A generalized linear model for social mobility

Contents

1. Introduction

The data used in this example deals with the occupation choices of sons compared to that of their fathers. Originally published by Biblarz & Rafferty (1993), the data set **occupations.lsf** (see the **MVABOOK\Chapter 4** folder) contains 5 variables:

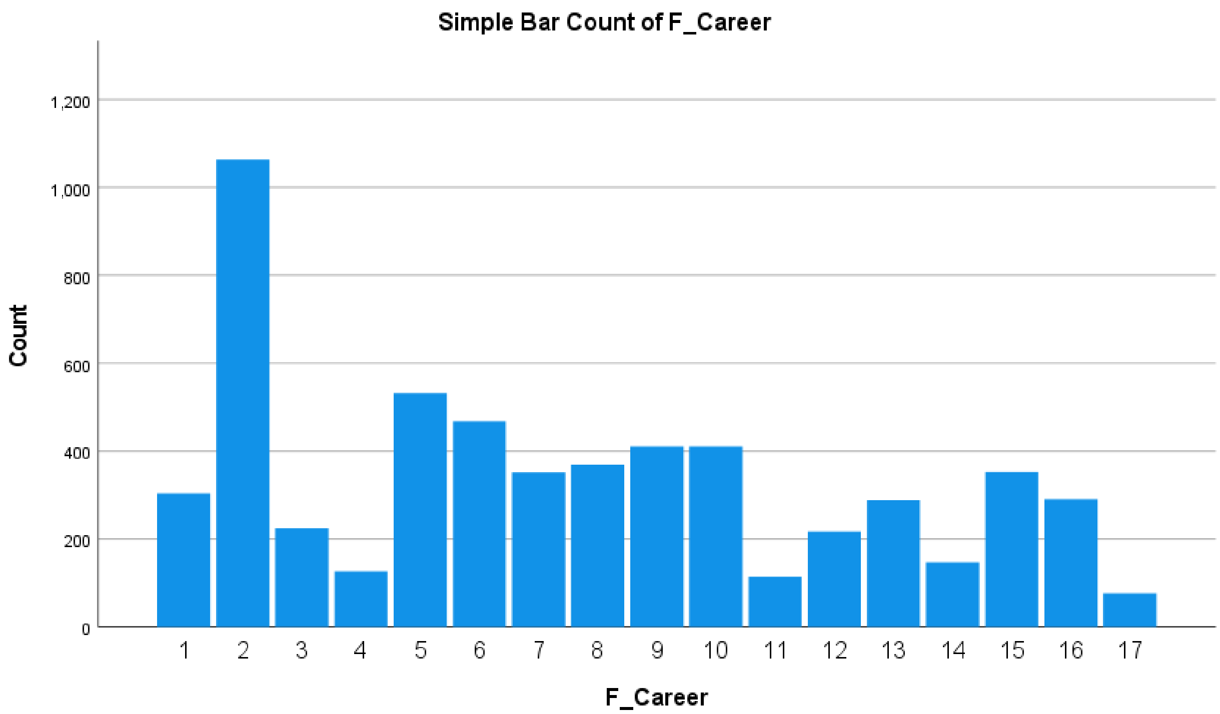
occupations.lsf

	F_Career	Son_ID	S_Choice	FamStruc	Race
1	1.00	1.00	1.00	0.00	0.00
2	1.00	2.00	1.00	0.00	0.00
3	1.00	3.00	1.00	0.00	0.00
4	1.00	4.00	1.00	0.00	0.00
5	1.00	5.00	1.00	0.00	0.00
6	1.00	6.00	1.00	0.00	0.00
7	1.00	7.00	1.00	0.00	0.00
8	1.00	8.00	1.00	0.00	0.00
9	1.00	9.00	1.00	0.00	0.00
10	1.00	10.00	1.00	0.00	0.00
11	1.00	11.00	1.00	0.00	0.00
12	1.00	12.00	1.00	0.00	0.00
13	1.00	13.00	1.00	0.00	0.00
14	1.00	14.00	1.00	0.00	0.00
15	1.00	15.00	1.00	0.00	0.00
16	1.00	16.00	1.00	0.00	0.00
17	1.00	17.00	1.00	0.00	0.00
18	1.00	18.00	1.00	0.00	0.00
19	1.00	19.00	1.00	0.00	0.00
20	1.00	20.00	1.00	0.00	0.00

- F_Career: The father's occupation, with possible occupation defined as 17 categories listed in the table below.
- Son_ID: An identifier of the son within the family unit, numbering the number of sons of a father.
- S_Choice: An indicator variable that assumes the value of 1 if the son do not have the same occupation as his father, and 2 otherwise.
- Famstruc: An indicator of whether the family structure is intact or not. If intact, the value 0 is assigned, if not intact a value of 1 was assigned.
- Race: Ethnicity indicator that assumes a value of 0 in the case of white respondents, otherwise 1.

Table: Career descriptions

Code assigned	Description
17	Professional, Self-employed
16	Professional, Salaried
15	Manager
14	Salesman, Non-retail
13	Proprietor
12	Clerk
11	Salesman, Retail
10	Craftsman - Manufacturing
9	Craftsman – Other
8	Craftsman – Construction
7	Service Worker
6	Operator, Non-manufacturing
5	Operator, Manufacturing
4	Laborer, Manufacturing
3	Laborer, Non-manufacturing
2	Farmer / Farm manager
1	Farm Laborer



When we look at cross-tabulation of the variables S_CHOICE and RACE for the two types of family structures, we see most of our respondents are from intact families. In intact families, sons were more likely to choose a different career, regardless of ethnicity. In non-intact families, the same pattern is observed, but it seems as if a higher percentage of non-white sons opted for a different career. A higher percentage of white respondents came from intact family structures.

S_Choice * Race Crosstabulation[a]

			Race		
			0	1	Total
S_Choice	1	Count	3775	323	4098
		% of Total	76.6%	6.6%	83.2%
	2	Count	770	59	829
		% of Total	15.6%	1.2%	16.8%
Total		Count	4545	382	4927
		% of Total	92.2%	7.8%	100.0%

a. FamStruc = 0

S_Choice * Race Crosstabulation[a]

			Race		
			0	1	Total
S_Choice	1	Count	559	134	693
		% of Total	67.8%	16.3%	84.1%
	2	Count	100	31	131
		% of Total	12.1%	3.8%	15.9%
Total		Count	659	165	824
		% of Total	80.0%	20.0%	100.0%

a. FamStruc = 1

We would like to investigate the relationship between a son's choice and his father's occupation – is the ratio of probabilities the same for all career categories. For example, is the son of a professional more or less likely to follow his father's career path than, for example, an operator in manufacturing?

It would also be of interest to evaluate the possible influence of both ethnicity and the family structure on these probabilities.

2. Logistic models

A logistic response model, in which the probability that a son will choose the same occupation as his father serves as outcome is used here. We opt to use the 17 possible occupations as level-2 units, within which sons of fathers with these occupations are then nested. We also use the family structure and ethnicity as predictors in this model.

The level-1 model can be expressed as:

$$P(son\ same_{ij} \mid occupation_i) = (1 + \exp[-(a_i + b_{i1}\ FamStruc_{ij} + b_{i2}\ Race_{ij}])$$

In the level-2 model, we allow the intercept and both predictor slopes to vary randomly over the occupations, i.e.,

$$a_i = \alpha + u_i,$$
$$b_{i1} = \beta_1 + v_{i1},$$
$$b_{i2} = \beta_2 + v_{i2}$$

The syntax for this model is shown below.

occupations1.prl

```
MGlimOptions Converge=0.0001 MaxIter=100 MissingCode=-999999
             Method=Quad NQUADPTS=10 ;
Title=Fathers and Sons Occupations;
SY=occupations.lsf;
ID2=F_Career;
Distribution=BER;
Link=LOGIT;
Intercept=Yes;
DepVar=S_Choice;
CoVars=FamStruc Race;
RANDOM2=intcept FamStruc Race;
```

At the beginning of the output file, the following information is given. We note that the largest observed number of fathers were farmers or farm managers (1063). From the descriptive statistics, we see the observed probability of a son choosing a different path given his father's occupation is 0.83.

```
   Model and Data Descriptions

Sampling Distribution                    = Bernoulli
Link Function                            = Logistic
PROB(Success)= 1.0/[1.0+EXP(-ETA)]

Number of Level-2 Units                    17
Number of Level-1 Units                    5751
Number of Level-1 Units per Level-2 Unit =
304  1063   225   127   532   468   352   369   411   411   115   217
289   147   353   291    77

o============================================================o
| Descriptive statistics for all the variables in the model |
o============================================================o
                                                    Standard
Variable      Minimum     Maximum        Mean     Deviation
--------      -------     -------        ----     ---------
S_Choic1       0.0000      1.0000      0.8331        0.3729
S_Choic2       0.0000      1.0000      0.1669        0.3729
intcept        1.0000      1.0000      1.0000        0.0000
FamStruc       0.0000      1.0000      0.1433        0.3504
Race           0.0000      1.0000      0.0951        0.2934
```

The estimated regression weights at convergence indicate that neither family structure nor ethnicity seem to have a significant effect on a son's choice of occupation, as both effects have a *p*-value above 0.05.

```
-2lnL (deviance statistic) =       4905.91825
    Akaike Information Criterion       4923.91825
    Schwarz Criterion                  4983.83241
```

Estimated regression weights

Parameter	Estimate	Standard Error	z Value	P Value
intcept	-1.7850	0.1774	-10.0625	0.0000
FamStruc	-0.2813	0.1919	-1.4660	0.1426
Race	-0.0882	0.2147	-0.4109	0.6811

Odds Ratio and 95% Odds Ratio Confidence Intervals

Parameter	Estimate	Odds Ratio	Bounds Lower	Upper
intcept	-1.7850	0.1678	0.1185	0.2376
FamStruc	-0.2813	0.7548	0.5182	1.0994
Race	-0.0882	0.9156	0.6011	1.3945

We next consider the estimates of the variances and covariances of the intercept and covariates. Again, none of these bar the random variation in the intercept is statistically significant. We conclude that family structure and ethnicity hasa a significant effect on the probability of a son choosing the same occupation as his father.

Estimated level 2 variances and covariances

Parameter	Estimate	Standard Error	z Value	P Value
intcept/intcept	0.4833	0.1836	2.6316	0.0085
FamStruc/intcept	0.0922	0.1137	0.8108	0.4175
FamStruc/FamStruc	0.1344	0.1423	0.9445	0.3449
Race/intcept	-0.1712	0.1413	-1.2119	0.2256
Race/FamStruc	0.0851	0.1103	0.7714	0.4405
Race/Race	0.1796	0.1567	1.1460	0.2518

As a result, we may want to consider the more parsimonious model where only the intercept is allowed to vary randomly over the occupational groups. The syntax for that is shown below:

occupations2.prl

```
MGlimOptions Converge=0.0001 MaxIter=100 MissingCode=-999999
             Method=Quad NQUADPTS=10 ;
Title=Fathers and Sons Occupations;
SY=occupations.lsf;
ID2=F_Career;
Distribution=BER;
Link=LOGIT;
Intercept=Yes;
DepVar=S_Choice;
CoVars=FamStruc Race;
RANDOM2=intcept;
```

Selected output for this model is as follows:

-2lnL (deviance statistic) = 4913.37243

```
Akaike Information Criterion       4921.37243
Schwarz Criterion                  4948.00095
```

Estimated regression weights

Parameter	Estimate	Standard Error	z Value	P Value
intcept	-1.7720	0.1711	-10.3563	0.0000
FamStruc	-0.1031	0.1085	-0.9499	0.3422
Race	-0.0998	0.1299	-0.7680	0.4425

Estimated level 2 variances and covariances

Parameter	Estimate	Standard Error	z Value	P Value
intcept/intcept	0.4500	0.1679	2.6797	0.0074

When we compare the deviances obtained for the two models, we note that the more comprehensive model had a lower deviance (4905 compared to 4913 for the simpler model). It has to be kept in mind though that this small reduction in deviance was obtained at the cost of estimating 5 additional parameters, which leads us to conclude that the simpler model describes the data adequately. This conclusion is in line with the reported AIC values: the AIC value for the simpler model is actually slightly lower than for the more complicated model.