, or to have further regression models run to deal with sample bias or skewed distributions, please submit your data to us (prior to 15 September). Prepare data files into appropriately labeled worksheets (indicating the wealth-type pairing), use the standard variable names, follow the format indicated in the worksheet WEALTH1_F-S, and provide no extraneous measures. We hope that this approach will minimize confusion. It might help to follow the model in the worksheet called LAND_F-S which you will find in *Kipsigis-IntergenSummary.xls*.

We also suggest that if you have known "missing" individuals these should be included in the data set so that a Heckman correction can eventually be run. For example, in a case where known fathers have missing (or out-migrated sons) a line would include measures of father's wealth and blank cells designating missing data for son's wealth, and a variable indicating missing status (see lines 28 and 29 in LAND_F-S in *Kips-IntergenSummary.xls*).

Finally we encourage you to submit a brief memo with your *IntergenSummary.xls* file to give some background to your study, and a brief ethnographic description of the population, and the output associated with estimating each reported β. Things to consider are potential biases in the sample, the estimation of measurement error, why you chose to calculate the wealth measures that you did, how and why you chose your measure of parental wealth (i.e., father, mother or parent) and how you dealt with parental averaging, etc. You might also include some thoughts on how you think the different wealth indices are transmitted in your population (although this is only for a later phase in the project). A draft of such a memo can be found on the webpage (*Kipsigis memo.doc*).

2 Model and estimation procedure

The natural logarithm of the wealth of the parent and offspring we denote y_p and

y, respectively. We are interested in the statistical model:

$$y_{ik} = \beta_0 + \beta_1 y_{pi,k} + X\beta_3 + \varepsilon_{ik} \quad (1)$$

where p denotes parent(s), and k is a clustering term indicating the child's membership in a group and the subscript {i,k} refers to the ith individual in the kth cluster. . The group may be a common mother, family, village or community. If there is information on an offspring's group at multiple social levels, then use the group at the highest social level recorded (e.g. use village rather than family if both types of information are available and there is more than one village represented in the sample). X are covariates. Using ordinary least squares, estimate the parameters β_1, β_2, and the vector of parameters β_3 (if one or more covariate is prescribed). Since age of offspring and parent is likely associated with wealth (maybe even in a nonlinear fashion), the list of covariates are:

$$\text{covariates} = X = \{A,\ A^2,\ A_p,\ A_p^2,\ (A - \bar{A})\, y_p\} \quad (2)$$

The variable A is the age of the offspring and A_p is the age of the parent (ages are not logged). $\bar{A}$ is the sample mean age of the offspring. This is the preferred model, and we recommend its use for all wealth analyses. The rationale for $(A - \bar{A})y_p$ is that β should be estimated at the mean age of the offspring. This is done by controlling for the interaction between parent's wealth and offspring's deviation from the mean offspring's age. That is to say, when this interaction is included, the estimate from equation (1) of the derivative of offspring wealth with respect to parental wealth evaluated at the average offspring age (i.e. the thing we want to know) is just β_1. We are not interested in the coefficients on any of the age (alone) variables; their function is simply as controls.

For wealth measures based on fertility and reproductive success we recommend a different procedure. For those of you (e.g. the historians) dealing entirely with individuals who have completed their reproduction we propose that you simply regress offspring wealth on parent wealth, with no age controls and no interactions. For those of you dealing with mixed samples (for example, ethnographers who have a mix of post reproductive and currently reproductive individuals), we suggest a slightly different procedure from above that takes account of any remaining years of possible reproduction, i.e. the offspring may not have completed the reproductive period. The effect is to correct the β estimate for the age of typical last reproduction (A_{lr}). Use the covariates:

$$X = \{A,\ A^2,\ A_p,\ A_p^2,\ (A_{lr} - A),\ (A_{lr} - A)\, y_p\} \quad (3)$$

If $A > A_{lr}$ enter a zero. Please note the value of A_{lr} you use in the summary statistics file (*IntergenSummary.xls*). The rationale for including $(A_{lr} - A)$ as a separate term in the

model is that it when interaction terms are used the main effects must also be included in the model otherwise very distorted beta estimates for parental wealth emerge.[3]

For those of you dealing with mixed samples you may be tempted to examine betas in the reproductively complete subsamples for which age controls are unnecessary (as outlined above). If there has been no major secular change we would expect beta estimates for the full sample and the reproductively complete sample to be similar; if you do this check, you should report both estimates.

Recall that all standard errors should be clustered with the k variable. Unless you have distinct villages or geographic clusters, then the k variable should just be the family cluster. In STATA this is just the cluster option to your REG command. In R this should come from the robcov library. In the Appendix (A.3) we show STATA code for running Eqn 1 and the preliminaries.

2.1 Nonlinearities

Nonlinearities may be found in your data set. For example, the rich may pass their inheritance to their children better (or less reliably) than the middle class or poor. To deal with this simply, we ask that everyone generate a scatterplot of their data and check whether or not there is eyeball detectable nonlinearity in either the logged and the raw values. Please include these scatter plots with the data analysis you send.
We suggest that people with large datasets also construct nonparametric plots (kernreg or lowess in STATA, loess in R) of y against y_p. If your data are highly clustered, use the "jitter" option to show the plotted values better. But remember, our main objective in Phase 1 is to produce a single comparable estimate of β, so even if you find nonlinearities please report the single β measure, specifically the slope at the mean of parent's wealth, as specified in Eqn 1.

2.2 Migration and selection bias

A potentially serious problem is the selective representation of offspring in the second generation; this may reflect migration, or any other dynamic rendering members of the second generation missing from our data sets. This is a problem only if those who are missing differ systematically (meaning, in ways correlated with parental wealth) from those present. For example we may be observing only the outcomes for locally successful offspring (the unsuccessful having migrated to an urban area); but if all of the offspring of the wealthy are successful and few of the offspring of the poor are

[3] The procedure here is:
a) Calculate ($A_{lr} - A$) in a way that is appropriate for your data (for Pimbwe we generated a variable 45-A for women and 55-A for men).
b) Recode ($A_{lr} - A$) to 0 for anyone who has reached age of last reproduction.
c) De-mean the term ($A_{lr} - A$), namely, subtract its sample mean from each observation.
d) Compute the interaction between the de-meaned ($A_{lr} - A$) and Yp.
f) Run the original model adding the new variable. See stata code.

successful, the very poor offspring of poor parents will be missing from our sample and as a result we will underestimate the degree of intergenerational transmission if we do not take account of the selection problem. If we know how many sons there are and how many migrated, then we could adjust for this selection using the so called Heckman procedure. This is one of those cases in which the objective of comparability across data sets has led us (after some thought and consultation) to the conclusion that we should not do these corrections (although we will want to explore the effect of doing them where we can at a later date). The basic idea of the Heckman correction is explained in an appendix (A.4). If you think you have a selection problem of this or related kind, please describe what it is and all that you can determine about its empirical importance (an example of such an illuminating report on sample bias is in Greg Clark's paper).

2.3 Inequality estimates

The coefficient of variation (standard deviation divided by the mean) and the Gini coefficients of the unlogged wealth of offspring and parent should be calculated, as well as the variance of the natural log of your wealth variables. These are all standard measures of inequality. There is a Gini command in both Stata and R (found in the reldist library for R, for STATA this requires downloading by typing "SSC install ginidesc").

3 Measurement Error

Two sources of measurement error are observation error and temporal variation; the former results from noisy measures, whereas the latter results from measures that vary over time, for example a man's cattle holding fluctuate due to disease, rainfall, raiding, etc. Ideally we should correct for both since we are interested in "permanent" or lifetime wealth.

Suppose that observed parental wealth at a given period in time y_p is

$$y_p = y_{p*} + e_p \tag{4}$$

where y_{p*} is the true permanent wealth and e_p is either measurement error or the difference between the transient (observed) and permanent wealth. We would like to be able to estimate $y = a + \beta y_{p*} + \ldots$ but cannot. When $y_p \neq y_{p*}$ the estimated relationship $\hat{\beta}$ will fall short of the true relationship β. If we know the statistical association between the observed and the true measure of wealth we can correct for this bias. Our estimated $\hat{\beta}$ will be

$$\hat{\beta} = \beta \frac{Var(y_{p*})}{Var(y_{p*}) + Var(e_p)}. \tag{5}$$

The ratio of the true variance (the numerator) to the total variance (the denominator) is the square of the correlation coefficient (r) between the true measure and the observed measure. Our strategy is to get an estimate of the correlation between the observed and the true wealth, and then adjust the estimate $\hat{\beta} / r^2 = \beta$. For example if r =0.7, then the true β is more than twice the estimated β.

To address the problem of noise from temporal variation, if we know the year to year variation in wealth, then we estimate r from these data (see below). If there is, instead, observation error, then we can use alternative sources of data to estimate Var(e_p). If you can't find an alternate measure, try to guess a correlation, (0 > r > 1), between the true measure and the observed variable, and correct the estimated β accordingly.

The error variance can be estimated in various ways. First, suppose that for or a given individual we have the number of cows for n years, that is $c_1, c_2, .. c_n$. Define the mean of that series $\bar{c}$ as the "permanent" wealth measure we wish to capture. Represent each c_i for i = 1, 2, …n as a noisy measure of $\bar{c}$. For the sample as a whole the correlation of each c_i with $\bar{c}$ is the correlation of an observed and a true measure. The mean of these observed-true correlations (for all individuals) is the measure (or r) we should use.

Suppose we have no measure of the 'true' (no analogue to $\bar{c}$). But as in Dominica we have two measures (from two different respondents) of the land of a given individual t_1 and t_2. For all the individuals in the sample correlate the t's given by source 1 and source 2. This correlation between the two noisy measures is the square of the correlation between the observed and the true measure, r. The same procedure could be used for the cows example above. The correlation between cows in one year and cows in another year is similar to the correlation between t_1 and t_2 in the Dominica example.

In both cases transient wealth and observation noise, if you have more than two measures of the same thing you should take the mean correlation of all the pairwise measures and then divide the estimated β coefficient by this mean. Be sure to report the estimated correlation between the true and observed when you report your data and estimates.

4 Recalculation of coefficients

In an effort to ensure maximum comparability across data sets, we would like to recalculate the beta coefficients using code specifically designed for this project. A template for pasting you data is found on a worksheet in the *IntergenSummary.xls*, which is distributed with this memo. We suggest submitting separate worksheet for each pairings, so analysts need make no sample selection decisions.

References

Irons, W. (1979). Cultural and biological success. In Chagnon, N. A. and Irons, W., editors, *Evolutionary biology and human social behavior*, pp.252–272. Duxbury Press, North Scituate.

Johnstone, J. and DiNardo, J. E. (1997). *Econometric Methods*. McGraw Hill, New York.

Wooldridge, J. M. (2001). *Econometric Analysis of Cross Section and Panel Data.*

A. Appendix

A.1 Handling zeros: log correction procedure

Taking the natural log of a zero number is problematic because the result is negative infinity so follow these procedures. The key idea is that in logs it does not matter if wealth is measured in dollars or in pennies. If unlogged wealth measures are large (100 or greater) add 1 to all observations before taking the natural logarithm. If unlogged wealth values are small (< 100), simply adding 1 will create a distortion, so first multiply all values by 100, then add 1, and then take the natural logarithm.

There may be cases where more than 30% of your wealth measure is zero. For example, half or less of a rural sample of men may have received one or more years of education. For such wealth outcomes we may ultimately recommend another method, such as using a Tobit regression equation. For the moment, however, we simply ask you to estimate β as explained above and note the proportion zeros for each measure on the beta summary worksheet in *IntergenSummary.xls*.

A.2 Data with missing parents

In some populations you may prefer to estimate β's for par-s or par-d in addition to estimates for mother-child and father-child separately. In such a procedure you may encounter cases where information is missing from a father or a mother. Since we are interested in the lifetime wealth accumulated by a son or daughter, the intuitions of the ethnographer/historian should decide whether to (a) use the available parent's wealth only, (b) roughly estimate the missing parent's wealth using the relationship between mother's and father's wealth (where both sets of information are available) in the population, (c) omit these cases with missing data from the analysis entirely, or (d) chose to estimate only the f-s pairing (and drop cases where the father's information is missing).

Use option (a) if the offspring's inheritance in the population is the sum or average of the parent's wealth, rather than the wealth of the highest or lowest parent and the missing partner "went missing" before he/she had an influence on inheritance (e.g. a mother who died in childbirth, or a father who disappeared when the offspring was an infant). Option (b) may be the way to go if both parents likely did have an effect on the

offspring's wealth and the statistical association between the two parents' wealth is high enough to give a good rough approximation to the missing parent's wealth.

If the unobserved unlogged wealth of the missing parent is $y_{p,m}$, (m stands for missing) it can be estimated as some multiple of the wealth of the parent on whom we have data. Suppose it's the dad whose wealth data are missing. Then take the parental pairs for whom all data are available and estimate the wealth of the dad as a function of the wealth of the mom (ordinary least squares regression), then for the missing dads, use the estimated equation to infer the predicted wealth of the dad (just plug in the mom's wealth value in the equation and use the predicted amount as the inferred wealth of the missing dad. Do the equivalent thing for missing moms. A simpler method (but less satisfactory) is just to assume that the wealth of the missing parent is the wealth of the present parent times various values 0, 0.5, or 1.0. This simply attributes to the missing parent no wealth, half the wealth of the present parent, and an amount of wealth equal to the present parent, respectively.

Use option (c) if the number of missing parents is low enough relative to the sample size so omitting observations will not dramatically alter estimates and standard errors. If unsure, compare the outcomes from this option with those of options (a) or (b). Option (d) is best if the number of "complete couples" is small enough relative to the sample size that an estimated correlation coefficient may be far from the true value.

A.3 An example of STATA code

```
use "C:\Documents and Settings\mbmulder\My Documents\all mine\SFI Ineq
project\beta analyses\bet_abosi_2nd\kips(abosi)_par-s.dta" , clear
des

sum S_AGE
*to find the mean of 43.6 years*

gen SAM = S_AGE - 44
gen AALAND = SAM*FL_LAND
gen AASTOCK = SAM*FL_STOCK
generate SAR = 55 - S_AGE
replace SAR=0 if SAR<=0
generate SARD = SAR – [[[mean value of SAR]]
generate AARS = SARD*FL_RS
gen AAED = SAM*F_ED

sum S_LAND SL_LAND, detail
sum S_STOCK SL_STOCK, detail
sum S_RS SL_RS, detail
sum S_ED, detail
sum S_AGE, detail
sum S_RS SL_RS, detail

*need gini: download by in command line typing "ssc install ginidesc" -
then run ginidesc (varname)
```

```
*requires version 9 stata and requires running each line separately
(why?)
ginidesc S_LAND
ginidesc S_STOCK
ginidesc S_RS
ginidesc S_ED

reg SL_LAND FL_LAND F_AGE F_AGE2 S_AGE S_AGE2 AALAND, cluster(FCODE)
reg SL_STOCK FL_STOCK F_AGE F_AGE2 S_AGE S_AGE2 AASTOCK, cluster(FCODE)
reg SL_ED FL_ED F_AGE F_AGE2 S_AGE S_AGE2 AAED, cluster(FCODE)
reg SL_RS FL_RS F_AGE F_AGE2 S_AGE S_AGE2 SARD AARS, cluster(FCODE)

lowess S_LAND F_LAND, jitter(7)
lowess SL_LAND FL_LAND, jitter(7)
lowess S_STOCK F_STOCK, jitter(7)
lowess SL_STOCK FL_STOCK, jitter(7)
lowess S_ED F_ED, jitter(10)
lowess SL_ED FL_ED, jitter(10)
lowess S_RS F_RS, jitter(7)
lowess SL_RS FL_RS, jitter(7)
```

A.4 Selection bias and the Heckman correction

For Phase I of the project we consider this extension as a form of recreational econometrics. We will do Heckman corrections and other extensions in Phase II. Suppose the problem (as in the example given above) is outmigration. The idea is that you first estimate a migration equation using a probit:

$$\Pr(\text{migrate}) = \alpha_0 + \alpha_1 y_p + Y_\gamma + \mu \qquad (4)$$

Where Y is some set of covariates that you think predict migration. Then you compute an inverse mills ratio, which is the marginal probit effect divided by the predicted probablity of migration, λ for each unmigrated observation. Then you estimate (1) including λ as a covariate (i.e. in X).. In STATA, there is a command called “heckman” that implements all of the above;(note that you must first use logistic regression to determine the variable(s) that predict migration, and then include these in the “selection” portion of the Heckman model; note too that the selection portion must include some predictors that are not in the main equation). In R, however, you need to load the package micEcon. In it, there is a “heckit” command.