Kinship Correlations and Intergenerational Mobility*

M. Dolores Collado
Universidad de Alicante

Ignacio Ortuño-Ortin†
Universidad Carlos III

October 2016
Very preliminary draft

Abstract

We propose a new methodology to estimate long-run intergenerational socioeconomic mobility. The specification is general enough to encompass the standard model as well as the specification recently proposed by Gregory Clark. Our approach does not require to have information about the variable of interest for individuals in several generations and make use of the correlations among individuals with different degrees of kinship in the same generation. In our empirical application we use census data from a Spanish region and find a high degree of persistence that corroborates some of Clark's findings.

1 Introduction

The analysis of the degree of socioeconomic intergenerational mobility has attracted the attention of many economists in recent years (see for example Chetty et al 2014). Part of the interest on this topic is due, at least in the case of income mobility, to its possible relation with the increasing income inequality experienced recently in some economies (Corak 2013). An additional factor to explain this interest is the existence of recent studies showing that mobility in the long-run is perhaps much lower than what most economists used to think (Long and Ferrie 2013, Clark 2014, Lindahl et al 2014). This recent literature has started to change the standard view about mobility across multiple generations, which used to assume that the correlation between grandparents and grandchildren outcomes is basically the square of the parent-offspring correlation. Since for most relevant outcomes such as income or education, parent-offspring correlations are always moderate, economists had often assumed that the correlation between individuals in one generation and their ancestors in different generations decreases really fast as we go back in time, so that after, say, three or four generations the link is already very weak. However, recent empirical studies suggest a much higher persistence rate in socioeconomic status and a significant link with grandparents and even with great-grandparents (Lindahl et al 2014).

An important contribution in this area has been the work by Clark (2014) who claims that mobility across several generations, for income as well as for other outcomes, is low due to the existence of a latent variable, the "underlying social competence" of families, which is inherited from parents and has a high persistence rate. If such latent variable indeed plays an important role in the transmission of socioeconomic status the standard regressions of offspring outcomes against

*We are very grateful to Gregory Clark and Jan Stuhler for very helpful comments and suggestions

†Corresponding author: Departament of Economics, Universidad Carlos III de Madrid, C/ Madrid 126, 28903-Getafe (Madrid), Spain. iortuno@eco.uc3m.es

parent's outcomes will be downward biased and the true persistence will be higher than suggested by the regression coefficient. Clark (2014) asses the role of such latent variable using a methodology based on the use of surnames. His approach requires information on the outcome of interest for individuals in several generations. Using data from a series of countries and periods of time Clark finds a very low degree of intergenerational mobility. Furthermore, the degree of mobility is very similar across countries and time. Lindahl et al (2014) and Braun and Stuhler (2016) also find a low degree of long-run intergenerational mobility but not as low as in Clark (2014).

The main problem with the approach adopted in these works is the data requirement, because for many countries is difficult to obtain comparable information for more than two generations about outcomes such as income or educational levels. For instance, for many countries, we typically find that there is very little variation in years of formal education for older generations because the majority of the population had just basic education. Here we propose a new approach to asses the degree of long-run intergenerational mobility that does not require information on previous generations. To apply our methodology we just need "horizontal" information, that is, information about individuals of the same generation, or very close generations, who are relatives of a certain degree, for example siblings, cousins, second cousins, parent-child, uncle-nephew. The idea behind our method is quite simple. Say that we would like to asses the link between grandparents and grandsons but we don't have data for grandparents to directly measure it. But if instead we have good data for cousins we can infer the grandparents-grandsons link from the cousins links. Thus, horizontal information can overcome the lack of vertical information. In particular, we compute the correlation for years of schooling for different degrees of kinship (brothers, fathers-sons, first-cousins and uncles-nephews) using census data from a Spanish region. If we have enough of these moments we can calibrate all the parameters of a reduced form model on intergenerational mobility. Our results from this calibration exercise are very much consistent with the high persistence hypothesis proposed by Clark. In particular we find that the persistence rate for the "underlying social competence" of families is around 0.8. Consistent with this result, our approach predicts that the educational levels of individuals in the current generation are still correlated in a non-negligible magnitude with the socioeconomic status of their ancestors as much as four or five generations back in time[1].

Our approach is also related to the literature on siblings correlations (See Jäntti and Jenkins 2015 for a recent literature review). Most of the papers in this literature aim at estimating the impact of family background on an observable outcome such as income, education, etc. The family background is a latent component that accounts for all factors shared by siblings that are orthogonal to the parental outcome. We extend these models by decomposing the family background into an inheritable and a non-inheritable component. The idea is that by using correlations on outcomes of relatives of different degrees of kinship we are able to disentangle the non-inheritable part of family background that is only shared by sibling to the inheritable part that is also partially shared by cousins, second cousins, etc., through their common ancestor. We also allow the inheritable component to be correlated to the parental observable outcome. Our decomposition of the family background into the inheritable and non-inheritable components is similar to the nature and nurture decomposition. There are several papers that aim at estimating the relative importance of nature and nurture by looking at the correlations in observed outcomes for different type of siblings like MZ twins, DZ twins, siblings, half siblings, adoptees, etc (See Björklund and Salvanes 2011 for a

[1] Collado et al (2014) analyze long-run mobility in the same Spanish region using census data from the XIX and the XX century. They find a higher level of mobility than the one in this paper. This discrepancy might be explained because they only consider two socioeconomic levels whereas than here individuals are classified according to 10 possible levels of education (years of schooling)

literature review).

Our approach requires data for a large sample of individuals and their relatives. However, in some cases such information about relatives is not available. For that reason, we propose a second approach that can partially overcomes such problem whenever the data contains the surnames of individuals.[2] Under this second approach we don't need to identify relatives with certainty but just in a probabilistic way. Thus, using the frequencies of surnames we are able to estimate the probability that two individuals with the same surname are relatives of a certain degree (brothers, cousins, and so on). In this way, we compute "expected moments" and provide an alternative way to calibrate the model. Moreover, we use this second approach as a test to validate our first approach.

Thus, we propose two new methods to asses the degree of long-run mobility that can be seen as complementaries to the one used recently by several economists. Our first method requires information about relatives whereas the second method requires information about surnames. We believe that our methods have an important advantage since they do not require information on individuals in previous generations, and therefore, they can be applied to study long-run intergenerational mobility in many countries in which there is no comparable data on individuals in several generations.

The results suggest that long-run intergenerational mobility might be quite low. Because we only calibrate a reduced form model it's difficult to get policy conclusions from our findings. However, the fact that the latent variable underlying the social competence of families explains a high part of the variance in levels of education suggests that public intervention policies should pay more attention to the role of the family.

The paper proceeds as follows. Section 2 sets out the basic model and develops our first method. Section 3 presents our main empirical findings. Section 4 provides our second method and additional empirical findings. Section 5 contains an additional validation test of the first method. Section 6 extends the basic model to account for assortative mating and the potential influence of mothers. Section 7 concludes. We include some additional information about the models in the Appendixes.

2 Theory

Suppose that y is the outcome of interest in our economy, for example income, education or wealth. Since in our empirical exercise such outcome will be the level of education henceforth we identify y with years of schooling but all our theoretical results are valid to study other outcomes as, for example, income. We want to study the link of such variable y between individuals and their ancestors. The model presented here considers only males, but it will be extended in Section 6 to cover individuals of any gender. We consider a reduced form of Becker-Tomes (1979) model similar to the one in Solon (2014)

$$y_t^i = \beta y_{t-1} + z_t^i + x_t + u_t^i \tag{1}$$

where $t-1$ denotes the father's generation and t the children's generation, y_{t-1} denotes years of schooling of his father, z_t^i denotes a latent variable that is inherited from the parents, x_t is a shock shared by all brothers in the family which is uncorrelated with the other variables (in particular with z_t), and u_t^i is an individual's white-noise error term. In principle the variable z_t^i might include common genes and family values and depending on the whether there are perfect credit markets

[2] The way we use surnames here is completely unrelated to the way surnames are used in Clark (2014), Collado et al (2014), Guell et al (2015), and Chetty et al (2014b), or the way names are used in Olivetti and Paserman (2015). Here we only use surnames to establish the probability that two people bearing the same surname are relatives.

or not father's wealth could be also part of it. The variable x_t could capture factors like the type of neighborhood, common friends and perhaps the influence of one sibling on another. These are factors that siblings might share but are not inherited from parents.

The latent variable z_t^i is often omitted in this type of analysis and it has been introduced by Clark (2014) who sees it as the "underlying social competence" of families and assumes that

$$z_t^i = \gamma z_{t-1} + e_t^i \tag{2}$$

where z_{t-1} denotes the father's value of such latent variable and e_t^i is an individual white-noise term.[3] Thus, the "underlying social competence" is passed from fathers to their sons with persistence rate γ.

The "standard approach" does not consider the existence of such variable z while Clark's approach assumes that $\beta = 0$.[4] We take a more general view and a priory do not exclude any possibility and let the data determine which model is the correct one. If the standard approach is the correct one we should find that z is zero (or close to zero) whereas if Clark's model is the correct one we should find very low values of β and significant values of z and γ.

We suppose we are in the steady state and therefore the persistence parameters β and γ, the distribution of z_{t-1} and y_{t-1} and all the covariances remain the same across generations. Under the standard approach the parameter β is estimated by regressing child's years of schooling on parental years of schooling. Since z is unobservable, estimating γ in Clark's model requires to have observations not only on sons years of schooling and fathers years of schooling but also on grandparents years of schooling (see Clark 2014). Unfortunately, in many cases it's difficult to get good data on the outcome of interest for a large sample of individuals from more than two different generations. We propose a new methodology that only requires information on the years of schooling of individuals in two generations, and sometimes only information from one generation. The idea behind our method is quite simple: if the model specified by (1) and (2) is correct and we have the necessary data, we can compute the correlations on years of schooling for different degrees of kinship, for example the correlation for brothers, father-son, first-cousins, second-cousins, uncle-nephew and so on. If we have enough of these moments we can calibrate all the parameters of the model.[5] To compute some of these moments we need information about individuals from the same generation (brothers, first-cousins, second cousins,...) and if we have information about a previous generation we might also compute correlations for father-sons and uncle-nephews. In some cases one can have data on grandparents and compute the grandfather-grandson correlation.

Write as σ_y^2, σ_z^2 and σ_x^2 the variances of y, z and x. We can write the covariance in years of schooling between brothers in generation t as

$$Cov_b(y_t^i, y_t^j) = \beta^2\sigma_y^2 + \frac{2\beta\gamma\sigma_z^2}{1-\beta\gamma} + \gamma^2\sigma_z^2 + \sigma_x^2$$

and the correlation as

$$\rho_b = \frac{Cov_b(y_t^i, y_t^j)}{\sigma_y^2} \tag{3}$$

In Appendix A we show the corresponding correlations for father-son, grandfather-grandson, uncle-nephew, first-cousins, second-cousins and third-cousins. If we know at least four of the previous correlations we can calibrate the model to determine the values of those unknowns.

[3] We assume that siblings errors e_t^i are uncorrelated. Our results are robust to imposing the restriction that siblings get the same realization of e_t.

[4] Clark (2014) does not need to include x because he does not use data on brothers, cousins, etc.

[5] If we have enough moments we can consider an even more general model in which the parameter β in the current generation could be different from the one in the previous generation.

We might not find an exact solution to such system of equations within the range of values that we consider feasible in our economy. In that case we determine $\beta, \gamma, \sigma_z^2$ and σ_x^2 by solving the following minimization problem

$$Min_{\{\beta,\gamma,\sigma_z^2,\sigma_x^2\}\in F} \quad \sum_{i\in C} p_i(\rho_i - \overline{\rho}_i)^2 \tag{4}$$

where $\overline{\rho}_i$ is the value of the observed correlation, p_i is the sample size used to calculate correlation ρ_i, F is the set of feasible values for the four unknowns, and C is the set of correlations for which we have reliable data (for example brothers, cousins, second-cousins, fathers-son)..

3 Empirical Application I

In this section we apply the method proposed in section 2 to calibrate the model using census data from the Spanish region of Cantabria.

3.1 The data

To apply our methodology we need data on extended families. The 2001 population census for Spain, which is available nationwide, does not allow to identify families unless they are living in the same house. However, for the region of Cantabria we have information on the full name of each person and we can use this information to identify fathers and sons, brothers, uncles and nephews, and cousins. The census contains information, among other variables, on the gender, age and educational level of all individuals living in the region (526, 339 persons). We define the t-generation as all males born in Cantabria between 1956 and 1976 (71, 479 males and 68, 830 females) and the $(t-1)$-generation as their parents. Surnames in Spain are passed from parents to children according to the following rule: A newborn person, regardless of gender, receives two surnames that will keep for life. The first surname is the father's first surname and the second the mother's first surname. This name convention allows us to identify fathers and mothers. For each person i in generation t we define the set of potential parents as all the couples born before 1956 such that the husband first surname coincides with person i first surname and the wife first surname coincides with person i second surname. Then, we say that we identify the parents if there is only one couple in the set of potential parents and the age difference between both parents and the son is at least 16 years. We identify the parents for 25, 860 males and 24, 610 females which is approximately 36.2% and 35.8% of the male and female population respectively. We use the information on the educational level to assign years of schooling to each person following Calero et.al.[6] We measure the years of schooling as deviations from the corresponding mean in each generation.

The matched sample almost 2 years younger than the unmatched one. The reason is that the older a person is the more likely the parents are not living together or one of them has died. Since the matched sample is younger it is also more educated (0.8 more years of schooling than the unmatched sample)

[6] We assign 2 years of education to those who did not complete primary education, 5 years to primary education, 8 to compulsory education, 10 to vocational training, 12 to secondary education, 15 to sort university degrees, 17 to long university degrees other than engineering and medicine, 18 for engineers and medical doctors and 19 for Ph.D. All our results are robust to other reasonable ways to assign years of education as, for example, assigning 0 years of education to those who did not complete primary education, 4 years to primary education, 9 to vocational training and 11 to secondary education.

Table 1								
	Men				Women			
	Matched		Unmatched		Matched		Unmatched	
	Mean	St. Dev	Mean	St. Dev	Mean	St. Dev	Mean	St. Dev
Age	33.61	5.91	35.42	6.16	33.70	5.92	35.50	6.15
Years of schooling	10.53	3.71	9.71	3.64	10.99	3.71	10.11	3.69
Number of observations	25,860		45,619		24,610		44,220	

Once we have identified parents and children, siblings are immediately identified. Finally, we identify siblings in the parents generation when there are only two individuals in that generation sharing the same two surnames. Once siblings in the parents generation are identified, uncles and nephews, and cousins are immediately identified. The strategy to identify siblings in the parents' generation is quite conservative in the sense that it is unlikely that we identify as brothers individuals who actually are not brothers, but we pay the price of having smaller sample sizes for cousins and uncles-nephews than for fathers-sons or brothers.

3.2 The benchmark case

We use the sample of males and the correlations between brothers, cousins (fathers are brothers), fathers and sons, and nephews and uncles (brothers of the fathers). The empirical covariances are first computed for each family and then averaged across families as suggested in Solon, Page and Duncan (2000). The empirical correlations are obtained by dividing the empirical covariances by the product of the standard deviations.[7] The empirical correlations and the number of families and pairs used to compute those correlations are presented in Table 2.

Table 2				
	Brothers	Father-son	Cousins	Uncle-nephew
Correlations	0.467	0.379	0.196	0.232
Number of families	6,022	17,663	746	1,921
Number of pairs	11,109	25,860	1,654	2,843

These correlations are within the values estimated in some other developed countries (see Hertz 2007 and Bjorklund and Salvanes 2011). We solve the minimization problem (4) with the four moments to obtain[8]

Table 3			
β	γ	σ_z^2	σ_x^2
0	0.790	6.586	2.303

We next compare the empirical correlations with the predicted correlations for these values of

[7] Notice that the standard deviation of y is 3.705 for the current generation and 3.831 for the parents generations. Therefore, the empirical correlations for fathers and sons, and uncles and nephews would have been slightly larger if we would have divided the covariances by the variance of y

[8] We use Mathematica to solve all the minimization problems in this paper. The codes and the details of all the computations are available upon request.

$\beta, \gamma, \sigma_z^2$ and σ_x^2

Table 4				
Correlations	Brothers	Father-son	Cousins	Uncle-nephew
Observed	0.467	0.379	0.196	0.232
Predicted	0.467	0.374	0.187	0.236
% Error	0%	-0.025%	-4.784%	1.859%

Since we don't have data on the correlation for other relatives, as grandfather-grandson, we cannot compare it with the correlation predicted by the model. However, we can compare the square of the father-son correlation with the grandfather-grandson correlation predicted by the model

Predicted grandfather-son	$(\text{father-son})^2$
0.299	0.144

This result is in accordance with Clark's view and with some recent empirical evidence (Lindahl et al 2015). The grandfather-grandson correlation is much stronger than the squared of the father-son correlation.

It's useful to asses how much of the total variance of y_t is explained by the different components of the model.

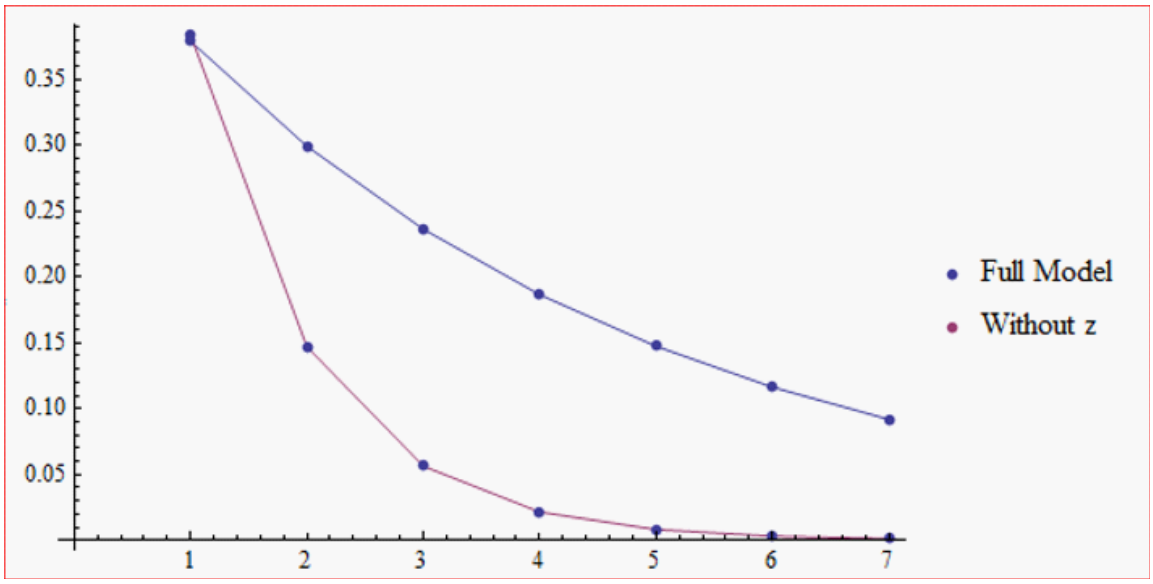
$$\begin{aligned} \sigma_y^2 &= \beta^2\sigma_y^2 + \sigma_z^2 + 2\beta Cov(y_{t-1}, z_t) + \sigma_x^2 + \sigma_u^2 \\ &= \beta^2\sigma_y^2 + \sigma_z^2 + 2\frac{\beta\gamma\sigma_z^2}{1-\beta\gamma} + \sigma_x^2 + \sigma_u^2 \end{aligned}$$

The part of the variance σ_y^2 explained directly by the father's years of schooling is $\beta^2\sigma_y^2$. The part directly explained by the latent variable z is σ_z^2 while the part explained by the shocks shared by brothers is σ_x^2. We standardize years of schooling so that $\sigma_y^2 = 1$ and obtain the following decomposition

Table 5				
Total explained	$\beta^2\sigma_{y_{t-1}}^2$	σ_z^2	σ_x^2	$2\beta Cov(y_{t-1}, z_t)$
0.648	0	0.480	0.168	0

Thus, the results in our benchmark case favour Clark's view that long-run mobility is much lower than suggested by most economists, and that a large share of the persistence is explained by an inherited latent variable with a high rate of persistence ($\gamma = 0.79$).

It's important to mention that both x and z are essential to obtain a satisfactory calibration of the model. Thus, if we drop x from the model and repeat our previous procedure we again obtain a very high value of the persistence parameter γ. However, in this case the (over-identified) model performs quite poorly at predicting the correlations. This is not surprising since previous works have already shown the importance of this type of shock to understand the correlations between brothers (Bjorklund and Salvanes 2011). If we now drop z, we obtain a non negligible $\beta = 0.384$ but the fit regarding cousins and uncle-nephew correlations is very poor. The predictions based on



these two models are presented in Table 6

Table 6				
Correlations	Brothers	Father-son	Cousins	Uncle-nephew
Observed	0.467	0.379	0.196	0.232
Dropping x Predicted	0.367	0.397	0.313	0.339
% Error	-21.47%	4.691%	59.82%	46.09%
Dropping z Predicted	0.475	0.384	0.070	0.182
% Error	1.794%	1,190%	-64.33%	-21.42%

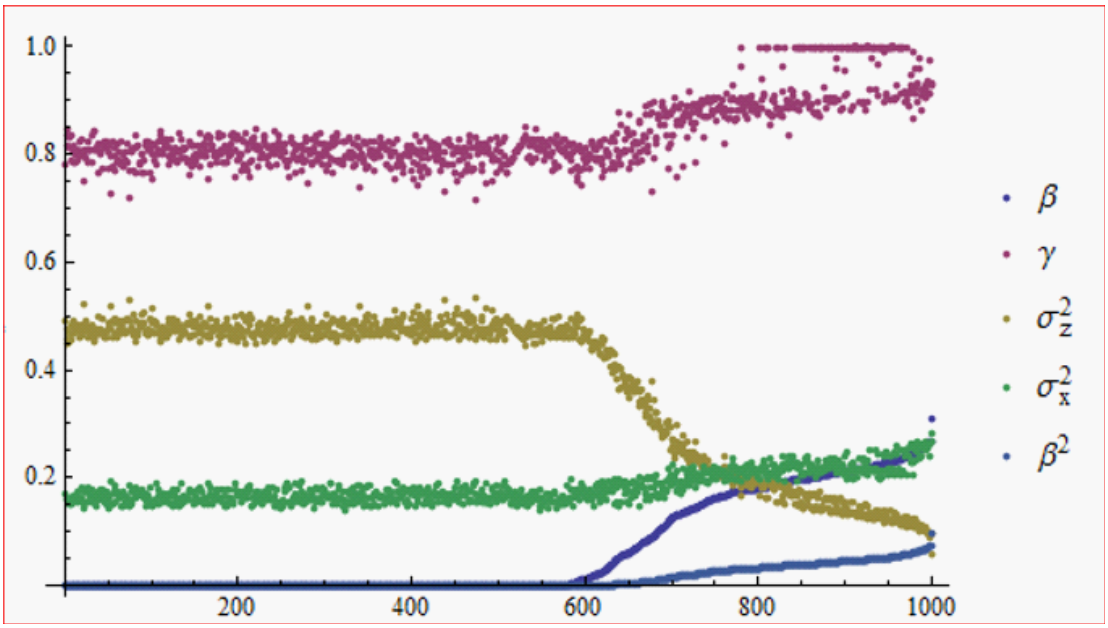
To better appreciate the consequences of these findings, Figure 1 shows the predicted correlations for individuals at the current generation and their ancestors, i.e. their fathers, grandfathers, great-grandfathers, etc., based on the full model and on the model without z.

Figure 1

As it is very clear from the figure, the persistence based on the model without z is low, so that after a few generations the influence of ancestors vanishes almost completely. Our approach, however, provides a more pessimistic view about intergenerational mobility in the long run. Thus, we find that, under the assumption of stability of the parameters of the model, the correlation between the levels of y of individuals in the current generation and the levels of y of their ancestors seven generations back in time is still as high as 9%.

3.3 Robustness checks

One possible concern with our previous analysis is about the robustness of our findings to changes in the values of the observed empirical correlations. For this reason we repeat our procedure for 1,000 different sets of values of the four correlations $\overline{\rho}_b, \overline{\rho}_{fs}, \overline{\rho}_{un}, \overline{\rho}_{c1}$. These values are obtained by



carrying out 1,000 random draws from our original sample, each draw selecting 75% of the original individuals in the current generation.[9] Table 7 reports the mean values of the four unknowns obtained under this procedure and compares it with the ones reported in the above benchmark case.

Table 7				
	β	γ	σ_z^2	σ_x^2
Mean value	0.065	0.840	5.136	2.500
Benchmark case	0	0.790	6.586	2.303

Table 8 shows the mean correlations predicted by the model and the observed ones

Table 8				
Mean correlations	Brothers	Father-son	Cousins	Uncle-nephew
Predicted	0.460	0.377	0.215	0.257
Observed	0.471	0.383	0.207	0.246

Figure 2 shows for these 1,000 subsamples the values of β and γ and the standardized σ_z^2, σ_x^2 as well as $\beta^2\sigma_y^2$ (the part of the variance of y_t directly explained by y_{t-1}).[10] The different cases are ordered according to the obtained values of β.

Figure 2

The basic facts that arise from this robustness check exercise are: i) The persistence parameter is always quite high and in almost all the cases greater than 0.75;[11] ii) The largest values of β are

[9] Since the number of pairs of uncle-nephew is not that large we consider that draws of 75% of the whole sample are better than draws of 50%. The results for the 50% case, which are in the vast majority of cases very similar to the ones reported here, are available upon request.

[10] Notice that we have normalized σ_z^2 and σ_x^2 to σ_y^2.

[11] Only 18% of the estimated γ are smaller 0.79, the value found for the benchmark case.

around 0.2, but even for those cases the part of the total variance of y_t explained directly by y_{t-1} is small and in all the simulations but one, smaller than the part of the variance explained by the latent variable z.

Thus, the main findings and conclusions obtained in the benchmark case are robust to these changes in the values of the observed correlations.[12]

4 A second approach

In many cases the data does not contain explicit information about the different degrees of kinship of the individuals. In our previous empirical application we could determine kinship for a large sample of individuals thanks to the fact that in that census each individual has two surnames. In many countries, however, names contain just one surname so that we can't use surnames to identify relatives. In this section we propose a methodology to overcome such lack of information for cases in which the data contains the surname of each individual. We first need to estimate the probability that two randomly chosen individuals from the same generation who bear the same surname are brothers, first-cousins, second-cousins and so on. Such probabilities will depend on the total number of individuals bearing such surname and the population growth (in the parent's generation). Once we have established those probabilities we can follow an approach similar to the previous one but using expectations of moments.

Let's define the "size" of a surname, in a given generation, as the total number of individuals in that generation bearing such surname. Appendix C explains how to compute $r_k(g,n)$, the probability that, given a population growth rate of g, two individuals bearing the same surname of size n are relatives of degree $k = 1, 2, 3, ..$,where $k = 1$ indicates that they are brothers, $k = 2$ denotes first-cousins, $k = 3$ denotes second-cousins, and so on. These probabilities are calculated under the assumption that all individuals with the same surname share a common ancestor (in the paternal line) with such surname. This might seem as a strong assumption since in many real cases two individuals bearing the same surname, even in the same geographical area, are not related (an example of this could be the Welsh surname Jones). However, most of those surnames are often of a very large "size" and the relevant probabilities are very low in any case. Thus, our analysis will be more accurate when there are enough surnames that are not too large in size.

Once we have those probabilities $r_k(g,n)$ for surname sizes $n = 2, 3, .., S$, our second methodology works as follows. Assume that the size of the surname and the variable y are independent random variables. Take all individuals in the current generation with a given surname of size n. Let $(y^1, y^2, ..., y^n)$ be their vector of individual years of schooling. We consider for each individual i his value y^i as a random variable with variance σ_y^2 (the variance in the whole population in that generation). Consider all n^2 pairs of individuals that can be formed from those n individuals (with repetitions). Among those n^2 pairs we have n pairs with degree of kinship $k = 0$, i.e. pairs formed by the same individual. We write the share of such pairs as $P_0(n) = \frac{n}{n^2} = \frac{1}{n}$. Say that the probability that two randomly chosen **different** individuals are brothers is given by $r_1(g,n)$. Then, the share of pairs who are brothers is $P_1(g,n) = (1 - P_0(n))r_1(g,n)$. In general we have that the share of pairs with degree of kinship $k > 0$ is

$$P_k(g,n) = (1 - P_0(n))r_k(g,n)$$

[12] We have carried out additional robustness checks. In particular we first have repeated the same robustness check as the one here but for draws of 50% of the original sample, and second we have solved our minimization problem for other 256 economies obtained by considering values of the correlations within a ±10% deviation from the benchmark case values. The results are again similar and are provided upon request.

It's true that for random variables $y^1, y^2, ..., y^n$

$$Var(\sum_{i=1}^{n} y^i) = \sum_{i,j} Cov(y^i, y^j) \tag{5}$$

We write the covariance for two individuals with degree of kinship k as $Cov_k(y^i, y^j)$, $s \geq 1$, and in the right-hand term of (5) the share of such pairs is $P_k(g, n)$. Thus, we have

$$Var(\frac{\sum_{i=1}^{n} y^i}{n}) = P_0(n)\sigma_y^2 + \sum_{k=1}^{\infty} P_k(g, n)Cov_k(y^i, y^j) \tag{6}$$

Equation (6) is the main one under this approach. Notice that in most countries one can estimate the parameters $P_k(g, n)$ and σ_y^2 from available data about population growth and the distribution of y, respectively.

The left-hand term in (6) can also be estimated if we have individual information for a large enough number of different surnames of size n. Thus, suppose that we have information for all individuals with surnames from the set $S_n = \{s_n^1, s_n^2, ..., s_n^{N_n}\}$ all of them of size n. Say that the observed mean years of schooling for individuals with surname s_n^i is $\overline{y}^{s_n^i}$ and the observed mean value among all individuals with surnames from the set S_n is $\overline{y}^n$. We then use the sample variance to estimate the population variance and

$$Var(\frac{\sum_{i=1}^{n} y^i}{n}) = \frac{1}{N_n - 1} \sum_{i=1}^{N_n} (\overline{y}^{s_n^i} - \overline{y}^n)^2 \tag{7}$$

Since $\beta \leq 1$ and $\gamma \leq 1$ the term $Cov_k(y^i, y^j)$ in (6) becomes small fast. For that reason in the empirical application we only consider degrees of kinship up to $k = 6$, i.e. up to fifth degree cousins. Thus, using (7) and a maximum degree of kinship of $k = 6$, instead of (6) we consider the following equations

$$\frac{1}{N_n - 1} \sum_{i=1}^{N_n} (\overline{y}^{s_n^i} - \overline{y}^n)^2 = P_0(n)\sigma_y^2 + \sum_{k=1}^{6} P_k(g, n)Cov_k(y^i, y^j), \ n = 2, 3, ...S \tag{8}$$

where S is the maximum surname size considered. In our empirical exercise we take $S = 15$. Notice that for large values of the surname size n the expected covariance $P_0(n)\sigma_y^2 + \sum_{k=1}^{4} P_k(g, n)Cov_k(y^i, y^j)$ should approach to zero.

The rest of the method is similar to the one in the first approach. If we have really good data about a large number of surnames of size up to $S \geq 4$ we solve for $\beta, \gamma, \sigma_z^2$ and σ_x^2 from the set of equations in (8). In some cases the solution is not in the feasible set F and we then solve for

$$Min_{\{\beta,\gamma,\sigma_z^2,\sigma_x^2\} \in F} \sum_{n=2}^{S} N_n \left(\frac{\sum_{i=1}^{N_n} (\overline{y}^{s_n^i} - \overline{y}^n)^2}{N_n - 1} - P_0(n)\sigma_y^2 - \sum_{k=1}^{4} P_k(g, n)Cov_k(y^i, y^j) \right)^2 \tag{9}$$

In many cases one can also obtain very accurate information about some moments, in particular about the correlations $\overline{\rho}_b, \overline{\rho}_{fs}$. In that case we might want to impose in the minimization program (9) the additional constraints $\overline{\rho}_b = \rho_b$, and $\overline{\rho}_{fs} = \rho_{fs}$.

4.1 Empirical Application II

In this section we apply our second method to calibrate again our model using census data from the Spanish region of Cantabria. In this case, however, we don't restrict our sample to the set of relatives that can be identified thanks to the fact that all individuals bear two surnames. Here we take all the 25-45 years old male individuals born in the region of Cantabria (71,515 individuals). For each individual we consider only his first surname. We obtain the following frequencies of surnames sizes from size 2 up to size 15.

Table 9

Size	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Frequency	800	508	312	240	171	123	94	87	66	75	52	42	37	26

Appendix C shows how we compute the probabilities $r_k(g,n)$. In the benchmark case of this section we assume a population growth rate of $g = 1.15$. To compute those probabilities we also assume that the number of male descendants of any individual follows a Poisson. We check that the predictions of our demographic model are very close to the data provided by the Spanish Statistical Institute. As an example, Table 10 shows, for surnames of size two and for surnames of size 10, the probability that two randomly chosen individuals bearing the same surname are relatives of different degrees

Table 10

	Brothers	Cousins	Second-cousins	Third-cousins	Forth-cousins
$n = 2$	0.772	0.151	0.045	0.016	0.007
$n = 10$	0.136	0.175	0.194	0.163	0.116
Probabilities $r_k(g,n); g = 1.15$					

We compute for each of the different surname sizes the sample variance $\frac{\sum_{i=1}^{N_n}(\overline{y}^i_{s_n}-\overline{y}^n)^2}{N_n-1}$. We next solve the minimization problem (9) with the additional constraints that the correlations for brothers and for father-son have to be equal to the observed correlations, i.e. $\overline{\rho}_b = \rho_b$, and $\overline{\rho}_{fs} = \rho_{fs}$, where those empirical correlations are the same as in the previous section ($\overline{\rho}_b = 0.459$, and $\overline{\rho}_{fs} = 0.374$). Table 11 shows the result under this alternative approach and compares it with the one obtained in the benchmark case of the previous approach

Table 11

	β	γ	σ_z^2	σ_x^2
Alternative approach	0	0.800	0.468	0.158
Previous approach (benchmark case)	0.099	0.850	0.301	0.174

The two calibration approaches produce reasonably similar results. The second approach gives even more weight to the latent variable z. Since this approach depends also on the way we construct the probabilities $r_k(g,n)$, we have carried out some robustness checks on the way we model the demography. The results are quite similar and are available upon request.

It's interesting to see that for these values of $\beta, \gamma, \sigma_z^2$ and σ_x^2, the grandfather-grandson correlation predicted by the model is 0.3, even higher than the predicted with the values obtained under the first approach. The correlations predicted for cousins and uncle-nephew are 0.192 and 0.24, very close to the empirical ones obtained in the first exercise, 0.197 and 0.262 respectively.[13]

[13] Notice that the set of families used in the first exercise is a subset of the sample used in this second case, so that in principle the correlations could be different.

In Figure 2 we put together the observed values of the variances $\frac{\sum_{i=1}^{N_n}(\overline{y}^{s_n^i}-\overline{y}^n)^2}{N_n-1}$ and the predicted values under this second method. The figure also shows the predicted values if we assume that $\sigma_z^2 = 0$.

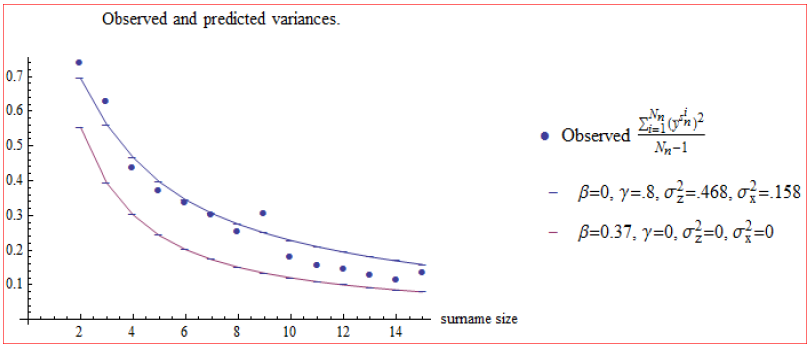


Figure 3

Thus, we believe that this second approach provides a useful methodology for cases in which we don't know the degree of kinship of individuals but do know their surnames.

5 Additional Verifications

Here we carry out an additional validation test of our first method proposed in section 2. We check how well our method predicts the expected covariances used in the second approach presented in section 4. Thus, we can use the values obtained in our benchmark case in section 2, $\beta = 0.099, \gamma = 0.850, \sigma_z^2 = 0.2301$ and $\sigma_x^2 = 0.174$, to compute the expected covariances $P_0(n)\sigma_y^2 + \sum_{k=1}^{4} P_k(g,n)Cov_k(y^i, y^j)$ in (8), and compare them with the empirical values $\frac{1}{N_n-1}\sum_{i=1}^{N_n}(\overline{y}^{s_n^i} - \overline{y}^n)^2$.

Figure 4 shows the result of this prediction exercise. In general the predictions are quite close to the observed data. Recall that the empirical values are obtained from a sample of individuals different from the one used in the calibration of the model, and that is why we can see this exercise as a validation test. Notice also that for surname sizes greater than 7 the empirical values seem to be more noisy, this is perhaps due to the fact that, as Table 8 shows, the sample sizes in those cases are relatively low. The figure also shows the prediction if we assume that $\sigma_z^2 = \sigma_x^2 = 0$ and β is given by the correlation fathers-sons.

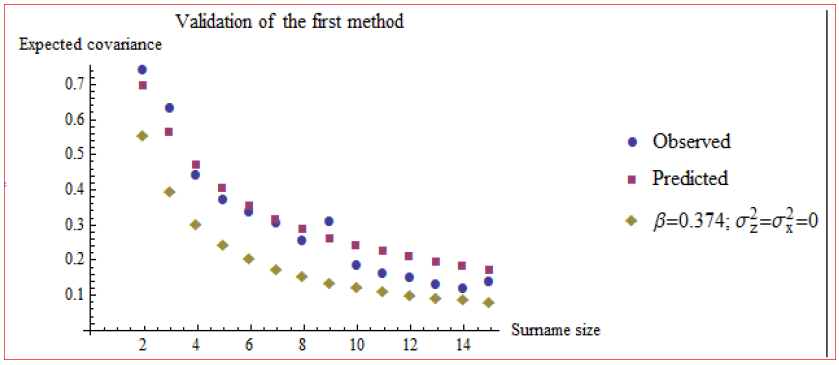


Figure 4

6 A model with assortative mating

The model we were considering did not take into account the potential influence of the mother in the outcome of the children. We now extend the previous model to incorporates mothers and assortative mating. We assume that the value of the output y for an individual from generation t is given by

$$y_t^k = \beta^k \widetilde{y}_{t-1} + z_t^k + x_t^k + u_t^k$$

where the superscript k stands for males ($k = m$) and for females ($k = f$). We assume that

$$\widetilde{y}_{t-1} = \frac{y_{t-1}^m + y_{t-1}^f}{2}$$

and the socioeconomic status of the child, z_t^k, depends on the father z_{t-1}^m as well as on the mother z_{t-1}^f

$$\begin{aligned} z_t^k &= \gamma^k \widetilde{z}_{t-1} + e_t^k \\ \widetilde{z}_{t-1} &= \frac{z_{t-1}^m + z_{t-1}^f}{2} \end{aligned}$$

Regarding the shocks, we assume that x_t^k is shared by all siblings of the same gender, can be correlated across siblings of different gender and is uncorrelated with the other variables (in particular with z_t and y_{t-1}). Finally u_t^k is an individual's white-noise error term.

We assume there is assortative mating both in years of schooling and in socioeconomic status. In particular we consider the following matching functions:

$$\begin{aligned} z_{t-1}^f &= r z_{t-1}^m + \omega_{t-1} \\ y_{t-1}^f &= \tau y_{t-1}^m + \varepsilon_{t-1} \end{aligned}$$

where ω_{t-1} and ε_{t-1} might be correlated but are uncorrelated to z_{t-1}^m and y_{t-1}^m. Thus, r measures the degree of assortative mating on the socioeconomic status (z), and τ the degree of assortative mating in years of schooling.

This model has 12 unknown parameters: β^m, γ^m, $\sigma^2_{z^m}$, $\sigma^2_{x^m}$, β^f, γ^f, $\sigma^2_{z^f}$, $\sigma^2_{x^f}$, $\sigma_{x^m x^f}$, r, τ and $\sigma_{w\varepsilon}$, and therefore we need at least 12 correlations between relatives of different kinship to calibrate these parameters. The inclusion of females into the model allows us to use the following 20 correlations: husband and wife, brothers, sisters, brother-sister, three types of male cousins (fathers are brothers, mothers are sisters, and father and mother are brother an sister) and analogously three of female cousins, four types of male-female cousins (fathers are brothers, mothers are sisters, father of the male is brother of the mother of the female, and mother of the male is sister of the father of the female), son-father, daughter-father, two types of nephew-uncle (brother of the father and brother of the mother) and analogously two of nice-uncle.[14] Notice that we have not included the correlations between individuals in generation t and their female relatives in the previous generation. The reason is that the education level of women in generation $t-1$ is not comparable to the education

[14] The formulas for these correlations as functions of the parameters are presented in Appendix B.

level of the current generation. Very few women born before 1956 went to university or even to high school, and the standard deviation of their years of schooling is much lower than for the current generation.[15]

6.1 Empirical Application III

We calibrate the parameters β^m, γ^m, $\sigma^2_{z^m}$, $\sigma^2_{x^m}$, β^f, γ^f, $\sigma^2_{z^f}$, $\sigma^2_{x^f}$, $\sigma_{x^m x^f}$, r, τ and $\sigma_{w\varepsilon}$ by solving the following minimization problem

$$Min_{\{\beta^m,\gamma^m,\sigma^2_{z^m},\sigma^2_{x^m},\beta^f,\gamma^f,\sigma^2_{z^f},\sigma^2_{x^f},\sigma_{x^m x^f},r,\tau,\sigma_{w\varepsilon}\}\in F} \quad \sum_{i\in C} p_i(\rho_i - \overline{\rho}_i)^2$$

where ρ_i are the theoretical correlations, $\overline{\rho}_i$ the empirical correlations, p_i the number of families used to calculate each correlation, F is the set of feasible values for the unknown parameters, and C is the set of correlations for which we have reliable data (husband-wife, brothers, sisters, brother-sister, etc.).

The calibrated parameters are presented in Table 12

Table 12			
β^m	γ^m	$\sigma^2_{z^m}$	$\sigma^2_{x^m}$
0.028	0.801	6.084	2.378
β^f	γ^f	$\sigma^2_{z^f}$	$\sigma^2_{x^f}$
0.000	0.809	5.800	2.129
$\sigma_{x^m x^f}$	r	τ	$\sigma_{w\varepsilon}$
1.732	0.976	0.541	0.000

The picture we obtain is again consistent with Clark's results. Both β^m and β^f are very close to zero, whereas γ^m and γ^f are around 0.8. This means that the observable outcome is transferred from parents to children indirectly through the latent variable z, which is very persistent. Another remarkable result is the large degree of assortative mating in z.

Regarding the fitting, we have computed the predicted correlations based on this parameters and we compare them with the empirical correlations. The results are presented in Table 13. The fit is reasonable taking into account that we try to match 18 moments using 12 parameters. As we expected, the empirical correlations based on many pairs of observations, which are likely to be quite accurate, are very close to the predicted ones, whereas those based on a smaller number of pairs are less close. This result is not only due to the weights used, a similar fit arises when we use equal weights (ES ESTO CIERTO?).

[15] The standard deviation of year of schooling for women is 3.004 in the parents generation and 3.711 in the current generation.

Table 13				
	N. pairs	Empirical	Predicted	Error (%)
husband-wife	21170	0.540	0.540	0.000%
brothers	11109	0.467	0.467	-0.006%
sisters	10316	0.437	0.437	0.000%
brother-sister	21017	0.414	0.414	-0.010%
male cousins				
(fathers are brothers)	1654	0.196	0.184	-6.064%
(mothers are sisters)	1539	0.213	0.197	-7.690%
(father and mother are brother and sister)	2586	0.209	0.190	-8.972%
female cousins				
(fathers are brothers)	1428	0.170	0.177	3.987%
(mothers are sisters)	1322	0.193	0.189	-1.906%
(father and mother are brother and sister)	2200	0.207	0.183	-11.622%
male-female cousins				
(fathers are brothers)	2919	0.186	0.180	-3.026%
(mothers are sisters)	2624	0.168	0.193	14.823%
(father-male is brother of mother-female)	2332	0.208	0.187	-10.259%
(mother-male is sister of father-female)	2425	0.150	0.186	24.269%
son-father	25860	0.379	0.379	-0.116%
daughter-father	24610	0.360	0.360	0.005%
nephew-uncle (brother of the father)	4640	0.232	0.236	1.754%
nephew-uncle (brother of the mother)	3457	0.228	0.243	6.515%
nice-uncle (brother of the father)	4467	0.216	0.228	5.467%
nice-uncle (brother of the mother)	3273	0.247	0.236	-4.554%

We now decompose the variance of y into its different components as

$$\sigma^2_{y^k} = (\beta^k)^2\sigma^2_{\widetilde{y}} + \sigma^2_{z^k} + \beta^k Cov(\widetilde{y}^k_{t-1}, z^k_t) + \sigma^2_{x^k} + \sigma^2_{u^k}$$

The results of these decompositions for males and females are presented in Table 14.[16] We can see that the model explains 62.5% of variance in years of schooling for males and 57.5% for females, with z and x accounting respectively for around 70% and 30% of the explained variance.

Table 14					
	Total explained	$\beta^2\sigma^2_{\widetilde{y}_t}$	σ^2_z	σ^2_x	$2\beta Cov(\widetilde{y}_{t-1}, z_t)$
Males	0.625	0.0005	0.444	0.173	0.0075
Females	0.575	0.0000	0.420	0.155	0.0000

Finally we compute the long-run mobility predicted by our model. Since we account for the influence of both parents, long-run mobility is not uniquely defined. We are going to use the paternal line and compute the predicted correlations for males in the current generation with their ancestors through the paternal line (i.e. the father, the father of the father, etc.). In Figure ? we compare those predicted correlations with the corresponding predictions based on a simple model without z (INCLUIR EL GRÁFICO).

[16] We standardize the different components to $\sigma^2_{y^m}$ for males and to $\sigma^2_{y^f}$ for females.

7 Conclusions

We have proposed a method to asses the degree of intergenerational mobility which takes into account the possibility that a substantial part of the persistence in socioeconomic status might be due to the existence of a latent variable that is inherited from parents. The method is based on the correlations between a series of relatives and does not demand much information about individuals in previous generations. Our findings suggest that indeed such latent variable plays a very important role and is the reason why persistence in socioeconomic status is much stronger than what is commonly thought. Thus, our results are in line with Clark (2014) claims about the low degree of social mobility in the long run. However, our exercise doesn't provide any new information in favor or against the possibility that the degree of intergenerational mobility is constant across different economies and time.

We have applied our method to asses the degree of intergenerational mobility in a Spanish region and have focused exclusively on the paternal line. The extension to other regions and countries and the inclusion of both genders and assortative mating are important tasks which are left for future research.

References

[1] Becker, Gary S., and Nigel Tomes (1979) "An equilibrium theory of the distribution of income and intergenerational mobility." *Journal of Political Economy* 87, pp 1153-1189.

[2] Bjorklund and Salvanes (2011) "Education and Family Background: Machanisms and Policies", in Eric A. Hanushek, S. Machin and L. Woessmann, eds. *Handbook of the Economics of Education*, vol3, pp 201-47, Amsterdam: North-Holland.

[3] Braun, S. and J. Stuhler (2016), "The Transmission of Inequality Across Multiple Generations: Testing Recent Theories with Evidence from Germany"

[4] Calero, J. J. O. Escardíbul, S. Waisgrais and M. Mediavilla, "Desigualdades socioeconómicas en el sistema educativo español". Secretaría General Técnica. Centro de Publicaciones. Ministerio de Educación y Ciencia (https://sede.educacion.gob.es/publiventa/desigualdades-socioeconomicas-en-el-sistema-educativo-espanol/investigacion-educativa/12330)

[5] Chetty, Raj, Nathaniel Hendren, Patrick Kline, and Emmanuel Saez, (2014) "Where is the land of opportunity? The geography of intergenerational mobility in the United States." *Quarterly Journal of Economics*, pp 2633-2679.

[6] Chetty, Raj, Nathaniel Hendren, Patrick Kline, Emmanuel Saez, and Nicholas Turner. (2014b). "Is the United States Still a Land of Opportunity?" *American Economic Review* 104 (5), pp. 141 47.

[7] Clark, Gregory (2014) *The son also rises: Surnames and the history of social mobility.* Princeton: Princeton University Press.

[8] Collado, M.D., I. Ortuño Ortín, and A. Romeu (2014), "Long-run intergenerational social mobility and the distribution of surnames".

[9] Corak, Miles (2013), Income Inequality, Equality of Opportunity, and Intergenerational Mobility", *Journal of Economic Perspectives*, 27,3, pp79-192

[10] Guell, M.C. Telmer and J. Rodriguez Mora, (2015), "The Informational Content of Surnames, the Evolution of Intergenerational Mobility and Assortative Mating", *Review of Economic Studies*, 82 (2): 693-735

[11] Hertz et al (2007) "The inheritance of educational inequality: International comparisons and fifty-year trends", *The B.E. Journal of Economic Analysis & Policy*

[12] Jäntti, M and S.P. Jenkins (2015) "Income mobility", Chapter 10, pp. 807–935, in Handbook of Income Distribution, Volume 2A, edited by A. B. Atkinson and F. Bourguignon, Elsevier.

[13] Lindahl, Mikale, Marten Palme, Sofia Massih, and Anna Sjorgen (2015) "Long-Term Intergenerational Persistence of Human Capital: An Empirical Analysis of Four Generations", *Journal of Human Resources*. 50, 1, 1-33.

[14] Long, J. and J. Ferrie, (2013), "Intergenerational occupational mobility in Great Britain and the United States since 1850", *American Economic Review*, 103(4): 1109-1137.

[15] Olivetti, C. and D. Paserman (2015), "In the name of the son (and the daughter): Intergenerational mobility in the United States, 1850-1940", *American Economic Review*, 105,8, pp 2695-2724

[16] Solon, Gary.(2014) "Theoretical models of inequality transmission across multiple generations", *Research in Social Stratification and Mobility*, 35, pp 13-18.

[17] Solon, Gary.(2015) "What Do We Know So Far about Multigenerational Mobility?", NBER, Working Paper 21053.

[18] Solon, G., M.E. Page and G.J. Duncan (2000) "Correlations between Neighboring Children in their Subsequent Educational Attainment", *The Review of Economics and Statistics*, 82, 3, 383-392

[19] Stuhler, Jan (2015), "Mobility Across Multiple Generations: The Iterated Regression Fallacy", IZA Discussion Paper 7072.

Appendix A

Consider a reduced form of Becker-Tomes (1979) model similar to the one in Solon (2014)

$$y_t^i = \beta y_{t-1} + z_t^i + x_t + u_t^i$$

where $t-1$ denotes the father's generation and t the children's generation, y_{t-1} is father's years of schooling, y_t^i is child i's years of schooling, z_t^i is the child i's status, x_t is not inherited, is uncorrelated with y_{t-1} and z_t^i but is shared among brothers, and u_t^i is a random error that is uncorrelated with y_{t-1} and z_t^i.

Status is partially inherited so that

$$z_t^i = \gamma z_{t-1} + e_t^i$$

where e_t^i that is not correlated across brothers. Notice that when $\beta = 0$ we are in Clark's model.

We assume that the second order moments of all variables are time invariant. We present below the formulas for the covariances in years of scholing for relatives of different degrees of kinship. The correlations are computed by dividing the covariances by the variance of y.

Covariances

Brothers

We first compute the covariances between y_{t-1} and z_{t-1}

$$\begin{aligned} Cov(y_t^i, z_t^i) &= Cov(\beta y_{t-1} + z_t^i, z_t^i) = \beta Cov(y_{t-1}, z_t^i) + \sigma_z^2 \\ &= \beta Cov(y_{t-1}, \gamma z_{t-1}) + \sigma_z^2 = \beta\gamma Cov(y_{t-1}, z_{t-1}) + \sigma_z^2 \end{aligned}$$

and in the steady state we have $Cov(y_t, z_t) = Cov(y_{t-1}, z_{t-1})$, so that

$$Cov(y_{t-1}, z_{t-1}) = \frac{\sigma_z^2}{1 - \beta\gamma}$$

and the covariance between brothers is

$$Cov_b(y_t^i, y_t^j) = \beta^2\sigma_y^2 + 2\beta\gamma Cov(y_{t-1}, z_{t-1}) + \gamma^2\sigma_z^2 + \sigma_x^2$$

Cousins

We first compute the following covariances for their fathers (who are brothers)

$$Cov_b(z_{t-1}^i, z_{t-1}^j) = \gamma^2\sigma_z^2$$

and

$$Cov_b(y_{t-1}^i, z_{t-1}^j) = \beta\gamma Cov(y_{t-2}, z_{t-2}) + \gamma^2\sigma_z^2$$

The covariance for male cousins whose fathers are brothers is

$$Cov_c(y_t^i, y_t^j) = \beta^2 Cov_b(y_{t-1}^i, y_{t-1}^j) + 2\beta\gamma Cov_b(y_{t-1}^i, z_{t-1}^j) + \gamma^2 Cov_b(z_{t-1}^{m,i}, z_{t-1}^{m,j})$$

Son-Father

$$Cov_{sf}(y_t^i, y_{t-1}) = \beta\sigma_y^2 + \gamma Cov(y_{t-1}, z_{t-1})$$

Nephew and uncle (brother of the father)

$$Cov_{neph-u}(y_t^i, y_{t-1}^j) = \beta Cov_b(y_{t-1}^i, y_{t-1}^j) + \gamma Cov_b(y_{t-1}^i, z_{t-1}^j)$$

Second cousins

We first compute the following covariances for their fathers (who are cousins)

$$Cov_c(z_{t-1}^i, z_{t-1}^j) = \gamma^2 Cov_b(z_{t-2}^i, z_{t-2}^j)$$

and

$$Cov_c(y_{t-1}^i, z_{t-1}^j) = \beta\gamma Cov_b(y_{t-2}, z_{t-2}) + \gamma^2 Cov_b(z_{t-2}^i, z_{t-2}^j)$$

The covariance for second cousins whose fathers are brothers is

$$Cov_{c2}(y_t^i, y_t^j) = \beta^2 Cov_c(y_{t-1}^i, y_{t-1}^j) + 2\beta\gamma Cov_c(y_{t-1}^i, z_{t-1}^j) + \gamma^2 Cov_c(z_{t-1}^i, z_{t-1}^j)$$

Third cousins

We first compute the following covariances for their fathers (who are second cousins)

$$Cov_{c2}(z^i_{t-1}, z^j_{t-1}) = \gamma^2 Cov_c(z^i_{t-2}, z^j_{t-2})$$

and

$$Cov_{c2}(y^i_{t-1}, z^j_{t-1}) = \beta\gamma Cov_c(y_{t-2}, z_{t-2}) + \gamma^2 Cov_c(z^i_{t-2}, z^j_{t-2})$$

The covariance for second cousins whose fathers are brothers is

$$Cov_{c3}(y^i_t, y^j_t) = \beta^2 Cov_{c2}(y^i_{t-1}, y^j_{t-1}) + 2\beta\gamma Cov_{c2}(y^i_{t-1}, z^j_{t-1}) + \gamma^2 Cov_{c2}(z^i_{t-1}, z^j_{t-1})$$

Appendix B

We now extend the previous model to incorporates mothers and assortative mating. We assume that the value of the output y for an individual i from generation t is given by

$$y^k_t = \beta^k \widetilde{y}^k_{t-1} + z^k_t + x^k_t + u^k_t \tag{10}$$

where the superscript $k = m$ stands for males and $k = f$ for females, and we assume that

$$\widetilde{y}^k_{t-1} = \alpha^k_y y^m_{t-1} + (1 - \alpha^k_y) y^f_{t-1}$$

and the "underlying social competence" of the child, z^k_t, depends on the father z^m_{t-1} as well as on the mother z^f_{t-1}

$$\begin{array}{c} z^k_t = \gamma^k \widetilde{z}^k_{t-1} + e^k_t \\ \widetilde{z}^k_{t-1} = \alpha^k_z z^m_{t-1} + (1 - \alpha^k_z) z^f_{t-1} \end{array} \tag{11}$$

Regarding the shocks, we assume that x^k_t is shared by all siblings of the same gender, can be correlated across siblings of different gender and is uncorrelated with the other variables (in particular with z_t). Finally u^k_t is an individual's white-noise error term.

We assume there is assortative mating both in years of schooling and in social competence. In particular we consider the linear projections of z^f_{t-1} and y^f_{t-1} on z^m_{t-1} and y^m_{t-1} respectively:

$$\begin{array}{rcl} z^f_{t-1} & = & r^m z^m_{t-1} + \omega^m_{t-1} \\ y^f_{t-1} & = & \tau^m y^m_{t-1} + \varepsilon^m_{t-1} \end{array}$$

where ω^m_{t-1} and ε^m_{t-1} might be correlated but are uncorrelated to z^m_{t-1} and y^m_{t-1}.

We use these matching functions to write years of schooling, y^k_t, and social status, z^k_t, as a function of father's years of schooling, y^m_{t-1}, and social status z^m_{t-1}. We write (11) as

$$z^k_t = G^k_m z^m_{t-1} + g^k_m \omega^m_{t-1} + e^k_t$$

where

$$\begin{array}{rcl} G^k_m & = & \gamma^k(\alpha^k_z + (1 - \alpha^k_z) r^m) \\ g^k_m & = & \gamma^m(1 - \alpha^k_z) \end{array}$$

and (10) as

$$y_t^k = B_m^k y_{t-1}^m + b_m^k \varepsilon_{t-1}^m + G_m^k z_{t-1}^m + g_m^k \omega_{t-1}^m + e_t^k + x_t^k + u_t^k$$

where

$$\begin{aligned} B_m^k &= \beta^k \left(\alpha_y^k + (1-\alpha_y^k)\tau^m \right) \\ b_m^k &= \beta^k (1-\alpha_y^k) \end{aligned}$$

All these expressions will be used to compute correlations between relatives that are related through their fathers. However, when we consider relatives that are related through their mothers, we need to consider y_{t-1}^k, as a function of mother's years of schooling, y_{t-1}^f, and social status z_{t-1}^f. We then also consider the linear projections

$$\begin{aligned} z_{t-1}^m &= r^f z_{t-1}^f + \omega_{t-1}^f \\ y_{t-1}^m &= \tau^f y_{t-1}^f + \varepsilon_{t-1}^f \end{aligned}$$

where ω_{t-1}^f and ε_{t-1}^f might be correlated but are uncorrelated to z_{t-1}^m and y_{t-1}^m, and

$$\begin{aligned} r^f &= r^m \frac{\sigma^2_{z^m_{t-1}}}{\sigma^2_{z^f_{t-1}}} \\ \tau^f &= \tau^m \frac{\sigma^2_{y^m_{t-1}}}{\sigma^2_{y^f_{t-1}}} \end{aligned}$$

We can then write (11) as

$$z_t^k = G_f^k z_{t-1}^f + g_f^k \omega_{t-1}^f + e_t^k$$

where

$$\begin{aligned} G_f^k &= \gamma^k (\alpha_z^k r^f + (1-\alpha_z^k)) \\ g_f^k &= \gamma^k \alpha_z^k \end{aligned}$$

and (10) as

$$y_t^k = B_f^k y_{t-1}^f + b_f^k \varepsilon_{t-1}^f + G_f^k z_{t-1}^f + g_f^k \omega_{t-1}^f + e_t^k + x_t^k + u_t^k$$

where

$$\begin{aligned} B_f^k &= \beta^k \left(\alpha_y^k \tau^f + (1-\alpha_y^k) \right) \\ b_f^k &= \beta^k \alpha_y^k \end{aligned}$$

We assume that the second order moments of all variables are time invariant. We present below the formulas for the covariances in years of scholing for relatives of different degrees of kinship. The correlations are computed by dividing the covariances by σ_m^2, σ_f^2 or $\sigma_m \sigma_f$ depending on the gender.

Covariances

Husband and wife

$$Cov_{h-w}(y^m_{t-1}, y^f_{t-1}) = Cov_{hw}(y^m_{t-1}, \tau^m y^m_{t-1} + \varepsilon^m_{t-1}) = \tau^m \sigma^2_{y^m}$$

Brothers

We first compute the covariances between y^k_{t-1} and z^k_{t-1} and z^j_t, $k, j = m, f$,

$$Cov(y^m_t, z^m_t) = B^m_m G^m_m Cov(y^m_{t-1}, z^m_{t-1}) + b^m_m g^m_m Cov(\varepsilon_{t-1}, \omega_{t-1}) + \sigma^2_{z^m}$$

and in the steady state we have $Cov(y^m_t, z^m_t) = Cov(y^m_{t-1}, z^m_{t-1})$, so that

$$Cov(y^m_{t-1}, z^m_{t-1}) = \frac{b^m_m g^m_m Cov(\varepsilon_{t-1}, \omega_{t-1}) + \sigma^2_{z^m}}{1 - B^m_m G^m_m}$$

and the covariance between brothers is

$$\begin{aligned} Cov_b(y^{m,i}_t, y^{m,j}_t) &= (B^m_m)^2 \sigma^2_{y^m} + 2B^m_m G^m Cov(y^m_{t-1}, z^m_{t-1}) + (b^m_m)^2 \sigma^2_{\varepsilon^m} \\ &\quad + (G^m_m)^2 \sigma^2_{z^m} + (g^m_m)^2 \sigma^2_{w^m} + 2b^m_m g^m_m Cov(\varepsilon^m_{t-1}, \omega^m_{t-1}) + \sigma^2_{x^m} \end{aligned}$$

Sisters

$$\begin{aligned} Cov_s(y^{f,i}_t, y^{f,j}_t) &= \left(B^f_m\right)^2 \sigma^2_{y^m} + 2B^f_m G^f Cov(y^m_{t-1}, z^m_{t-1}) + \left(b^f_m\right)^2 \sigma^2_{\varepsilon^m} \\ &\quad + \left(G^f_m\right)^2 \sigma^2_{z^m} + \left(g^f_m\right)^2 \sigma^2_{w^m} + 2b^f_m g^f_m Cov(\varepsilon^m_{t-1}, \omega^m_{t-1}) + \sigma^2_{x^f} \end{aligned}$$

Brother-sister

$$\begin{aligned} Cov_{b-s}(y^{m,i}_t, y^{f,j}_t) &= B^m_m B^f_m \sigma^2_{y^m} + B^m_m G^f_m Cov(y^m_{t-1}, z^m_{t-1}) + B^f_m G^m_m Cov(y^m_{t-1}, z^m_{t-1}) + b^m_m b^f_m \sigma^2_{\varepsilon^m} \\ &\quad + G^m_m G^f_m \sigma^2_{z^m} + g^m_m g^f_m \sigma^2_{w^m} + \left(b^f_m g^m_m + b^m_m g^f_m\right) Cov(\varepsilon^m_{t-1}, \omega^m_{t-1}) + \sigma_{x^m, x^f} \end{aligned}$$

Male cousins (fathers are brothers)

We first compute the following covariances for their fathers (who are brothers)

$$Cov_b(z^{m,i}_{t-1}, z^{m,j}_{t-1}) = (G^m_m)^2 \sigma^2_{z^m} + (g^m_m)^2 \sigma^2_{w^m}$$

and

$$\begin{aligned} Cov_b(y^{m,i}_{t-1}, z^{m,j}_{t-1}) &= B^m_m G^m_m Cov(y^m_{t-2}, z^m_{t-2}) + b^m_m g^m_m Cov(\varepsilon^m_{t-2}, w^m_{t-2}) \\ &\quad + (G^m_m)^2 \sigma^2_{z^m} + (g^m_m)^2 \sigma^2_{w^m} \end{aligned}$$

The covariance for male cousins whose fathers are brothers is

$$\begin{aligned} Cov_{mc\_fb}(y^{m,i}_t, y^{m,j}_t) &= (B^m_m)^2 Cov_b(y^{m,i}_{t-1}, y^{m,j}_{t-1}) + 2B^m_m G^m_m Cov_b(y^{m,i}_{t-1}, z^{m,j}_{t-1}) \\ &\quad + (G^m_m)^2 Cov_b(z^{m,i}_{t-1}, z^{m,j}_{t-1}) \end{aligned}$$

Male cousins (mothers are sisters)

We first compute the following covariances for their mothers (who are sisters)

$$Cov_s(z^{f,i}_{t-1}, z^{f,j}_{t-1}) = (G^f_m)^2 \sigma^2_{z^m} + (g^f_m)^2 \sigma^2_{w^m}$$

and

$$\begin{aligned} Cov_s(y_{t-1}^{f,i}, z_{t-1}^{f,j}) &= B_m^f G_m^f Cov(y_{t-2}^m, z_{t-2}^m) + b_m^f g_m^f Cov(\varepsilon_{t-2}^m, w_{t-2}^m) \\ &\quad + (G_m^f)^2 \sigma_{z^m}^2 + (g_m^f)^2 \sigma_{w^m}^2 \end{aligned}$$

The covariance for male cousins whose mothers are sisters is

$$\begin{aligned} Cov_{mc\_ms}(y_t^{m,i}, y_t^{m,j}) &= \left(B_f^m\right)^2 Cov_s(y_{t-1}^{f,i}, y_{t-1}^{f,j}) + 2B_f^m G_f^m Cov_s(y_{t-1}^{f,i}, z_{t-1}^{f,j}) \\ &\quad + \left(G_f^m\right)^2 Cov_s(z_{t-1}^{f,i}, z_{t-1}^{f,j}) \end{aligned}$$

Male cousins (father and mother are brother and sister)

We first compute the following covariances for their father and mother (who are brother and sister)

$$Cov_{b-s}(z_{t-1}^{m,i}, z_{t-1}^{f,j}) = G_m^m G_m^f \sigma_{z^m}^2 + g_m^m g_m^f \sigma_{w^m}^2$$

and

$$\begin{aligned} Cov_{b-s}(y_{t-1}^{m,i}, z_{t-1}^{f,j}) &= B_m^m G_m^f Cov(y_{t-2}^m, z_{t-2}^m) + b_m^m g_m^f Cov(\varepsilon_{t-2}^m, w_{t-2}^m) \\ &\quad + G_m^m G_m^f \sigma_{z^m}^2 + g_m^m g_m^f \sigma_{w^m}^2 \end{aligned}$$

$$\begin{aligned} Cov_{b-s}(y_{t-1}^{f,i}, z_{t-1}^{m,j}) &= B_m^f G_m^m Cov(y_{t-2}^m, z_{t-2}^m) + b_m^f g_m^m Cov(\varepsilon_{t-2}^m, w_{t-2}^m) \\ &\quad + G_m^m G_m^f \sigma_{z^m}^2 + g_m^m g_m^f \sigma_{w^m}^2 \end{aligned}$$

The covariance for male cousins whose father and mother are brother and sister is

$$\begin{aligned} Cov_{mc\_fb-ms}(y_t^{m,i}, y_t^{m,j}) &= B_m^m B_f^m Cov_{b-s}(y_{t-1}^{m,i}, y_{t-1}^{f,j}) + B_m^m G_f^m Cov_{b-s}(y_{t-1}^{m,i}, z_{t-1}^{f,j}) \\ &\quad + B_f^m G_m^m Cov_{b-s}(y_{t-1}^{f,i}, z_{t-1}^{m,j}) + G_m^m G_f^m Cov_{b-s}(z_{t-1}^{m,i}, z_{t-1}^{f,j}) \end{aligned}$$

Female cousins (fathers are brothers)

$$\begin{aligned} Cov_{fc\_fb}(y_t^{f,i}, y_t^{f,j}) &= (B_m^f)^2 Cov_b(y_{t-1}^{m,i}, y_{t-1}^{m,j}) + 2B_m^f G_m^f Cov_b(y_{t-1}^{m,i}, z_{t-1}^{m,j}) \\ &\quad + (G_m^f)^2 Cov_b(z_{t-1}^{m,i}, z_{t-1}^{m,j}) \end{aligned}$$

Female cousins (mothers are sisters)

$$\begin{aligned} Cov_{mc\_ms}(y_t^{f,i}, y_t^{f,j}) &= \left(B_f^f\right)^2 Cov_s(y_{t-1}^{f,i}, y_{t-1}^{f,j}) + 2B_f^f G_f^f Cov_s(y_{t-1}^{f,i}, z_{t-1}^{f,j}) \\ &\quad + \left(G_f^f\right)^2 Cov_s(z_{t-1}^{f,i}, z_{t-1}^{f,j}) \end{aligned}$$

Female cousins (father and mother are brother and sister)

$$\begin{aligned} Cov_{mc\_fb-ms}(y_t^{f,i}, y_t^{f,j}) &= B_m^f B_f^f Cov_{b-s}(y_{t-1}^{m,i}, y_{t-1}^{f,j}) + B_m^f G_f^f Cov_{b-s}(y_{t-1}^{m,i}, z_{t-1}^{f,j}) \\ &\quad + B_f^f G_m^f Cov_{b-s}(y_{t-1}^{f,i}, z_{t-1}^{m,j}) + G_m^f G_f^f Cov_{b-s}(z_{t-1}^{m,i}, z_{t-1}^{f,j}) \end{aligned}$$

Male-female cousins (fathers are brothers)

$$Cov_{m-fc\_fb}(y_t^{m,i}, y_t^{f,j}) = B_m^m B_m^f Cov_b(y_{t-1}^{m,i}, y_{t-1}^{m,j}) + \left(B_m^m G_m^f + B_m^f G_m^m\right) Cov_b(y_{t-1}^{m,i}, z_{t-1}^{m,j}) + G_m^m G_m^f Cov_b(z_{t-1}^{m,i}, z_{t-1}^{m,j})$$

Male-female cousins (mothers are sisters)

$$Cov_{m-fc\_ms}(y_t^{m,i}, y_t^{f,j}) = B_f^m B_f^f Cov_s(y_{t-1}^{f,i}, y_{t-1}^{f,j}) + \left(B_f^m G_f^f + B_f^f G_f^m\right) Cov_s(y_{t-1}^{f,i}, z_{t-1}^{f,j}) + G_f^m G_f^f Cov_s(z_{t-1}^{f,i}, z_{t-1}^{f,j})$$

Male-female cousins (father of the male is brother of the mother of the female)

$$Cov_{m-fc\_fb-ms}(y_t^{m,i}, y_t^{f,j}) = B_m^m B_f^f Cov_{b-s}(y_{t-1}^{m,i}, y_{t-1}^{f,j}) + B_m^m G_f^f Cov_{b-s}(y_{t-1}^{m,i}, z_{t-1}^{f,j}) + B_f^f G_m^m Cov_{b-s}(y_{t-1}^{f,i}, z_{t-1}^{m,j}) + G_m^m G_f^f Cov_{b-s}(z_{t-1}^{m,i}, z_{t-1}^{f,j})$$

Male-female cousins (mother of the male is sister of the father of the female)

$$Cov_{m-fc\_ms-fb}(y_t^{m,i}, y_t^{f,j}) = B_f^m B_m^f Cov_{b-s}(y_{t-1}^{m,i}, y_{t-1}^{f,j}) + B_m^f G_f^m Cov_{b-s}(y_{t-1}^{m,i}, z_{t-1}^{f,j}) + B_f^m G_m^f Cov_{b-s}(y_{t-1}^{f,i}, z_{t-1}^{m,j}) + G_f^m G_m^f Cov_{b-s}(z_{t-1}^{m,i}, z_{t-1}^{f,j})$$

Son-Father

$$Cov_{sf}(y_t^m, y_{t-1}^m) = B_m^m \sigma_{y^m}^2 + G_m^m Cov(y_{t-1}^m, z_{t-1}^m)$$

Son-Mother

$$\begin{aligned} Cov_{sm}(y_t^m, y_{t-1}^f) &= Cov(B_m^m y_{t-1}^m + b_m^m \varepsilon_{t-1}^m + G_m^m z_{t-1}^m + g_m^m \omega_{t-1}^m, \tau^m y_{t-1}^m + \varepsilon_{t-1}^m) \\ &= \tau^m B_m^m \sigma_{y^m}^2 + \tau^m G_m^m Cov(y_{t-1}^m, z_{t-1}^m) + b_m^m \sigma_{\varepsilon^m}^2 + g_m^m Cov(\omega_{t-1}^m, \varepsilon_{t-1}^m) \end{aligned}$$

Daughter-Father

$$Cov_{df}(y_t^f, y_{t-1}^m) = B_m^f \sigma_{y^m}^2 + G_m^f Cov(y_{t-1}^m, z_{t-1}^m)$$

Daughter-Mother

$$\begin{aligned} Cov_{dm}(y_t^f, y_{t-1}^f) &= Cov(B_m^f y_{t-1}^m + b_m^f \varepsilon_{t-1}^m + G_m^f z_{t-1}^m + g_m^f \omega_{t-1}^m, \tau^m y_{t-1}^m + \varepsilon_{t-1}^m) \\ &= \tau^m B_m^f \sigma_{y^m}^2 + \tau^m G_m^f Cov(y_{t-1}^m, z_{t-1}^m) + b_m^f \sigma_{\varepsilon^m}^2 + g_m^f Cov(\omega_{t-1}^m, \varepsilon_{t-1}^m) \end{aligned}$$

Nephew and uncle (brother of the father)

$$Cov_{neph-u\_bf}(y_t^{m,i}, y_{t-1}^{m,j}) = B_m^m Cov_b(y_{t-1}^{m,i}, y_{t-1}^{m,j}) + G_m^m Cov_b(y_{t-1}^{m,i}, z_{t-1}^{m,j})$$

Nephew and uncle (brother of the mother)

$$Cov_{neph-u\_bm}(y_t^{m,i}, y_{t-1}^{m,j}) = B_f^m Cov_{b-s}(y_{t-1}^{f,i}, y_{t-1}^{m,j}) + G_f^m Cov_{b-s}(y_{t-1}^{m,i}, z_{t-1}^{f,j})$$

Nephew and aunt (sister of the father)

$$Cov_{neph-au\_sf}(y_t^{m,i}, y_{t-1}^{f,j}) = B_m^m Cov_{b-s}(y_{t-1}^{m,i}, y_{t-1}^{f,j}) + G_m^m Cov_{b-s}(y_{t-1}^{f,i}, z_{t-1}^{m,j})$$

Nephew and aunt (sister of the mother)

$$Cov_{neph-au\_sm}(y_t^{m,i}, y_{t-1}^{f,j}) = B_f^m Cov_s(y_{t-1}^{f,i}, y_{t-1}^{f,j}) + G_f^m Cov_s(y_{t-1}^{f,i}, z_{t-1}^{f,j})$$

Nice and uncle (brother of the father)

$$Cov_{nice-u\_bf}(y_t^{f,i}, y_{t-1}^{m,j}) = B_m^f Cov_b(y_{t-1}^{m,i}, y_{t-1}^{m,j}) + G_m^f Cov_b(y_{t-1}^{m,i}, z_{t-1}^{m,j})$$

Nice and uncle (brother of the mother)

$$Cov_{nice-u\_bm}(y_t^{f,i}, y_{t-1}^{m,j}) = B_f^f Cov_{b-s}(y_{t-1}^{f,i}, y_{t-1}^{m,j}) + G_f^f Cov_{b-s}(y_{t-1}^{m,i}, z_{t-1}^{f,j})$$

Nice and aunt (sister of the father)

$$Cov_{nice-au\_sf}(y_t^{f,i}, y_{t-1}^{f,j}) = B_m^f Cov_{b-s}(y_{t-1}^{m,i}, y_{t-1}^{f,j}) + G_m^f Cov_{b-s}(y_{t-1}^{f,i}, z_{t-1}^{m,j})$$

Nice and aunt (sister of the mother)

$$Cov_{nice-au\_sm}(y_t^{f,i}, y_{t-1}^{f,j}) = B_f^f Cov_s(y_{t-1}^{f,i}, y_{t-1}^{f,j}) + G_f^f Cov_{b-s}(y_{t-1}^{f,i}, z_{t-1}^{f,j})$$

Variance decomposition

We have that

$$y_t^k = \beta^k \widetilde{y}_{t-1}^k + z_t^k + x_t^k + u_t^k$$

Then

$$\sigma^2_{y^k} = (\beta^k)^2 \sigma^2_{\widetilde{y}^k} + \sigma^2_{z^k} + \beta^k Cov(\widetilde{y}_{t-1}^k, z_t^k) + \sigma^2_{x^k} + \sigma^2_{u^k}$$

- $\sigma^2_{u^k}$ is obtained as a residual.
- $\beta^k, \sigma^2_{x^k}$ and $\sigma^2_{z^k}$ are directly estimated.
- $\sigma^2_{\widetilde{y}^k}$

  $$\sigma^2_{\widetilde{y}^k_{t-1}} = \left(\alpha_y^k\right)^2 \sigma^2_{y^m} + \left(1-\alpha_y^k\right)^2 \sigma^2_{y^f} + \alpha_y^k(1-\alpha_y^k)\tau^m \sigma^2_{y^m}$$

  and we use the estimates of α_y^k and τ^m, and the empirical values for $\sigma^2_{y^m}$ and $\sigma^2_{y^f}$
- $Cov(\widetilde{y}_{t-1}^k, z_t^k)$

  $$Cov(\widetilde{y}_{t-1}^k, z_t^k) = G_m^k \left[\alpha_y^k + (1-\alpha_y^k)\tau^m\right] Cov(y_{t-1}^m, z_{t-1}^m) + g_m^k(1-\alpha_y^k) Cov(\varepsilon_{t-1}^m, w_{t-1}^m)$$

  and we use the estimates of $\alpha_y^k, G_m^k, g_m^k, Cov(\varepsilon_t^m, w_t^m)$ and $Cov(y_t^m, z_t^m)$.

Appendix C

Let p_s be the probability that a surname picked up at random is of size s in a given generation[17] and let $p_{ns}(g)$ be the probability that a surname of size s in a given generation becomes of size n in the next generation, where g denotes the population growth rate.[18] Using Bayes rule we can compute the probability $q_{sn}(g)$ that a surname of size n in generation t was of size s in generation $t-1$.

$$q_{sn}(g) = \frac{p_{ns}(g)p_s}{\sum_{j=1}^{\infty} p_{nj}(g)p_j}$$

Then, using this conditional probabilities, we can compute the probability that two persons with the same surname in generation t are brothers, first-cousins, second-cousins, etc., depending on the surname size n and the population growth rate of g. Let $r_1(g,n)$ be the probability of being brothers, $r_2(g,n)$ the probability of being first-cousins,.., $r_k(g,n)$ the probability of being $(k+1)$th-cousins, etc. We have the following recursive formulas:

$$\begin{aligned}
r_1(g,n) &= \sum_{s=1}^{\infty} q_{sn}(g)\frac{1}{s} \\
r_2(g,n) &= \sum_{s=2}^{\infty} q_{sn}(g) r_1(g,s) \\
&\vdots \\
r_k(g,n) &= \sum_{s=2}^{\infty} q_{sn}(g) r_{k-1}(g,s) \\
&\vdots
\end{aligned}$$

[17] We assume that for all generations the distribution of surname sizes follow a p-law distribution. In our case we estimate the power parameter is 1.6

[18] In the empirical part of the paper we assume that for all generations the number of male descendants follow a Poisson distribution with population growth rate g.