Social Mobility Data

Llasmin Lopez

February 8, 2021

1 Introduction

The beta binomial model on this data set can be described as follows: For any litter, let n denote the number of implants and X denote the number implants that die. Then

$$Q(x,n) := P(X = x, n) = \binom{n}{x} \prod_{r=0}^{x-1}(\mu + r\theta) \prod_{r=0}^{n-x-1}(1 - \mu + r\theta) / \prod_{r=0}^{n-1}(1 + r\theta)$$

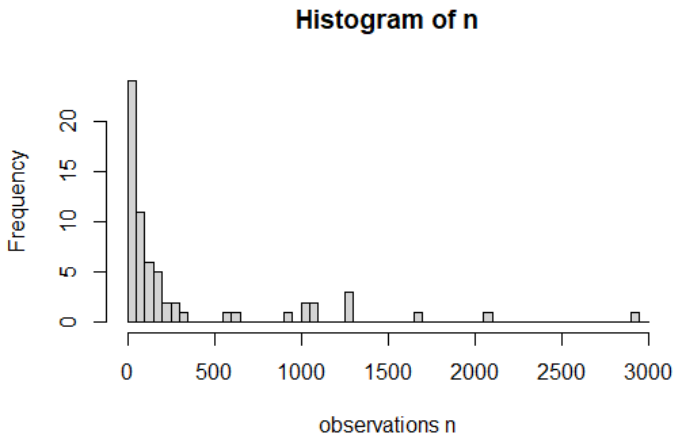
The United States of America are synonymous with the idea of social mobility, motivating many to migrate to America for the opportunity to work for a better position in life. In this analysis, we use the data Social Mobility to model and investigate the relationship between sons' occupation and that of their fathers' along with race and family background. We apply multinomial regression methods and compare the fits of both a Baseline Odds Model and a Proportional Odds Model as well as test the hypothesis that all coefficients are zero valued.

2 Data

The Social Mobility data was obtained from [Princeton] and consists of five variables: fathers' occupation, sons' occupation, race, and family background. Each subject is assumed to be independent. For occupation variables, occupations are categorized into four categories: farm, unskilled, skilled, and professional; these occupations are presented in a seemingly increasing order. The variable 'race' indicates whether a person is black or not, and the variable 'family background' indicates whether the family structure is intact or not.

With our goal of modeling, this data presents its' first challenge in the number of covariates. Sons' occupation is the response variable which leaves only three explanatory variables:

fathers' occupation, race, and family background. Also, it is not clearly stated whether race refers to that of the sons or that of the fathers, or if biracial families were excluded from the sample. There is also some ambiguity about whether the observed sons' occupation refers to the sons' first occupation or, instead, his most current occupation. Notable observations of the data were that many cases were observed fewer than five times and one particular case is an outlier with over twelve percent of total entries, shown here:



3 Methods

As previously mentioned, the response variable is a factor with four levels and we have multiple factor covariates thus, we fit a multinomial regression. Multinomial regression models relate categorical responses with more than two response categories to covariates. There are two models we consider: Baseline Odds Model and Proportional Odds Model.

It is often unclear which of these two models is the superior choice, we will use both methods and compare their results. In the Baseline Odds model, we specify a baseline category and our linear predictor is $\frac{\pi_{ij}}{\pi_{i1}} = \exp(\eta_{ij})$ for $\eta_{ij} = \mathbf{X}_i\beta_j, \beta_j \in \mathbf{R}^p, 2 \leq j \leq M$ where $M = 4$ in our model. Thus for the number of covariates p, we have $(M-1)p$ predictors.

The Proportional Odds model is applicable for both ordered and unordered categorical data however, it may prove to be more rewarding in cases of ordered data. In our data, the order of occupations is *farm, unskilled, skilled, professional*. Let index j represent the response categories (i.e. $j = 1, ..., M$ for $M = 4$). Then for z_{im} denoting new responses and y_{ij} observed ordinal responses, $z_{im} = \sum_{j=1}^{m} y_{ij}$. For $\mu_{im} = E(z_{im})$ and binary link functin g, our model is $g(\mu_{im}) = \beta_{0m} + \mathbf{X}_i\beta$.

3.1 Baseline Odds Model

We select "farm" to be the baseline. Then, we use RStudio to construct estimated models:

$$logit(\hat{P}(Y \leq 1)) = -2.53+0.557*(U)+0.977*(S)+1.736*(P)-0.594*(blackyes)-0.117*(nonintactyes)$$

$$logit(\hat{P}(Y \leq 2)) = -0.144+0.557*(U)+0.977*(S)+1.736*(P)-0.594*(blackyes)-0.117*(nonintactyes)$$

$$logit(\hat{P}(Y \leq 3)) = 1.2230+0.557*(U)+0.977*(S)+1.736*(P)-0.594*(blackyes)-0.117*(nonintactyes)$$

where fathers' occupation is denoted U if unskilled, denoted S for skilled, and denoted P for professional and Y values 1,2,3 correspond to sons' occupation.

Denote the slope parameters as $\beta_1 = (\beta_{1,1}, \beta_{1,2}, \beta_{1,3}, \beta_2, \beta_3)$.Consider the hypothesis that all slope parameters are zero. That is, $H_0 : \hat{\beta} = \mathbf{0}$, $H_A : \hat{\beta} \neq \mathbf{0}$. Then for alpha equal to 0.5, we apply the Gaussian approximation to obtain p-values for this test. The slope estimates, their p-values, and their 95% confidence intervals are shown below:

	Value	**p -value**	**Confidence Interval**
$\hat{\beta_{1,1}}$	0.557	1.655192e-49	[0.483, 0.631]
$\hat{\beta_{1,2}}$	0.9766	2.712850e-156	[0.905, 1.048]
$\hat{\beta_{1,3}}$	1.7356	0.000000e+00	[1.657, 1.815]
$\hat{\beta_2}$	-0.5938	3.873668e-39	[-0.683, -0.50]
$\hat{\beta_3}$	-0.1171	1.920759e-03	[-0.191, -0.043]

Table 1: Slope estimates

Since all p-values were significant to the level alpha, we reject the null.

To asses goodness-of-fit, we calculated Pearson residuals by category:

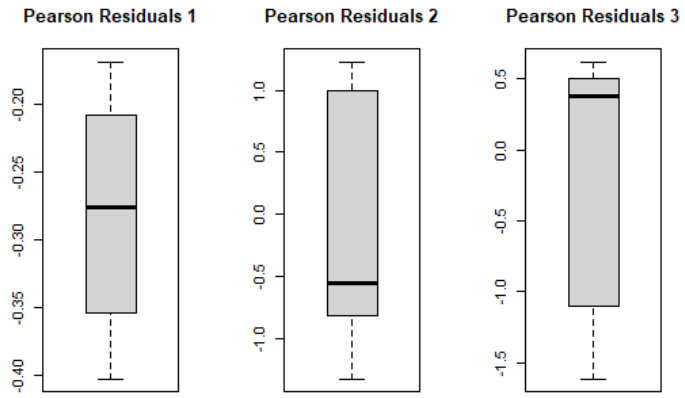
3.2 Proportional Odds Model

Under this model, not only do coefficients vary model to model based on ordinal, but the slope estimates also vary for each fit:

$$logit(\hat{P}(Y \leq 1)) = 0.569+2.70*(U)+2.45*(S)+1.88*(P)+1.0989*(blackyes)0.1370*(nonintactyes)$$

$$logit(\hat{P}(Y \leq 2)) = 0.656+2.63*(U)+2.79*(S)+2.28*(P)+0.3723*(blackyes)+0.1413*(nonintactyes)$$

$$logit(\hat{P}(Y \leq 3)) = 0.457+2.84*(U)+3.22*(S)+3.54*(P)+0.0517*(blackyes)-0.0515*(nonintactyes)$$



Denote the slope parameters as before and again consider the hypothesis that all slope parameters are zero. That is, $H_0 : \hat{\beta} = \mathbf{0}$, $H_A : \hat{\beta} \neq \mathbf{0}$. Then for alpha equal to 0.5, we obtain p-values for this test. The slope estimates, their p-values, and their 95% confidence intervals are shown below for the unskilled model:

	Value	**p -value**	**Confidence Interval**
$\hat{\beta_{1,1}}$	2.70	3.295951e-81	[2.4212, 2.9754]
$\hat{\beta_{1,2}}$	2.45	1.236940e-71	[2.1841, 2.7213]
$\hat{\beta_{1,3}}$	1.88	2.356951e-40	[1.6028, 2.1569]
$\hat{\beta_2}$	1.0989	9.327528e-16	[0.8309, 1.3669]
$\hat{\beta_3}$	0.1370	0.2413256	[-0.0922, 0.3661]

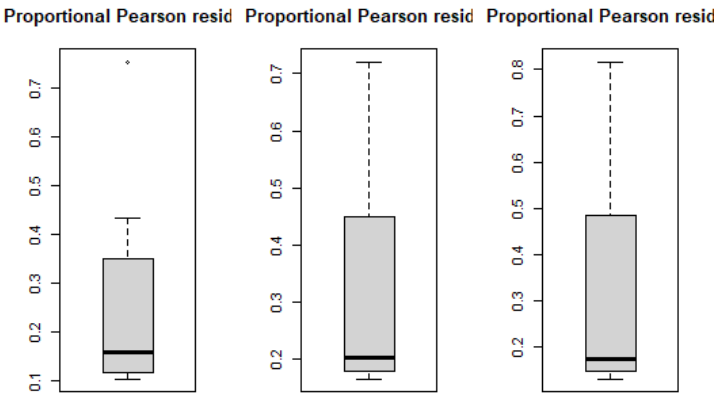
Table 2: Slope estimates under Unskilled

All but one coefficient had p-values significant to level alpha. Since the alternative was that at least one of the coefficients were not equal to zero, we still reject the null.

To asses goodness-of-fit, we calculated Pearson residuals by model:

4 Results

For both Baseline and Odds proportional models, we were able to reject the hypothesis of all coefficients being equal to zero, i.e. their corresponding covariates have some influence on the response variable. Comparing the Pearson residual plots for each model, it is clear there is some inconsistency. The residual plots under the Odds model are skewed and the first model



observes an outlier.

5 Discussion

Since it is easier to interpret the coefficients of the odds model, we will focus on interpreting those coefficients. Overall, as the occupation of the son improves, the odds of being more likely to have a father with a higher ordered occupation increases. This can be observed in each step with the exception of the son being a skilled worker and the odds of his father being an unskilled worker. It seems that sons are more likely to have the same profession as their fathers, followed by the proceeding occupation. This would appear to support the notion of social mobility.

It is difficult, at this point, to be confident in the fit of any of the models presented. It was noted earlier that one combination of predictors composed quite a bit of the data while other combinations of predictors occurred only a few times, if at all. Both of the methods used rely on a large enough sample size n to calculate estimates thus the results shown may be mislead.

References

[1] Biblarz, Timothy J., and Adrian E. Raftery. "The Effects of Family Disruption on Social Mobility." American Sociological Review, vol. 58, no. 1, 1993, pp. 97–109. JSTOR

[2] Bruin, J. 2006. newtest: command to compute new test. UCLA: Statistical Consulting Group. https://stats.idre.ucla.edu/stata/ado/analysis/.

[3] Rodriguez,"Generalized Linear Models" Princeton University.
https://data.princeton.edu/wws509/datasets/mobility

6 Appendix: R Code

```
# Social mobility data
# Llasmin Lopez

library(MASS)
library(nnet)
library(lawstat)

######## social mobility data ########
#####################################
  mobil = read.table("C:/Users/Llasmin/OneDrive/New One
  Drive/OneDrive/Documents/mobility.dat.txt")

  insertRow <- function(existingDF, newrow, r) {
    existingDF[seq(r+1,nrow(existingDF)+1),] <-
    existingDF[seq(r,nrow(existingDF)),]
    existingDF[r,] <- newrow
    existingDF
  }

  mobil = insertRow(mobil, c("professional","farm","yes","yes",0),52)

# Data Exploration
  mobil$fatherOccup = factor(mobil$fatherOccup,
                             levels =
                             c('farm','unskilled','skilled','professional')
                             )
  mobil$sonOccup = factor(mobil$sonOccup,
                          levels =
                          c('farm','unskilled','skilled','professional'))
  mobil$black = as.factor(mobil$black)
  mobil$nonintact = as.factor(mobil$nonintact)
  mobil$n = as.numeric(mobil$n)
  sum(mobil$n) # 21,107
  par(mfrow=c(1,1))
  hist(mobil$n, breaks = 50*c(0:60), xlab = "observations n",main = 'Histogram of n')
  summary(mobil[mobil$fatherOccup=='farm',5])
  boxplot(n ~ fatherOccup, data = mobil)
```

```
boxplot(n~ fatherOccup + black, data = mobil, ylab = "Range of n",
        xlab = "Fathers' Occupation : Race", main = 'Size vs Occupation
        stratified by Race')
    # outlier
    which.max(mobil[mobil$black=='no',5]) #31
    not_b = mobil[mobil$black == 'no',]
    not_b[31,] # father: professional,
               # son: professional,
               # black = no,
               # nonintact = no,
               # n = 2927 ~ 13.867% of cases
boxplot(n ~ fatherOccup + nonintact, data = mobil, ylab = "Range of n")
    # outlier
    which.max(mobil[mobil$nonintact=='no',5]) #31
    not_in = mobil[mobil$nonintact == 'no',]
    not_in[31,] # same outlier observed previously

boxplot(n ~ black + nonintact, data = mobil, ylab = "Range of n" )
boxplot(n ~ fatherOccup + black + nonintact, data = mobil, ylab = "Range of n" )



######### Multinomial Regression => 2 models must be explored ########
#####################################################################

### 1) Proportional odds model###----------------------------------------
mobil.plr <- polr(sonOccup ~ fatherOccup + black + nonintact, weights = n,
data = mobil)
#mobil2.plr <- polr(sonOccup ~ fatherOccup * black * nonintact, weights =
n, data = mobil)

summary(mobil.plr)
27#summary(mobil2.plr)

# n by M matrix of predicted prob
prd_prob_po = fitted(mobil.plr)
# vector of predicted labels
prd_labl_po = predict(mobil.plr)
```

```
head(data.frame(prd_labl_po, prd_prob_po))

## Testing
  (ctable = coef(summary(mobil.plr)))
  #p-values
  pv = pnorm(abs(ctable[,"t value"]), lower.tail = F)*2
  (ctable = cbind(ctable, "p value" = pv))
  # C.I.
  (ci <- confint.default(mobil.plr)) # assuming normality

  #2
  # n by M matrix of predicted prob
  prd_prob_po2 = fitted(mobil2.plr)
  # vector of predicted labels
  prd_labl_po2 = predict(mobil2.plr)

  head(data.frame(prd_labl_po2, prd_prob_po2))

  ## Testing
  #(ctable2 = coef(summary(mobil2.plr)))
  #p-values
  #pv2 = pnorm(abs(ctable2[,"t value"]), lower.tail = F)*2
  #(ctable2 = cbind(ctable2, "p value" = pv2))
  # C.I.
  #(ci2 <- confint.default(mobil2.plr)) # assuming normality

##Pearson residuals------------------------------------------------------

#reformatting data:
sonOccup <- rbind(matrix(mobil$n[1:16], byrow=F,
ncol=4),matrix(mobil$n[17:32], byrow=F, ncol=4),
                  matrix(mobil$n[33:48], byrow=F, ncol=4),
                  matrix(mobil$n[49:64], byrow=F, ncol=4))
colnames(sonOccup) <- c("farm","unskilled", 'skilled','professional')
mobil2 <- data.frame(sonOccup, mobil[c(1:4,17:20,33:36,49:52),c(1,3,4)])
head(mobil2)

obslabel <- t(apply(mobil2[,1:4], 1, function(x) {
  res <- numeric(4)
```

```
  res[which.max(x)] <- 1
  res
}))
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[seq_len(nrow(mobil2))*4,1:m]))
  (obs_m - fit_m) / sqrt(fit_m * (1 - fit_m))
})
summary(resP.plr)
head(resP.plr)
par(mfrow=c(1,3))
boxplot(resP.plr[,1], main='Pearson Residuals 1')
boxplot(resP.plr[,2], main = 'Pearson Residuals 2')
boxplot(resP.plr[,3], main = 'Pearson Residuals 3')
               (tvalue1 = sum(resP.plr[,1]))
               (tvalue2 = sum(resP.plr[,2]))
               (tvalue3 = sum(resP.plr[,3]))
                qchisq(0.95,9)
               pchisq(-4.488873,9)
               pchisq(-0.5750669,9)
               pchisq(-2.87531,9)

### 2) Baseline Odds Model###-----------------------------------------
mobil.bo <- multinom(sonOccup ~ fatherOccup + black + nonintact, weights =
n, data = mobil)
#mobil2.bo <- multinom(sonOccup ~ fatherOccup * black * nonintact, weights
= n, data = mobil)

# provides coefficient estimates and corresponding standard errors.
summary(mobil.bo, digit = 3)
#summary(mobil2.bo, digit = 3) #again not worth adding extra terms

# z values
  zval.bo <- coef(mobil.bo) / summary(mobil.bo)$standard.errors
  # two-sided p-values
  pval.bo <- 2 * pnorm(abs(zval.bo), lower.tail=FALSE)

  (ztable = coef(summary(mobil.bo)))
  (ztable = cbind(ztable, "p value" = pval.bo))
```

```
  # C.I.
  (CI <- confint(mobil.bo)) # assuming normality

prd_prob_bo = fitted(mobil.bo)
head(prd_prob_bo)
prd_labl_bo = predict(mobil.bo)
head(prd_labl_bo)

  # Pearson residuals------------------------------------------------------

# a list of (M-1) elements, each element contains the Pearson residuals for
one submodel
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" accordingly
  obs_m <- obslabel[rowSums(obslabel[,c(1,m)]) > 0, m]
  fit_m <- prd_prob_bo[rowSums(obslabel[,c(1,m)]) > 0,c(1,m)]
  fit_m <- fit_m[,2] / rowSums(fit_m)
  (obs_m - fit_m) / sqrt(fit_m * (1 - fit_m))
})
head(resP.bo)
resP.bo
par(mfrow=c(1,3))
boxplot(resP.bo[[1]], main = 'Proportional Pearson resid.1')
boxplot(resP.bo[[2]], main = 'Proportional Pearson resid.2')
boxplot(resP.bo[[3]], main = 'Proportional Pearson resid.3')
```