Estimating occupational mobility with covariates

Jørgen Modalsli[*]

May 11, 2015

Abstract

The Altham statistic is often used to calculate intergenerational associations in occupations in studies of historical social mobility. This paper presents a method to incorporate individual covariates into such estimates of social mobility, and to construct corresponding confidence intervals. The method is applied to an intergenerational sample of Norwegian data, showing that estimates of intergenerational mobility are robust to the inclusion of controls for father's and son's age.

Keywords: Intergenerational occupational mobility, Altham statistic
JEL codes: J62, N34, C46

1 Introduction

The Altham statistic (Altham, 1970; Altham & Ferrie, 2007) sees increasing use as an indicator of intergenerational occupational mobility in the historical economics literature (Long & Ferrie, 2007, 2013; Boberg-Fazlic & Sharp, 2013; Azam, 2013; Ferrie, 2005; Long, 2013). In such historical studies, based on census records or family reconstitution data, income data is usually not available, while data on occupation or social class does exist.[1]

The statistic is constructed from matrices tabulating fathers' and sons' occupations, using two-way odds ratios. Following the literature, we index father's occupations by i and l, and son's occupations by j and m, and let p_{ij} denote the probability of a child obtaining an occupation j given father's occupation i. The two-way odds ratio Θ_{ijlm} then compares the probabilities of two sons' occupations, given two fathers' occupations:

$$\Theta_{ijlm} = \log\left(\frac{p_{ij}/p_{im}}{p_{lj}/p_{lm}}\right) \tag{1}$$

[*] Statistics Norway, Research Department. I am grateful to Manudeep Bhuller, Terje Skjerpen and an anonymous referee for helpful comments. Support from the Norwegian Research Council is acknowledged.

[1] The Altham statistic does not depend on an unambiguous ordering of occupations or social classes. In situations where such an ordering is available, other tools can be used that take advantage of this additional information.

The Altham statistic $d(P, J)$ is defined as the square root of the sum of the squared deviations of two-way odds ratios from a hypothetical "full mobility" setting where said odds ratios are zero:[2]

$$d(P, J) = \left(\sum_{i=1}^{N} \sum_{j=1}^{N} \sum_{l=1}^{N} \sum_{m=1}^{N} [\Theta_{ijlm}]^2 \right)^{1/2} \tag{2}$$

where N refers to the number of occupation categories.[3]

In studies of intergenerational mobility using income data such as Solon (1992), a common approach is to regress son's (log) income on father's (log) income to obtain an estimate of the intergenerational association parameter. Further information on individuals can be incorporated in the regression to study intergenerational mobility "net of" covariates, for example to account for an age profile in income.[4]

The purpose of this note is to similarly extend the calculation of the Altham statistic $d(P, J)$ to adjust for covariates, giving a summary statistic of mobility net of these characteristics. This method will be applied to occupational mobility between fathers and sons in Norway between 1960 and 1980, controlling for the age composition in different occupation categories.

2 Modelling mobility

To study occupational choice with control variables, we use the canonical multinomial logit model (see Agresti (2002, p. 268) for a general description of multinomial logit models) where the child's occupation is the outcome. We consider a set of N occupations and set the first as the reference outcome. We denote occupation by o, let superscript f denote parent and s child, index individuals by q, and estimate a system of $N - 1$ equations for son's occupation, indexed by k:

$$\log \left(\frac{Pr(o_q^s = k)}{Pr(o_q^s = 1)} \right) = \alpha_k + \boldsymbol{\beta}_k' \boldsymbol{D}_q + \boldsymbol{\gamma}_k' \boldsymbol{X}_q \qquad k = 2, 3, ..., N \tag{3}$$

where $\mathbf{D}_q = \{D_{2,q}, D_{3,q}, ..., D_{N,q}\}$ is a vector of dummy variables where $D_{z,q} = 1$ if father's

[2] The above papers use the statistic in two settings: for comparison of two different mobility matrices $d(P, Q)$ and for comparing a mobility matrix to a hypothetical matrix of full mobility $d(P, J)$. For brevity, this article only deals with the second setting.

[3] For simplicity, an equal number of father's and son's occupations are considered throughout this paper, though the result is generalizable to the case where these are different.

[4] Formally, for father's log income y^f and son's log income y^s, indexing individuals by q, we have

$$y_q^s = \alpha + \beta^{\text{OLS}} y_q^f + \boldsymbol{\gamma}' \boldsymbol{X}_q + \epsilon_i$$

As explained by Solon (1992), an estimator of social mobility based on β^{OLS} has inherent biases and methods using instrumental variables or average income over several years should be preferred. However, in studies of historical data such methods are frequently not feasible because of data limitations. Moreover, occupations are more stable over the life cycle than income.

occupation is z and $D_{z,q} = 0$ otherwise. $\boldsymbol{\beta}_k$ and $\boldsymbol{\gamma}_k$ are parameter vectors; we use β_k^i to refer to the ith element of $\boldsymbol{\beta_k}$. For the sake of completeness, we also define parameters for the reference group, with α_1 set to zero and $\boldsymbol{\beta}_1'$ and $\boldsymbol{\gamma}_1'$ as vectors of zeros. Estimated probability ratios do not depend on the choice of reference category.

From Equation (3), we have, for the example of comparing the probability of a son getting occupations 3 vs. 4, given that the father holds occupation 2 and the son is 30 years old, with a dummy variable specification for son's age:

$$\log\left(\frac{Pr(o_q^s = 3|o_q^f = 2)}{Pr(o_q^s = 4|o_q^f = 2)}\right) = (\alpha_3 - \alpha_4) + (\beta_3^2 - \beta_4^2) + (\gamma_3^{30} - \gamma_4^{30}) \tag{4}$$

When there are no control variables $\boldsymbol{X}_q$, it can be shown that the estimation procedure yields the raw probabilities $(\widehat{Pr(o_q^s = j|o_q^f} = i) = p_{ij})$.[5]

The setup of the multinomial logit model in Equation (3) makes the estimated odds ratios invariant across subgroups defined by control variables. To see this, insert for the probabilities in (1) from (4) to get

$$\Theta_{ijlm} = (\beta_j^i - \beta_m^i) - (\beta_j^l - \beta_m^l) \tag{5}$$

For any set of covariates $\boldsymbol{X_q}$, including the empty one, the expression for $d(P, J)$ as expressed by parameters estimated with multinomial logit as in Equation (3) remains

$$\widehat{d(P,J)} = \left(\sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{l=1}^{N}\sum_{m=1}^{N}\left[(\hat{\beta}_j^i - \hat{\beta}_m^i) - (\hat{\beta}_j^l - \hat{\beta}_m^l)\right]^2\right)^{1/2} \tag{6}$$

Equation (6) can then be used as a basis for calculating intergenerational occupational mobility while controlling for age structure or other covariates. Further, the parameters $\boldsymbol{\gamma}_k'$ give information on the relationship between covariates and occupation outcomes.

Using the standard errors of the estimated coefficients from Equation (3), we can also construct confidence intervals for the estimates of the probabilities as well as the estimate of overall mobility.[6]

[5] See Appendix for proof.

[6] The confidence intervals are constructed using a bootstrap technique based on the covariance matrix from the logit estimation. See the Appendix for details.

3 Application

As an application, the methodology is used on an intergenerational transition matrix constructed from the Norwegian censuses of 1960 and 1980. Occupations are coded into four categories similar to those used by Long & Ferrie (2013); see Modalsli (2015) for further details. We restrict the sample to the native-born male population between 30 and 60 years of age in 1980, for which the father's identity is known and the father is between 30 and 60 years old in 1960, and use occupations reported in 1960 for fathers and 1980 for sons. The total sample population is 201,289 individuals, and the aggregate transition matrix is shown in Table 1.

Father's	Son's occupation				
occupation	W	F	S	U	Total
White collar (W)	32,005	476	10,448	1,117	44,046
	72.7%	*1.1%*	*23.7%*	*2.5%*	
Farmer (F)	11,215	9,878	17,484	2,588	41,165
	27.2%	*24.0%*	*42.5%*	*6.3%*	
Manual, skilled (S)	37,178	898	51,426	3,776	93,278
	39.9%	*1.0%*	*55.1%*	*4.0%*	
Manual, unskilled (U)	6,391	527	11,664	4,218	22,800
	28.0%	*2.3%*	*51.2%*	*18.5%*	
Total	86,789	11,779	91,022	11,699	201,289

Table 1: Father-son occupation transition matrix (cell count and row percentage), Norway, 1960-1980

Included in X	$d(P, J)$	Interval
No controls (reference)	22.3	(22.1 - 22.6)
Son's age (dummy variable)	22.0	(21.8 - 22.3)
Father's age (dummy variable)	21.9	(21.6 - 22.2)
Father's and son's age (dummy variables)	21.9	(21.7 - 22.2)
Father's and son's age (linear)	22.0	(21.7 - 22.2)
Father's and son's age (quadratic)	21.9	(21.6 - 22.2)

Table 2: Estimates of intergenerational occupational mobility (Norway 1960-1980) when controlling for age composition

The Altham statistic calculated from Table 1 using (2) is 22.3. We proceed to calculate the Altham statistic using covariates for father's and son's age using (3) and (6); the results are reported in Table 2. It is evident that the change in the Altham statistic from inclusion of age controls is only moderate, and that all 95% confidence intervals overlap.

While the estimate of intergenerational mobility in society as a whole does not change much when age controls are included, there can be substantial age variation in specific transition probabilities. This is illustrated in Figure 1, where transition probabilities are estimated with a model using dummies for father's occupation and son's age (the second line in Table 2) using the parameters obtained from Equation (3). The confidence bands are constructed using the

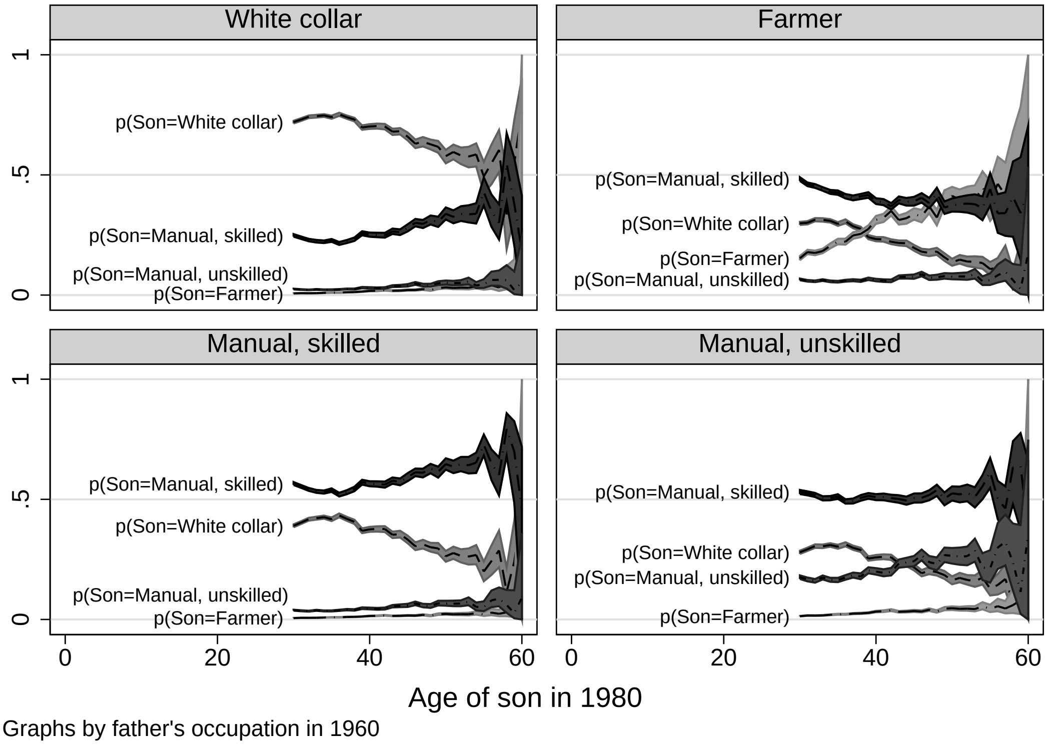

Figure 1: Predicted probability of son's occupation in 1980 (using (3)), given son's age in 1980 and father's occupation in 1960. 95% confidence intervals.

same method as for the intervals on the Altham statistic. It is evident from the figure that some occupations experience age variation in probabilities. Notably, older sons are more likely to be farmers, while younger sons are more likely to be white-collar workers. Because of missing family information for older cohorts, there are fewer individuals in the upper end of the age range, leading to less precise estimates for these ages.

The approach used here imposes some restrictions on the covariates. While a given age dummy can affect the probabilities of sons' occupations separately, the interaction with father's occupation only happens through a multiplicative (log-additive) interaction with the relevant β parameter.

4 Concluding comments

This paper has illustrated an approach for incorporating individual covariates into the Altham statistic commonly used to examine intergenerational mobility in historical samples. In the application used here, estimated mobility changes only moderately when controls are included, though specific transition probabilities are heterogeneous across age groups.

The approach can be extended to control for other types of covariates, such as the effect of regional characteristics or the occupation of other family members on occupational outcomes. As the availability of large historical data sets increases, there is likely to be further scope for the inclusion of covariates in historical analyses of intergenerational mobility.

A Appendix

Closed form of multinomial logit with no covariates

This section shows that the maximum likelihood estimate of $d(P, J)$ obtained using the multinomial logit model is equal to the expression given in Equations (1-2)

From Agresti (2002, p. 273), with population shares of son's occupation denoted π, individuals indexed by q and total population size Q:

$$\mathcal{L} = \log \prod_{q=1}^{Q} \left(\prod_{k=1}^{K} \pi_k(\boldsymbol{x}_q)^{y_{qk}} \right) = \sum_{q=1}^{Q} \left(\sum_{k=1}^{K} y_{qk}(\alpha_k + \boldsymbol{\beta}_{\boldsymbol{k}}^{'} \boldsymbol{x}_{\boldsymbol{q}}) - \log \left[1 + \sum_{k=1}^{K} \exp(\alpha_k + \boldsymbol{\beta}_{\boldsymbol{k}}^{'} \boldsymbol{x}_{\boldsymbol{q}}) \right] \right) \tag{7}$$

Here k indexes equation, that is, son's occupation, while $\boldsymbol{x}$ indexes individual covariates, that is, father's occupation. We can interpret $\mathcal{L}$ as the weighted sum of the contributions of all N^2 cells in the transition matrix

$$\mathcal{L} = \sum_{i=1}^{N}\sum_{j=1}^{N} Q_{ij} L_{ij} \tag{8}$$

Consider an individual where the father has occupation i and the son has occupation j. In this case $y_{qj} = 1$ and all other y_q's are zero. Moreover, the vector $\boldsymbol{\beta_k' x_q}$ becomes β_q^i. For this individual we then have the contribution term

$$L_{ij} = (\alpha_j + \beta_j^i) - \log\left[1 + \sum_{k=1}^{K} \exp(\alpha_k + \beta_k^i)\right] \tag{9}$$

Summing all the contributions, we get

$$\mathcal{L} = \sum_{i=1}^{N}\sum_{j=1}^{N} \big(Q_{ij}(\alpha_j + \beta_j^i)\big) - \sum_{i=1}^{N}\left(M_i \log\left[1 + \sum_{k=1}^{K} \exp(\alpha_k + \beta_k^i)\right]\right) \tag{10}$$

where $M_i = \sum_{j=1}^{N} Q_{ij}$.

Maximizing $\mathcal{L}$ with respect to $N + N^2$ parameters (the αs and βs, respectively) and reordering the first order conditions gives the system

$$\sum_{z=1}^{N} \frac{M_z \exp(\hat{\alpha}_j + \hat{\beta}_j^z)}{1 + \sum_{k=1}^{K} \exp(\hat{\alpha}_k + \hat{\beta}_k^z)} = \sum_{q=1}^{N} Q_{qj} \tag{11}$$

$$\frac{\exp(\hat{\alpha}_j + \hat{\beta}_j^i)}{1 + \sum_{k=1}^{K} \exp(\hat{\alpha}_k + \hat{\beta}_k^z)} = \frac{Q_{ij}}{\sum_{w=1}^{N} Q_{iw}} \tag{12}$$

To identify the system, we set $\hat{\alpha}_1 = 0$, $\hat{\beta}_j^1 = 0$ for all j, $\hat{\beta}_1^i = 0$ for all i. A total of 10 restrictions gives 20 free parameters to identify. We then insert from (12,$j = 1$) into (12,$j > 1$) to identify the remaining α parameters. Further insertion gets the expression for the βs and we obtain (for $j > 1$ and $i > 1$):

$$\hat{\alpha}_j = \log(N_{1j}/N_{11}) \tag{13}$$

$$\hat{\beta}_j^i = \log(N_{ij}/N_{i1}) - \log(N_{1j}/N_{11}) = \log\left(\frac{N_{ij}/N_{i1}}{N_{1j}/N_{11}}\right) \tag{14}$$

Inserting for the predicted probabilities in the multinomial logit model gives

$$
\begin{aligned}
Pr(\widehat{o_q^s = j | o_q^f} = i) &= \frac{\exp(\hat{\alpha}_j + \hat{\beta}_j^i)}{1 + \sum_{k=1}^{K} \exp(\hat{\alpha}_k + \hat{\beta}_k^i)} \qquad (15) \\
&= \frac{N_{ij}}{\sum_{w=1}^{N} Q_{iw}} \qquad (16)
\end{aligned}
$$

which are the empirical probabilities of son's occupation given father's occupation.

Confidence intervals

To estimate confidence intervals for $\widehat{d(P,J)}$, a parametric bootstrapping technique is used. The coefficient estimates and covariance matrix from Stata's `mlogit` command is used to draw parameter values 1000 times using Stata's `drawnorm` command. The values of the set of βs for each of the 1000 iterations are then used to calculate an Altham statistic. These are sorted increasingly, and the 25th and 975th values are used as an upper and lower bound for the 95% confidence intervals presented in Table 2.

The confidence bands in Figure 1 are constructed in a similar way, using the values from the draws of the $\boldsymbol{\gamma}$ parameters.

It should be noted that the significance tests in Altham & Ferrie (2007) and Long & Ferrie (2013) use a chi-square test based on a generalized linear model, as described in Agresti (2002, chapter 4.5), a different approach than the one used here to obtain confidence intervals.

References

Agresti, Alan. 2002. *Categorical Data Analyis.* Wiley Interscience.

Altham, Patricia M E. 1970. The Measurement of Association of Rows and Columns for an $r \times s$ Contingency Table. *Journal of the Royal Statistical Society Series B*, **32**(1), 63–73.

Altham, Patricia M E, & Ferrie, Joseph P. 2007. Comparing Contingency Tables: Tools for Analyzing Data from Two Groups Cross-Classified by Two Characteristics. *Historical Methods*, **40**(1), 3–16.

Azam, Mehtabul. 2013. Intergenerational Occupational Mobility in India. *IZA Discussion Paper*, **7608**.

Boberg-Fazlic, Nina, & Sharp, Paul. 2013. North and South: Social Mobility and Welfare Spending in Preindustrial England. *EHES Working Paper*, **37**(Apr.).

Ferrie, Joseph P. 2005. The End of American Exceptionalism ? Mobility in the United States Since 1850. *Journal of Economic Perspectives*, **19**(3), 199–215.

Long, Jason. 2013. The surprising social mobility of Victorian Britain. *European Review of Economic History*, **17**, 1–23.

Long, Jason, & Ferrie, Joseph. 2007. The path to convergence: Intergenerational occupational mobility in Britain and the US in three eras. *Economic Journal*, **117**, 61–71.

Long, Jason, & Ferrie, Joseph. 2013. Intergenerational Occupational Mobility in Great Britain and the United States since 1850. *American Economic Review*, **103**(4), 1109–1137.

Modalsli, Jørgen. 2015. Intergenerational mobility in Norway, 1865-2011. *Statistics Norway Discussion Paper*, **798**.

Solon, Gary. 1992. Intergenerational income mobility in the United States. *The American Economic Review*, **82**(3), 393–408.