Is It Origin, Destination or Mobility?

A Monte Carlo Simulation of the Diagonal Reference Model

Alessandro Procopio[1] Robin Samuel[2]

[1]alessandro.procopio@uni.lu [2]robin.samuel@uni.lu

Cambridge Social Stratification Research Seminar

27/08/2020

UNIVERSITY OF LUXEMBOURG
Department of Social Sciences

Introduction

In social stratification and mobility literature, there is a deeply rooted track of research on the consequences of social mobility, e.g., concerning attitudes, behaviors and psycho-social conditions (Friedman, 2014).

However, applied researchers faced challenges on corroborating the theory, due to identification problem (linear dependence of covariates) which affect estimates specific to origin, destination and mobility.

To overcome the identification problem, Sobel (1981, 1985) proposed the *Diagonal Reference Model* (DRM).

The DRM is considered the first-choice statistical methodology among sociologists and demographers to investigate consequences of social mobility (Van der Waal et al., 2017; Billingsley et al., 2018).

Although the number of fields in which the model is gaining interest is growing, use of the DRM produced a body of null or weak evidences concerning mobility effects $\rightarrow$ stark contrast with expectations derived from theory.

This suggests that also the DRM may still suffer of the identification problem.

Aims of the Study

Gap

In methodological literature, there is a lack of research on the model behavior when applied to real world data.

Aim

We address this gap through Monte Carlo Simulation (MCS) technique to assess potential benefits and limitations of the model. Specifically:

1. Evaluate the degree of estimation bias;
2. Evaluate the capability of the model to detect effects existing on the population.

The root of the Identification Problem

The most intuitive approach to empirically assess effects of social Origin (O), Destination (D) and Mobility (M) is to run ANOVA regression (Blalock 1966, 1967; Duncan 1966; Mason et al. 1973) of the form:

$$Y = \mu + \alpha_i O + \beta_j D + \gamma_k M + \epsilon_{ij}$$

μ is the grand mean of Y

α_i is the effect of the i^{th} Origin class

β_j is the effect of being currently in the j^{th} Destination class

γ_k is the effect of mobility.

One Model, Two Problems

We can rewrite the formula in matrix form as $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{b}$, where the dependent variable is $\boldsymbol{y}_{n\times 1} = (y_1, y_2, y_3, \ldots y_n)^T$, the regressor matrix $\boldsymbol{X}_{n\times p}$ and the matrix of coefficients as $\boldsymbol{b}_{p\times 1} = (\alpha, \beta, \gamma)^T$.

We encounter two problems in this set: one easy, one more complex.

The easy problem concerns the *overparameterization* of the model, which can be solved by constraining $\sum_{i=1}^{I}\alpha = \sum_{j=1}^{J}\beta = \sum_{k=1}^{K}\gamma = 0$.

Blue Monday of Estimators

Because of linear dependency between O, D and M, the rank (the number of linearly independent rows/columns) of $\boldsymbol{X}_{n \times p}$ is less than p.

The Best Linear Unbiased Estimator (BLUE) is $\boldsymbol{b}_{OLS} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$, but if $\boldsymbol{X}$ is *rank deficient* (and in our case it is), than we cannot find the inverse of the square matrix

$$\boldsymbol{X}^T\boldsymbol{X} = \begin{bmatrix} n & \sum_{o=1}^{n} x_o \\ \sum_{o=1}^{n} x_o & \sum_{o=1}^{n} x_o^2 \end{bmatrix}$$

In this case the square matrix $\boldsymbol{X}^T\boldsymbol{X}$ is said to be *singular*.
We don't have any unique estimator.

The DRM Approach

In ANOVA regression, we decompose y_{ijk} in two additive effects: α_i for origin and β_j for destination.

In DRM, y_{ijk} can be decomposed in μ_{ii}, the population means specific to the i^{th} Origin category and μ_{jj}, the population means specific to the j^{th} Destination category. For immobile individuals - diagonals of the mobility table - $\mu_{ij} = \mu_{ii} = \mu_{jj}$.

DRM is a **parametrically weighted** regression model (Yamaguchi, 2002) as the means μ_{ii} and μ_{jj} are weighted by ρ and $(1 - \rho)$. These quantify relative salience of Origin and Destination on off-diagonal cells mean values μ_{ij}.

Putting All Together

$$\hat{\mu}_{ij} = \rho\mu_{ii} + (1-\rho)\mu_{jj} + \sum_{w=1}^{W} \gamma_w M_{ijw} + \epsilon_{ijk}$$

$$\rho = \frac{e^{\delta_i}}{e^{\delta_i} + e^{\delta_j}}$$

$$(1-\rho) = \frac{e^{\delta_j}}{e^{\delta_i} + e^{\delta_j}}$$

μ_{ii} = mean at the i^{th} origin category;

μ_{jj} = mean at the j^{th} destination class;

M_{ijw} = Mobility variable(s) with effect magnitude γ_w;

ρ and $(1-\rho)$ = relative salience of origin to the current destination;

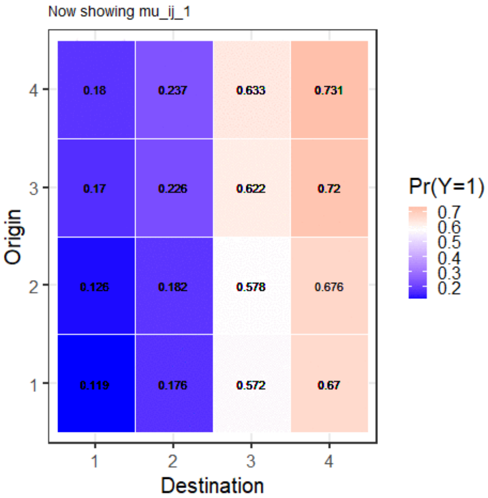
δ_i and δ_j = parameters to be estimated.

The Logic of the DRM

	Destination				
Origin	I	II	III	IV	all
I	μ_1	$\rho\mu_1 + r\mu_2$	$\rho\mu_1 + r\mu_3$	$\rho\mu_1 + r\mu_4$	$\rho\mu_1$
II	$\rho\mu_2 + r\mu_1$	μ_2	$\rho\mu_2 + r\mu_3$	$\rho\mu_2 + r\mu_4$	$\rho\mu_2$
III	$\rho\mu_3 + r\mu_1$	$\rho\mu_3 + r\mu_2$	μ_3	$\rho\mu_3 + r\mu_4$	$\rho\mu_3$
IV	$\rho\mu_4 + r\mu_1$	$\rho\mu_4 + r\mu_2$	$\rho\mu_4 + r\mu_3$	μ_4	$\rho\mu_4$
all	$r\mu_1$	$r\mu_2$	$r\mu_3$	$r\mu_4$	μ

$r = 1 - \rho$

Simulating Weighted Effects



MLE and NLS of the Model

Maximum Likelihood Estimation:

$$\mathcal{L}_{(\mu,\rho,\gamma,\sigma)} = \prod_{ijk} \Bigg((2\pi\sigma^2)^{-\frac{1}{2}} \exp\Bigg\{ -(2\sigma^2)^{-1} \times$$

$$\left(y_{ijk} - \sum_{i=1}^{I} \rho\mu_{\mathrm{ii}} \mathrm{X}_{.i} - \sum_{j=1}^{J} (1-\rho)\mu_{\mathrm{jj}} \mathrm{X}_{.j} - \sum_{w=1}^{W} \gamma_w \mathrm{X}_{ijw} \right)^2 \Bigg\} \Bigg)$$

Nonlinear Least Squares:

$$\operatorname*{arg\,min}_{f(\rho,\mu,\gamma,\sigma)} = \sum_{ijk} \left(y_{ijk} - \sum_{i=1}^{I} \rho\mu_{ii} X_{i.} - \sum_{j=1}^{J} (1-\rho)\mu_{jj} X_{.j} - \sum_{w=1}^{W} \gamma_w \mathrm{X}_{ijw} \right)^2$$

Our Experimental Design

Scenarios

We have constructed two scenarios:

Identity function: $\boldsymbol{X}\beta = \mu$ where $g(\mu) = \boldsymbol{X}\beta$

Logit function: $\boldsymbol{X}\beta = ln\left(\frac{\mu}{1-\mu}\right)$ where $g(\mu) = \frac{exp(\boldsymbol{X}\beta)}{1+exp(\boldsymbol{X}\beta)}$

Mobility Table Generation and Computation

Demirhan (2016) `rTableICC` package to generate random contingency table [1].

`gnm` package for computation of the model (Turner and Firth, 2015).

[1]Product Multinomial sampling: $p(y) = \frac{n!}{y!(n-y)!}\pi y(1-\pi)^{n-y}$ (Agresti, 2018)

The True Model

Continuous D.V.

$$y_{ijk} = \rho\mu_{ii} + r\mu_{jj} + \gamma_{up} Upward + \gamma_{down} Downward$$

Logit D.V.

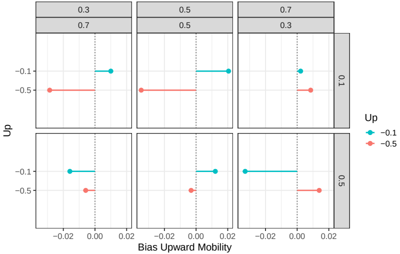
$$\pi_{ij} = \frac{exp(\rho\mu_{ii} + r\mu_{jj} + \gamma_{up} Upward + \gamma_{down} Downward)}{1 + exp(\rho\mu_{ii} + r\mu_{jj} + \gamma_{up} Upward + \gamma_{down} Downward)}$$

Population True Values

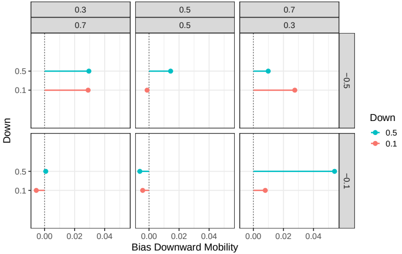
ρ	r	γ_{Up}	γ_{Down}
0.70	0.30	{-0.1, -0.5}	{0.1, 0.5}
0.50	0.50	{-0.1, -0.5}	{0.1, 0.5}
0.30	0.70	{-0.1, -0.5}	{0.1, 0.5}

Continuous Dependent Variable

Upward Mobility

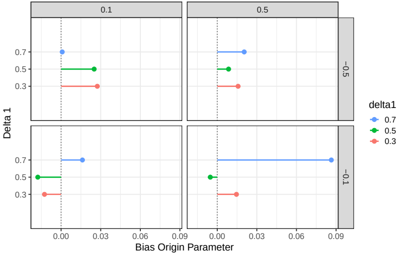


Downward Mobility

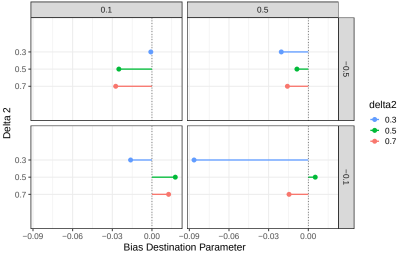


Weighting Parameters

e^{δ_i}

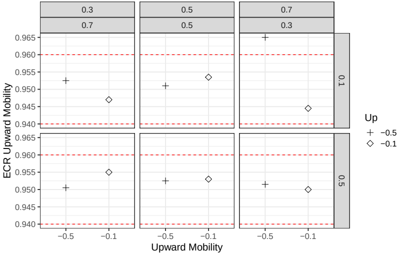


e^{δ_j}

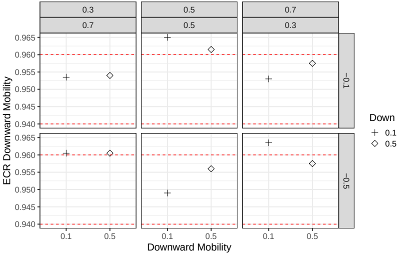


Empirical Coverage Rates

Upward Mobility

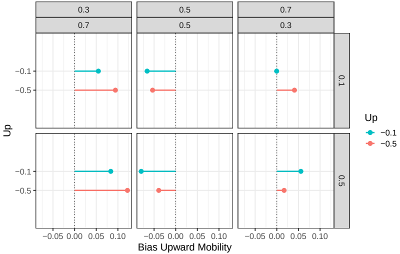


Downward Mobility

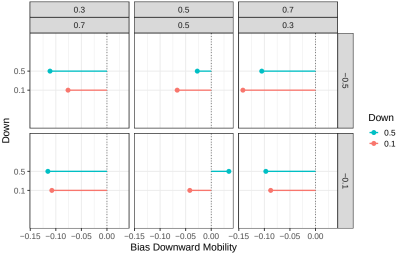


Logit

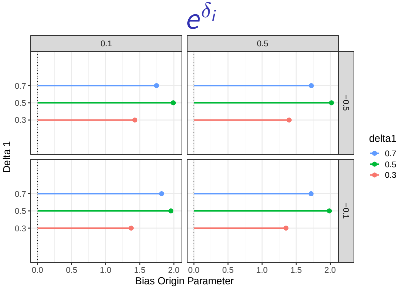
Upward Mobility

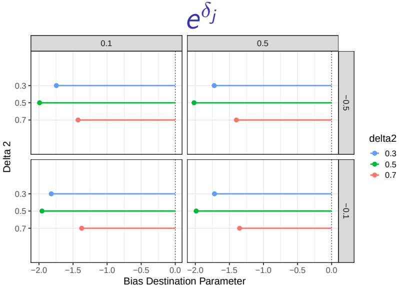


Downward Mobility



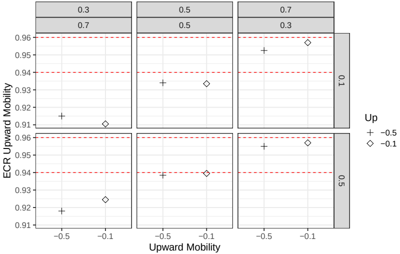
Weighting Parameters



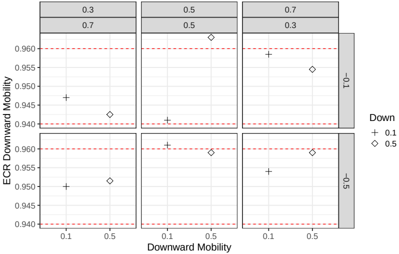


Empirical Coverage Rates

Upward Mobility



Downward Mobility



Discussion

Bias

In both scenarios, findings suggest mostly unbiased estimators.

Although greater bias in the non-linear logistic set.

Severe bias (still to clarify) affecting weighting parameters in logistic set.

ECR

ECR in linear scenario within the accpetable boundaries (except one case).

ECR in logistic scenario show under-coverage for upward mobility estimates.

When salience paremeters are equivalent, DRM tends to suffer from over-coverage.

Limits of this Study

Possible improvements are:

- Include a full-range mobility variable.
- Scenarios might include unobserved heterogeneity to simulate a more realistic dataset.
- Test how the model behaves when applied to longitudinal data.
- Inclusion of worked examples with real-world data.

Conclusions

What is good

DRM works better when the D.V. is continuous.

What is bad

Still methodological concerns when the simulation turns to non-linear logistic. Here identification problem can still "muddy the waters".

A last note

We should pay attention to the generalization of the results: as this is an experimentation, results might be specific to this set!

THANK YOU FOR YOUR TIME!

Bibliography

Agresti, A. (2018). *An introduction to categorical data analysis.* John Wiley & Sons.

Billingsley, S., Drefahl, S., and Ghilagaber, G. (2018). An application of diagonal reference models and time-varying covariates in social mobility research on mortality and fertility. *Social Science Research*, 75:73–82.

Blalock, H. M. (1966). The Identification Problem and Theory Building: The Case of Status Inconsistency. *American Sociological Review*, 31(1):52–61.

Blalock, H. M. (1967). Status inconsistency, social mobility, status integration and structural effects. *American Sociological Review*, pages 790–801.

Demirhan, H. (2016). rTableICC: An R Package for Random Generation of 22K and RC Contingency Tables. *The R Journal*, 8(1):48–63.

Duncan, O. D. (1966). Methodological issues in the analysis of social mobility. In *Social structure and mobility in economic development*, pages 51–97. Chicago, Aldine Publishing Company.

Friedman, S. (2014). The Price of the Ticket: Rethinking the Experience of Social Mobility. *Sociology*, 48(2):352–368.

Mason, K. O., Mason, W. M., Winsborough, H. H., and Poole, W. K. (1973). Some methodological issues in cohort analysis of archival data. *American sociological review*, pages 242–258.

Sobel, M. E. (1981). Diagonal Mobility Models : A Substantively Motivated Class of Designs for the Analysis of Mobility Effects. *American Sociological Review*, 46(6):893–906.

Sobel, M. E. (1985). Social Mobility and Fertility Revisited : Some New Models for the Analysis of the Mobility Effects Hypothesis. *American Sociological Review*, 50(5):699–712.

Turner, H. and Firth, D. (2015). Generalized nonlinear models in R: An overview of the gnm package. *R Journal*, pages 1–61.

Van der Waal, J., Daenekindt, S., and de Koster, W. (2017). Statistical challenges in modelling the health consequences of social mobility: the need for diagonal reference models. *International Journal of Public Health*, 62(9):1029–1037.

Yamaguchi, K. (2002). Regression Models with Parametrically Weighted Explanatory Variables. *Sociological Methodology*, 32:219–245.