¶ Statistical Intersectionality of Male Occupational Inheritance

Sta 223: Midterm Project

Richard Paul Yim

22 February 2023

1 Introduction

Intersectionality is the idea that social identities such as race, class and ethnicity, *intersect* to impact an individual's privileges and social experiences (e.g., discrimination, biases). Within the broad scope of questions that can be characterized as problems of intersectionality, a particular question of interest is male occupational inheritance: how does an individuals socioeconomic history affect his current occupation, and effectively his economic wealth. With respect to this question we present a case study on male occupational inheritance from data made available by the Data and Program Library Service at the University of Wisconsin-Madison utilizing the generalized linear model.

2 Data

In this section we briefly go over the background of the dat, and relevant preprocessing steps for analysis.

2.1 Background

The original dataset was collected by the US Bureau of the Census at the University of Chicago and the University of Wisconsin-Madison. The survey data is sampled from 1962 to 1973 across the continental United States with $n = 21107$ data points. The format of the original data set is presented as a $17 \times 17 \times 2 \times 2$ table where each cell is a frequency count, where the first two dimensions correspond to 17 different occupational outcomes with respect to father's occupation the son's occupation. The third and second dimensions correspond to whether a family is black—black being at least one black family member between father and son where mixed is considered black—, and whether the family is nonintact—family disruption by absentee parent, death of parent, divorce, etc... With such complexity in covariates, we perform some data preprocessing before continuing with modeling and analysis.

Some additional elaboration on occupations: among the 17 classes, these classes are numerically ordered. For example, "Professional, Self-Employed" and "Professional-Salaried" are numerically coded very high at values 17 and 16, respectively; down to "Farmer/Farm Manager" and "Farm Laborer" at 2 and 1, respectively. The original data set also includes differences between the son's first occupation and his current occupation as separate covariates as well. (The full and precise characterization of the original data set can be found in Biblarz and Raferty, 1993.)

2.2 Encoding

This data has a fair amount of complexity with all the possible covariates and classifications of occupation. For the purpose of our analysis we are interested in understanding expected frequencies or counts of intersectional variations as defined by the son's and father's occupation, race and family disruption. Effectively, with this four dimensional frequency table, there is a lot of complexity in the way among 17 possible occupations per son and father occupational pairs. To reduce this complexity, we introduce some intuitive binnings of occupation class: among the 17 occupations, four natural classes were shown to appear: 1. professionals, 2. skilled workers, 3. unskilled workers and 4. farmers. This effectively reduces our table dimensions from $17 \times 17 \times 2 \times 2$ to $4 \times 4 \times 2 \times 2$. (More precisely, relative to the original ordinal occupation encodings, occupations 1-2 are grouped as farm laborers; class 3-6 are classified as unskilled; 7-11 as skilled and 12-17 as professional/managerial.)

Furthermore, in our study we do not consider the son's first occupation, only his current occupation. Altogether we formally define our starting design matrix with main effects as

$$\mathbf{X} = [\mathbf{1}\ \mathbf{X}_{F1}\ \mathbf{X}_{F2}\ \mathbf{X}_{F3}\ \ \mathbf{X}_{S1}\ \mathbf{X}_{S2}\ \mathbf{X}_{S3}\ \ \mathbf{X}_3\ \mathbf{X}_4],$$

where $\mathbf{X}_{Fi}$ and $\mathbf{X}_{Si}$ for $i = 1, 2, 3$ are dummy encoded variables (indicator variables) with value 1 corresponding to a father and son being in that class, respectively (it should also be noted that when $\mathbf{X}_{Si} = 0$ and/or $\mathbf{X}_{Fi} = 0$ for all $i = 1, 2, 3$ the data point corresponds to a baseline in farmer occupation). $\mathbf{X}_3$ corresponds to whether the family is of black heritage with 1 being coded as yes; $\mathbf{X}_4$ corresponds to whether the family is nonintact—with no history of family disruption—with 1 being coded as yes, nonintact. We let the first column of our design matrix be a column of ones which will correspond to our slope in our generalized linear model—to be discussed in the next section. Additionally, this design matrix has 64 rows corresponding to $4 \times 4 \times 2 \times 2$ possible combinations of male occupation inheritance intersectionality as we've defined.

Before continuing on to model selection, it's important to note the efficacy and reason behind using a generalized linear model or GLMs: GLMs are highly interpretable. With an effective model specification, a lot of inference and interpretable conclusions can be derived from modeling our data and research problem with a GLM, where our parameters β have nice asymptotic properties as our sample size in our data grows and our covariates are linearly independent, a result that can be theoretically and formally proven with the Lindeberg-Feller central limit theorem.

3 Modeling

In this section we discuss our choice of generalized linear model and formally state the regression problem; discuss the model selection procedures; as well as outlier detection considerations.

3.1 Generalized Linear Model

At first view one may be tempted to immediately characterize this problem as a classification type regression. In other words, one may be tempted to use a multinomial regression model. However, from our natural and intuitive binning specification between the four occupation class we are left with considerable imbalances and severe lack of uniformity in frequencies for the son's occupation, which would be the natural response, if we asked the question:

> How is son's occupation impacted by intersectional dynamics (e.g., race, family disruption, father's occupation)?

However, this misses the essence of our problem which is to study and understand occurences/frequencies of total and combined intersectional characteristics. In other words, we want the question:

> On average, how frequent do certain intersectional outcomes occur within our sample data, and how are they representetative of the population at the time of sampling?

With this question in mind, we characterize the following initial regression model:

$$\log(\mathbb{E}(\mathbf{Y}|\mathbf{X})) = \mathbf{X}\beta,$$

where β is a nine dimensional parameter vector where each entry i corresponds to the ith covariate/column in our design matrix $\mathbf{X}$, defined in the previous section; and $\mathbf{Y}$ is our response of counts/frequencies according to each profile of intersectionality occuring within our data (again, of which there are 64). This generalized linear model with what's known as a "log link" is called a Poisson regression model, logarithmically relates the conditional expectation of *count* response data to a linear combination of covariates in the design matrix.

Before continuing to our formal modeling procedures, we'd like to emphasize again that even if were considering the first research question posed above, in our arguably preferable and intutitive occupation binning specification, the counts are unbalance in what would be our expected labeling. In particular, out of the sons in the data, there are 885 farmers, 5436 unskilled workers, 6311 skilled workers and 8475 professionals. Specifically, the consequence of such imbalance is poor model fitting with multinomial regression approaches,

so our data and research question is ill-posed for multinomial regression. In particular, we include code in the Appendix where the corresponding runs for what are known as proportional odds and baseline odds models produce poor fit according to certain diagnostic requirements, which we demonstrate use in the section 3.2, the next section.

3.2 Model Selection

Often the model that we begin with will not be the final model under consideration. While we began with the model in Section 3.1, we ended up with an entirely different set of covariates and a different design matrix and parameter vector, $\mathbf{X}'$ and β', respectively, which is essentially derived from our original design matrix $\mathbf{X}$. Namely, our final model is of the following form:

$$\log(\mathbb{E}(Y|\mathbf{X}')) = \mathbf{X}'\beta',$$

where $\mathbf{X}'$ is a design matrix of 31 columns and 64 columns. This model was arrived at by using a bidirectional stepwise Bayesian Information Criterion (BIC) regression. This regression procedure starts from the initial model and iteratively adds and removes covariates from the base model according to a scope of covariates bounded above by a larger model, and below by a smaller model. The metric for deciding to add or remove covariates is determined by the BIC score, where the lower the score, the lower the mix of model complexity and explanation of variance by our predictors. BIC effectively scores how complex our model is and how much variance is induced by our model, such that lower BIC is better (lower model complexity, greater explainablity of response by the covariates). Specifically, from our base we define a smaller scope model that consists of just the intercept as

$$\log(\mathbb{E}(Y|\mathbf{X}_{small})) = \mathbf{1}\beta_0,$$

and a larger model including second order interactions, or

$$\log(\mathbb{E}(Y|\mathbf{X}_{large})) = \mathbf{X}^{(2)}\beta^{(2)},$$

where $\mathbf{X}^{(2)}$ consists of 64 rows, like before, and 31 covariates including 8 choose 2 interaction variables across each of the 8 covariates in the intitial model. The corresponding parameter vector $\beta^{(2)}$ is a 31 dimensional vector, again with each entry corresponding to a column in $\mathbf{X}^{(2)}$. We report that we start with a BIC of 4868.912 in our initial main effects model, to a BIC of 636.7593 in our final model. (The reason we choose BIC stepwise regression is because it allows us to generally aim for less complex models with lesser covariates.) However, even with this procedure we actually end up with a full second order interaction model with main effects.

Some very important things to note with respect to our coefficient estimates: we see that many of the coefficients are statistically significant in our full coefficients table (referenced in Section 5.1.1 of the Appendix, along with confidence two-sided confidence intervals in 5.1.2). The conclusion section will provide inference and interpretability in our model, leaving our analysis at the accompanied by our discussion. With our model selection decided, in our next section we discuss goodness-of-fit. (Note: in this case $\mathbf{X}' = \mathbf{X}^{(2)}$ and $\beta' = \beta^{(2)}$.)

3.3 Goodness-of-Fit

One diagnostic to measure goodness-of-fit between the covariates and response is by using residual deviance statistic. This statistic effectively measures the discrepancy of our fitted values against the true values according to the covariates in our model. As a standard procedure we begin with the null model of just the intercept, adding each covariate in our full final model to compute corresponding residual deviances when we iteratively add each covariate up until we arrive at the final model that is being studied. We show a deviance table representing the change in residual deviance as we add predictors through each row, row-by-row, in Section 5.1.2 in the Appendix. Residual deviance in our table is decreasing at a statistically significant level with $\alpha = 0.01$. (Table 2 in the Appendix shows the corresponding deviance statistics between models and the p-values from our χ^2 test, and they are found to all be statistically significant deviances). If we compare the residual deviance of the null model consisting of just the slope, at a residual deviance of 41330.49, to the final model with all second order interactions and main effects at a residual deviance of 110.03, we see that our

final model has very low deviance with a corresponding residual deviance degrees of freedom of 33, indicating that our model has good fit according to this diagnostic.

In addition to studying model deviances by adding covariates one-by-one, we can also study the deviance residual plots against fitted values. The idea behind this diagnostic is that if we fit the point cloud of deviance residuals, we can fit a smoothing spline to see if there is any nonlinear nonzero relationship between fitted values and deviance residuals, which may indicate that our Poisson regression model with log link does not accurately describe a linear relationship between the log of the expected value of counts conditioned on our covariates. Referring to Figure 1, we can see that we have sufficiently good model fit where the smoothing spline in red is relatively flat with no strong nonlinearities.

With these two diagnostics, we can safely conclude that our final model of all second order interactions including the main effects, and of the main effects, has good fit; our model is a valid and justified model. (As mentioned in 3.1, the multinomial regression was ill-posed and it was further demonstrated by using the residual versus fitted values diagnostic that even with high smoothing penalty constant, there were vey persistent and strong nonlinearities in the multinomial regression model.)

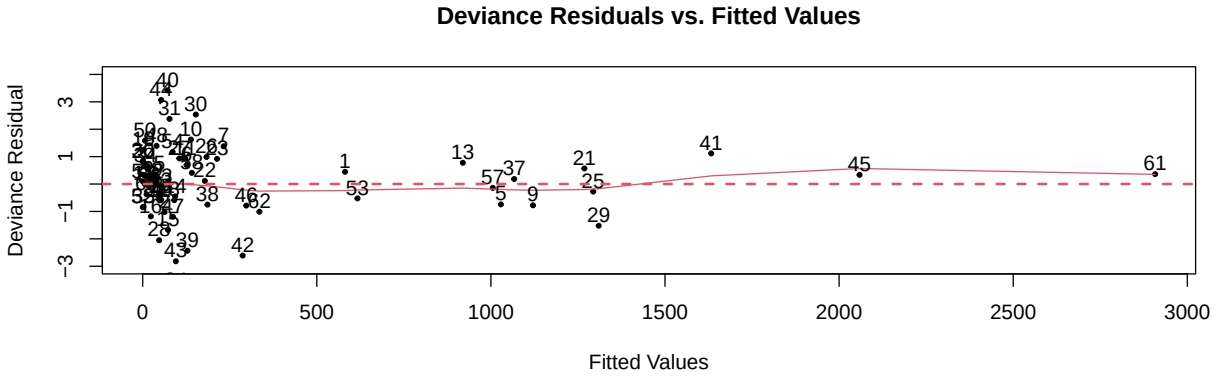


Figure 1: Goodness of fit with residuals against fitted values. No nonlinear nonzero interactions.

3.4 Outlier Detection

With all the discussion about model selection and goodness-of-fit, we briefly cover a point of outliers in our model. We detected one outlier in our data in particular the data point consisting of father and son occupation being professional, and not a black or nonintact family. The corresponding count for this intersection corresponds to maximum count across any observation at $n = 2927$. Based off of our intial main effects model, it is found that through leverage analysis and Cook's distances, that this intersection is *statistically* an outlier.

Referring to Figure 2, the first corresponds to the leverage analysis plot. Leverage values for points are effectively measurements for how much a deviation is incurred when an observation is included in a model. Points with high leverage depending on their placement perturb the slopes of a model, or essentially perturb the parameter vector. Among the points of *high leverage*, which are classified as points with leverage greater than $2\frac{p}{n}$, where p is the number of covariates and n is the number of observations, we can look at the corresponding Cook's distance (a measurement of residual error of the data point scale by a function of its leverage) to determine if the corresponding points of high leverage or influence are good for our model.

Despite this particular data point appearing statistically to be an outlier, for our purposes we still keep it in our model because in the broader context of our research problem, this a very large intersection that should be included and modeled, and may actual impose conservative estimates of our regression coefficients. Therefore, we keep this outlier in our final model.

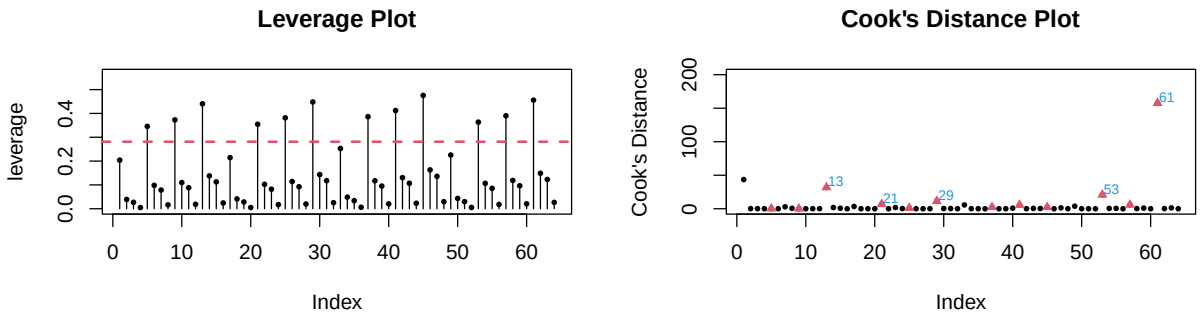


Figure 2: Leverage plots and Cook's distance from final model. Data point 61 corresponds to the outlier.

4 Conclusion

In this section we perform some scientific hypotheses and finish with a short discussion. We propose, test and report the results three scientific hypothesis based off of our final model:

1. Is there any effect from the predictors on population count? Or

$$H_0 : \sum_{i=1}^{31} |\beta_i| = 0 \text{ v.s. } H_1 : \sum_{i=1}^{31} |\beta_i| > 0.$$

   Referring to our full Coefficients table as well as our confidence interval table in the Appendix, we see that many of the coefficients are nonzero at a statistically significant level of $\alpha = 0.01$, such that we can reject the null hypothesis here that there is no effect on our model by our covariates, in favor of the alternative hypothesis that indeed there is an effect. Again, this just validates our consideration for whether the covariates have any affect on the response.

2. Is there is a positive association with the interaction covariate where the both the son's occupation and father's occupation are professionals? Or

$$H_0 : \beta_{sP:fP} \leq 0 \text{ v.s. } H_1 : \beta_{sP:fP} > 0,$$

   where $\beta_{sP:fP}$ is the coefficient corresponding to the interaction effect of both son and father being professionals. The reason we are interested in this interaction is because it addresses our outlier concern from before. For this test, we perform a one-sided z-test, and find that the corresponding p-value is 3.26e-145, so we can reject the null hypothesis that this coefficient is less than or equal to zero, accepting the alternative that it is positive at a statistically significant level.

3. Is the parameter for fatherProfessional greater than the parameter for sonUnskilled given an interaction of family disruption? Or

$$H_0 : \beta_{fP:nonintact} \leq \beta_{sU:nonintact} \text{ v.s. } H_1 : \beta_{fU:nonintact} > \beta_{sP:nonintact}.$$

   Looking in Section 5.1.5 we have a 99% confidence ellipsoid that intersect both the null and alternative parameter sets, indicating that we cannot reject the null hypothesis, and indeed the slope for professional father interacting on nonintact is less than or equal to the slope for unskilled son interacting with nonintact.

From 1. we've essentially confirmed that at the very least we are able to model effects on frequencies in each intersection based off of our data and model 2. regarding the largest cohort or intersection of both father and son professionals, we see that there is a sort accumulation of occupational inheritance occurring at very least explaining perhaps why this intersection belongs to the largest group 3. The slope of effect of a professional father on disrupted families is no greater than that of an unskilled son from a disrupted family, indicating as an example that family disruption renders all social mobility affects null.

5 Appendix

5.1 Additional Tables and Figures

5.1.1 Full Coefficients Table

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	6.3651	0.0406	156.84	0.0000
fatherOccupunskilled	-2.4891	0.1369	-18.18	0.0000
fatherOccupskilled	-2.4159	0.1319	-18.31	0.0000
fatherOccupprofessional	-2.3915	0.1339	-17.86	0.0000
sonOccupunskilled	0.5710	0.0497	11.50	0.0000
sonOccupskilled	0.6568	0.0492	13.36	0.0000
sonOccupprofessional	0.4586	0.0509	9.00	0.0000
blackyes	-2.5858	0.1294	-19.99	0.0000
nonintactyes	-2.2228	0.1078	-20.62	0.0000
fatherOccupunskilled:sonOccupunskilled	2.6987	0.1414	19.09	0.0000
fatherOccupskilled:sonOccupunskilled	2.4524	0.1370	17.90	0.0000
fatherOccupprofessional:sonOccupunskilled	1.8802	0.1414	13.30	0.0000
fatherOccupunskilled:sonOccupskilled	2.6326	0.1414	18.62	0.0000
fatherOccupskilled:sonOccupskilled	2.7920	0.1361	20.51	0.0000
fatherOccupprofessional:sonOccupskilled	2.2827	0.1395	16.37	0.0000
fatherOccupunskilled:sonOccupprofessional	2.8429	0.1422	20.00	0.0000
fatherOccupskilled:sonOccupprofessional	3.2220	0.1365	23.60	0.0000
fatherOccupprofessional:sonOccupprofessional	3.5428	0.1382	25.63	0.0000
sonOccupunskilled:blackyes	1.0992	0.1371	8.02	0.0000
sonOccupskilled:blackyes	0.3725	0.1402	2.66	0.0079
sonOccupprofessional:blackyes	0.0502	0.1432	0.35	0.7258
blackyes:nonintactyes	1.1557	0.0572	20.21	0.0000
fatherOccupunskilled:blackyes	-0.2958	0.0643	-4.60	0.0000
fatherOccupskilled:blackyes	-0.6294	0.0669	-9.40	0.0000
fatherOccupprofessional:blackyes	-1.4729	0.0953	-15.46	0.0000
fatherOccupunskilled:nonintactyes	0.1380	0.0625	2.21	0.0273
fatherOccupskilled:nonintactyes	0.3531	0.0596	5.93	0.0000
fatherOccupprofessional:nonintactyes	0.1277	0.0662	1.93	0.0537
sonOccupunskilled:nonintactyes	0.1236	0.1170	1.06	0.2907
sonOccupskilled:nonintactyes	0.1315	0.1166	1.13	0.2596
sonOccupprofessional:nonintactyes	-0.0649	0.1174	-0.55	0.5807

5.1.2 Two-Sided Confidence Intervals of coefficients

	2.5 %	97.5 %
(Intercept)	6.28	6.44
fatherOccupunskilled	-2.77	-2.23
fatherOccupskilled	-2.68	-2.17
fatherOccupprofessional	-2.66	-2.14
sonOccupunskilled	0.47	0.67
sonOccupskilled	0.56	0.75
sonOccupprofessional	0.36	0.56
blackyes	-2.85	-2.34
nonintactyes	-2.44	-2.02
fatherOccupunskilled:sonOccupunskilled	2.43	2.99
fatherOccupskilled:sonOccupunskilled	2.19	2.73
fatherOccupprofessional:sonOccupunskilled	1.61	2.17
fatherOccupunskilled:sonOccupskilled	2.36	2.92
fatherOccupskilled:sonOccupskilled	2.53	3.07
fatherOccupprofessional:sonOccupskilled	2.02	2.57
fatherOccupunskilled:sonOccupprofessional	2.57	3.13
fatherOccupskilled:sonOccupprofessional	2.96	3.50
fatherOccupprofessional:sonOccupprofessional	3.28	3.82
sonOccupunskilled:blackyes	0.84	1.38
sonOccupskilled:blackyes	0.10	0.65
sonOccupprofessional:blackyes	-0.22	0.34
blackyes:nonintactyes	1.04	1.27
fatherOccupunskilled:blackyes	-0.42	-0.17
fatherOccupskilled:blackyes	-0.76	-0.50
fatherOccupprofessional:blackyes	-1.66	-1.29
fatherOccupunskilled:nonintactyes	0.02	0.26
fatherOccupskilled:nonintactyes	0.24	0.47
fatherOccupprofessional:nonintactyes	-0.00	0.26
sonOccupunskilled:nonintactyes	-0.10	0.36
sonOccupskilled:nonintactyes	-0.09	0.36
sonOccupprofessional:nonintactyes	-0.29	0.17

5.1.3 Deviance Table

Table 1: Deviance Table.

	Df	Deviance	Resid. Df	Resid. Dev	Pr(>Chi)
NULL			63	41330.49	
fatherOccup	3	187.47	60	41143.02	0.0000
sonOccup	3	7453.18	57	33689.84	0.0000
black	1	16701.54	56	16988.30	0.0000
nonintact	1	12554.62	55	4433.68	0.0000
fatherOccup:sonOccup	9	3055.88	46	1377.80	0.0000
sonOccup:black	3	513.62	43	864.18	0.0000
black:nonintact	1	390.94	42	473.24	0.0000
fatherOccup:black	3	302.42	39	170.82	0.0000
fatherOccup:nonintact	3	41.73	36	129.09	0.0000
sonOccup:nonintact	3	19.06	33	110.03	0.0003

5.1.4 Pearson Residual versus Fitted Values Plot

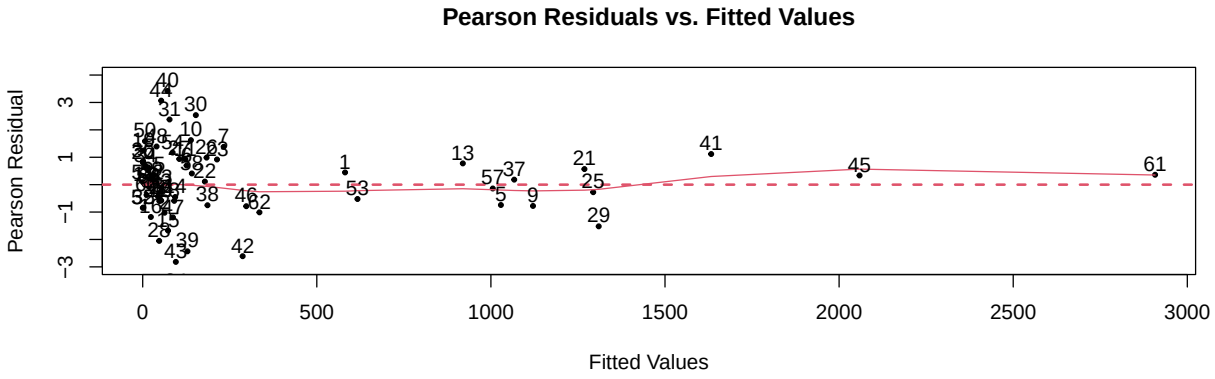
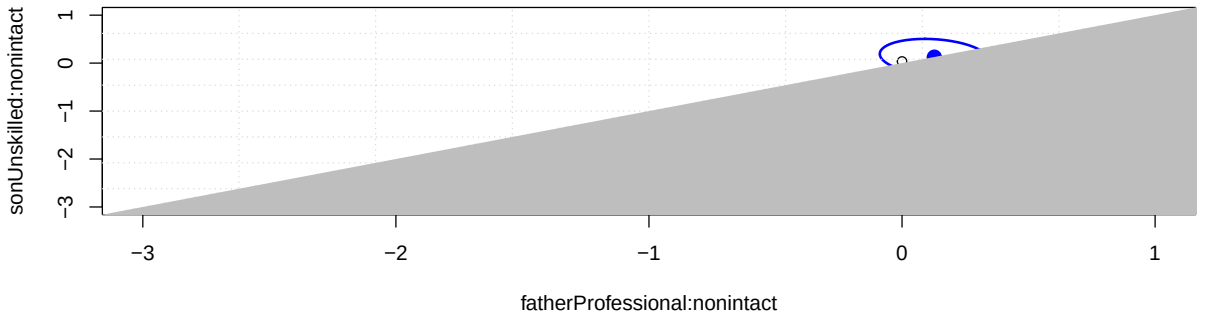


Figure 3: Plot of Pearson Residual values

5.1.5 Confidence ellisoid for sonUnskilled:nonintact and fatherProfessional:nonintact covariates



5.2 References

1. https://grodri.github.io/glms/datasets/#mobility
2. http://lib.stat.cmu.edu/datasets/socmob
3. T.J. Biblarz, A.E. Raftery. *The Effects of Family Disruption on Social Mobility.* American Sociological Review. Vol. 58, No. 1 (Feb., 1993), pp. 97-109.

5.3 Code

```
knitr::opts_chunk$set(echo = TRUE)
# ======================================================================
# Data
# ======================================================================
# Data reading and cleaning
library(tidyr)
library(dplyr)
# read read data
read <- read.table('mobility.txt', header = T)
mobility <- read[2:6]
# add missing data
mobilityFull <- rbind(mobility[1:51,],c("professional", "farm","yes","yes",0),
                      mobility[-(1:51),])
# type casting
factorcasting <-lapply(mobilityFull[,1:4],factor)
mobilityFull[,1:4] <- factorcasting
mobilityFull[,5] <- as.numeric(mobilityFull$n)
mobilityFull[,5] <- as.numeric(mobilityFull$n)
mobilityFull$fatherOccup <- factor(mobilityFull$fatherOccup, levels=c("farm",
                                          "unskilled", "skilled", "professional"))
mobilityFull$sonOccup <- factor(mobilityFull$sonOccup, levels=c("farm",
                                          "unskilled", "skilled", "professional"))
rownames(mobilityFull) <- 1:64

# spread data for goodness-of-fit analysis
mobilityRes <- mobilityFull %>%
  group_by(sonOccup) %>%
  spread(sonOccup, n, fill = 0) %>%
  as.data.frame
# counts
sum(mobilityFull[mobilityFull$sonOccup=="farm",]$n)
sum(mobilityFull[mobilityFull$sonOccup=="unskilled",]$n)
sum(mobilityFull[mobilityFull$sonOccup=="skilled",]$n)
sum(mobilityFull[mobilityFull$sonOccup=="professional",]$n)

sum(mobilityFull[mobilityFull$fatherOccup=="farm",]$n)
sum(mobilityFull[mobilityFull$fatherOccup=="unskilled",]$n)
sum(mobilityFull[mobilityFull$fatherOccup=="skilled",]$n)
sum(mobilityFull[mobilityFull$fatherOccup=="professional",]$n)

sum(mobilityFull[mobilityFull$nonintact=="yes",]$n)
sum(mobilityFull[mobilityFull$nonintact=="no",]$n)
sum(mobilityFull[mobilityFull$black=="yes",]$n)
sum(mobilityFull[mobilityFull$black=="no",]$n)
```

```r
# ======================================================================
# model fitting
# ======================================================================
# full model
mobility.glm <- glm(n~., family=poisson(), data=mobilityFull)
res.P = residuals(mobility.glm,type="pearson")
res.D = residuals(mobility.glm,type="deviance")

# ======================================================================
# Model Selection
# ======================================================================
library(MASS)
# bic first order model
model.null <- glm(n~1, family=poisson(), data=mobilityFull)
model.full <- glm(n~.^2, family=poisson(), data=mobilityFull)
mobility2.glm <- stepAIC(mobility.glm, direction="both",
            scope=list(lower=model.null, upper=model.full), k=log(dim(mobilityFull)[1]))
anova(mobility2.glm, test="Chisq")

confint(mobility2.glm)
library(MASS)
# bic first order model
model.null <- glm(n~1, family=poisson(), data=mobilityFull)
model.full <- glm(n~.^2, family=poisson(), data=mobilityFull)
mobility2.glm <- stepAIC(mobility.glm, direction="both",
              scope=list(lower=model.null, upper=model.full), k=log(dim(mobilityFull)[1]))
anova(mobility2.glm, test="Chisq")

# ======================================================================
# Goodness-of-fit
# ======================================================================
res2.P = residuals(mobility2.glm,type="deviance")
res2.D = residuals(mobility2.glm,type="deviance")

# Deviance X fitted
plot(mobility2.glm$fitted, res2.D, xlab="Fitted Values",ylab="Deviance Residual",
     ylim = c(-3,4), pch = 19, cex = 0.5)
lines(smooth.spline(mobility2.glm$fitted.values, res2.D, spar=1.5), col=2)
text(mobility2.glm$fitted, res2.D+0.4, labels = seq(length(res2.D)))
abline(h=0,col=2,lwd=2,lty=2)
title("Deviance Residuals vs. Fitted Values")

# ======================================================================
# Outlier Check
# ======================================================================
par(mfrow=c(1,2))
# leverage points
leverage <- hatvalues(mobility.glm)
W <- diag(mobility.glm$weights)
X <- model.matrix(mobility.glm)
Hat = sqrt(W) %*% X %*% solve(t(X) %*% W %*% X) %*% t(X) %*% sqrt(W)
all(abs(leverage - diag(Hat)) < 1e-15)
```

```r
p <- length(coef(mobility.glm))
n <- nrow(mobilityFull)
plot(names(leverage), leverage, xlab="Index", type="h", ylim = c(0,2*p/n*2))
points(names(leverage), leverage, pch=16, cex=0.6)
abline(h=2*p/n,col=2,lwd=2,lty=2)
infPts <- which(leverage>2*p/n)
title("Leverage Plot")

# Cook's distance
cooks <- cooks.distance(mobility.glm)
plot(cooks, ylab="Cook's Distance", pch=16, cex=0.6,ylim=c(0,200))
points(infPts, cooks[infPts], pch=17, cex=0.8, col=2)
susPts <- as.numeric(names(sort(cooks[infPts], decreasing=TRUE)[1:5]))
text(susPts, cooks[susPts], susPts, adj=c(-0.1,-0.1), cex=0.7, col=4)
dispersion <- 1
all(abs(cooks - (res.P/(1 - leverage))^2 * leverage/(dispersion * p) < 1e-15))
title("Cook's Distance Plot")
# ========================================================================
# Pearson X fitted
# ========================================================================
plot(mobility2.glm$fitted, res2.P, xlab="Fitted Values",ylab="Pearson Residual",
     ylim = c(-3,4), pch = 19, cex = 0.5)
lines(smooth.spline(mobility2.glm$fitted.values, res2.P, spar=1.5), col=2)
abline(h=0,col=2,lwd=2,lty=2)
text(mobility2.glm$fitted, res2.P+0.4, labels = seq(length(res2.P)))
title("Pearson Residuals vs. Fitted Values")
# confidence ellipsoid
library(car)
x=seq(-4,4,0.01)
plot(c(0),c(0),ylim=c(-3,1), xlim=c(-3, 1),col="white",
     ylab="sonUnskilled:nonintact", xlab="fatherProfessional:nonintact")
points(0,0.035,bg="red")
confidenceEllipse(mobility2.glm, which.coef = c(28,29), levels=0.99, add=T)
grid(8,8)
y2 <- x
y1 <- rep(-10,length(x))
# Fill area between lines
polygon(c(x, rev(x)), c(y2, rev(y1)),
        col = "gray", lty = 0)

# one-sided interaction (alt greater than 0)
pnorm(summary(mobility2.glm)$coefficients[18,3], lower.tail = F) # p-value

# ========================================================================
# Proportional Odds
# ========================================================================
library(MASS)
mobilityPLR <- polr(formula = sonOccup ~
                      (fatherOccup + black + nonintact)^2, data = mobilityFull, weights = n)
summary(mobilityPLR)

# Model Selection
stepAIC(mobilityPLR)
```

```r
stepAIC(mobilityPLR, k=log(dim(mobilityFull)[1]))
mobilityPLR.AIC <-polr(formula = sonOccup ~ fatherOccup + black + nonintact + fatherOccup:nonintact,
     data = mobilityFull, weights = n)

# ======================================================================
# Baseline Odds
# ======================================================================
library(nnet)
mobility.base = multinom(sonOccup~ (fatherOccup+nonintact+black+n)^2, data=mobilityFull)
summary(mobility.base)

# Model Selection
stepAIC(mobility.base)
stepAIC(mobility.base, k=log(dim(mobilityFull)[1]) )
mobilityBASE.AIC <- multinom(formula = sonOccup ~ n, data = mobilityFull)

source("mainMultinomial.R")
mobilityPLR.AIC <-polr(formula = sonOccup ~ fatherOccup + black + nonintact + fatherOccup:nonintact,
     data = mobilityFull, weights = n)

# ======================================================================
# Proportional Odds Residuals
# ======================================================================
# Pearson residuals
obslabel = cbind(mobilityFull$sonOccup=="farm", mobilityFull$sonOccup=="unskilled",
                 mobilityFull$sonOccup=="skilled", mobilityFull$sonOccup=="professional")*1
obslabel <-obslabel[c(TRUE,rep(FALSE,3)), ]
prd_prob_po <- fitted(mobilityPLR.AIC)
prd_prob_po <- unique(prd_prob_po)
resP.plr <- sapply(1:(ncol(obslabel)-1), function(m) {
  obs_m <- rowSums(as.matrix(obslabel[,1:m]))
  fit_m <- rowSums(as.matrix(prd_prob_po[,1:m]))
  (obs_m - fit_m) / sqrt(fit_m * (1 - fit_m))
})
fitted_m <- sapply(1:(ncol(obslabel)-1), function(m){
  rowSums(as.matrix(prd_prob_po[,1:m]))
})
# plot
if(0){
par(mfrow=c(1,3))
plot(fitted_m[,1], resP.plr[,1], xlab="po.fitted probability for class 1", ylab="Pearson residual",
     sub = "spar = 0.8")
lines(smooth.spline(x = fitted_m[,1], y = resP.plr[,1], spar=0.8), col=2); abline(h=0,lty=2)
plot(fitted_m[,2], resP.plr[,2], xlab="po.fitted probability for class 1&2", ylab="Pearson residual",
     sub = "spar = 1.4")
lines(smooth.spline(x = fitted_m[,2], y = resP.plr[,2], spar=1.4), col=2); abline(h=0,lty=2)
plot(fitted_m[,3], resP.plr[,3], xlab="po.fitted probability for class 1&2&3", ylab="Pearson residual",
     sub = "spar = 1.4")
lines(smooth.spline(x = fitted_m[,3], y = resP.plr[,3], spar=1.4), col=2); abline(h=0,lty=2)
}


# ======================================================================
```

```r
# Baseline Odds Residuals
# ================================================================
# baseline
prd_prob_bo2 = fitted(mobilityBASE.AIC)
fitted_m = sapply(2:(ncol(obslabel)), function(m){
  fit_m <- prd_prob_bo2[rowSums(obslabel[,c(1,m)]) > 0,c(1,m)]
  fit_m <- fit_m[,2] / rowSums(fit_m)
})
baseN <- 2
resP.bo <- sapply(2:ncol(obslabel), function(m) {
  # baseline is column 1 here
  # otherwise you should replace "1" with the corresponding index and adjust the range of "m" according
  obs_m <- obslabel[rowSums(obslabel[,c(baseN,m)]) > 0, m]
  fit_m <- prd_prob_bo2[rowSums(obslabel[,c(baseN,m)]) > 0,c(1,m)]
  fit_m <- fit_m[,baseN] / rowSums(fit_m)
  (obs_m - fit_m) / sqrt(fit_m * (1 - fit_m))
})
if(0){
# plot
par(mfrow=c(1,3))
plot(fitted_m[,1], resP.bo[,1], xlab="po.fitted probability for class 2 vs 1", ylab="Pearson residual",
     sub = "spar = 0.9")
lines(smooth.spline(x = fitted_m[,1], y = resP.bo[,1], spar=1.1), col=10); abline(h=0,lty=2)
plot(fitted_m[,2], resP.bo[,2], xlab="po.fitted probability for class 3 vs 1", ylab="Pearson residual",
     sub = "spar = 1.5",ylim=c(-10,10))
lines(smooth.spline(x = fitted_m[,2], y = resP.bo[,2], spar=.1), col=10); abline(h=0,lty=2)
plot(fitted_m[,3], resP.bo[,3], xlab="po.fitted probability for class 4 vs 1", ylab="Pearson residual",
     sub = "spar = 1.5",ylim=c(-10,10))
lines(smooth.spline(x = fitted_m[,3], y = resP.bo[,3], spar=1.1), col=10); abline(h=0,lty=2)
}
```