Live Session Week 4: Analyzing Nominal and Ordered Data

Devesh Tiwari

June 6, 2017

Agenda

1. Review question (10 minutes)
2. Breakout 1: EDA part I (15 mins)
3. Breakout 2: EDA part II (25 mins)
4. Breakout 3: Coefficient interpretation (25 mins)
5. Breakout 4: Comparing Nominal v Ordered logistic (time remaining)

Group Discussion 1: 10 minutes

You are a data scientist for a retail company and are interested in understanding which customer characteristics are correlated with customers' purchasing behavior. A different team handed you a dataset which contains several covariates of customer characteristics; this dataset also classifies a customer as being those who *never purchase*, *sometimes purchase*, and *purchases a lot*.

Which model would you rather use, a nominal response model or an ordinal logistic regression model? Why?

Introduction

In this exercise, we will explore the relationship between voters' self identified party affiliation and their demographic characteristic. In particular, we seek to answer whether voters' age, race, and gender influence their party choice. For this exercise we will use the data from the **American National Election Survey**, which conducted a survey several months prior to the 2016 American Presidential elections. Note that the original survey data uses survey weights, which we will not be using.

The dataset "*w271_summer2017_anes.csv*" contains a handful of variables from the survey, and these variables have been cleaned and modified for this exercise. This dataset contains the following variables:

Variable Name	**Explanations**
ftwhite, ftblack, ftmuslim, ftpolice	Feeling thermometer variables where respondents are asked to rate their favorability of whites, blacks, muslims, and police on a 0 − 100 scale.
Presjob	A seven point scale indicating respondents' evaluation of President Obama. 1 = Very strongly approve; 7 = Very strongly disapprove
Srv_spend	Seven point scale representing the degree to which respondents believe that the government should provide or should not provide services: 1 = Government should provide many fewer services; 7 = Government should provide many more services.
crimespend	A seven point scale representing degree to which respondents think that the federal government should or should not increase federal spending on crime. 1 = Increased a great deal; 7 = Decreased a great deal
ideo5	A five point scale of respondents' self reported ideology. 1 = Very liberal; 5 = Very conservative
party	Categorical variable indicating respondents' party affiliation: Democrat, Independent, Republican
age	Respondents' age, as of 2016.
race_white	Dummy variable taking a value of one if the respondent is white and is zero otherwise.
female	Dummy variable taking a value of one if therespondent is female and is zero otherwise.

Breakout Session 1: Exploratory Data Analysis (10 mins in breakout, 5 mins together)

We do not have time to conduct a thorough EDA. However, think about the type of plots or tables you would need to create in order to conduct an EDA that would help you adjudicate between using a nominal response model and an ordinal logistic model. Answer this question for a categorical variable and a continous variable.

If you have time, choose a (A) Categorical independent variable and a (B) Continuous independent variable from the dataset which has been provided today. Remember, the DV is *party*.

```
rm(list = ls())
require(knitr)
opts_chunk$set(tidy.opts=list(width.cutoff=60),tidy=TRUE)

require(vcd)
```

```
## Warning in library(package, lib.loc = lib.loc, character.only = TRUE,
## logical.return = TRUE, : there is no package called 'vcd'
```

```
require(nnet)
require(car)
require(MASS)

path <- "~/Documents/Projects/MIDS/Summer 2017/live_sessions/week4"
setwd(path)

df <- read.csv("w271_summer2017_anes.csv",
               stringsAsFactors = FALSE, header = TRUE, sep = ",")

names(df)
```

```
##  [1] "ftwhite"    "ftblack"    "ftmuslim"   "presjob"    "srv_spend"
##  [6] "crimespend" "ideo5"      "ftpolice"   "party"      "age"
## [11] "race_white" "female"
```

```
# INSERT CODE HERE
```

Breakout Session 2: EDA part 2 (15 mins in breakout rooms; 10 mins together)

In the code below, I create some contingency tables between the dependent variable and *ideo5* (ideology) and *crimespend*. Take a look and answer the following questions

1. Is *party* independently related to each of the two independent variables?
2. Is their an ordered relationship between *party* and each of the two independent variables? What does your exploration tell you about the relationship between these variables? Hint: You might find the prop.table function helpful.

```
party.ideo.table <- xtabs(~party + ideo5, data = df)
party.ideo.table
```

```
##              ideo5
## party           1   2   3   4   5
##   Democrat    109 146 129  39   8
##   Independent  24  47 159  85  27
##   Republican    5  10  61 122  72
```

```
party.crimespend.table <- xtabs(~party + crimespend, data = df)
party.crimespend.table
```

```
##              crimespend
## party           1   2   3   4   5   6   7
##   Democrat     83 104  97 134  24  11   6
##   Independent  39  66  82 119  30  19  24
##   Republican   23  67  69  90  17   5   9
```

```
## Insert code to conduct tests of independence
```

Breakout Session 3: Comparing nominal response models to ordinal logistic (15 mins in breakout rooms; 10 mins together)

In the code below, I create two bi-variate models in which *crimespend* is the only independent variable. The first model is as a nominal response model and the second model is an ordered logistic regression.

NOTE: I am treating *crimespend* an a continous variable, which strictly speaking, it is not! I am doing this for convenince, but in real life and lab situations, you should justify this choice *and* explore models in which *crimespend* is coded as a factor variable.

Take a look at the code and output and come prepared to answer the following questions:

1. Interpret the relationship between *crimespend* and partisanship in the nominal response model. Why are their two equations? Is the relationship between *crimespend* and partisanship statistically significant?
2. Interpret the relationship betweeen the two variables in the ordinal logisitc regression model. Is the relationship between *crimespend* and partisanship statistically significant?

```
mod.nominal <- multinom(party ~ crimespend, data = df)
```

```
## # weights:  9 (4 variable)
## initial  value 1228.248539
## final  value 1190.269918
## converged
```

```
mod.ordinal <- polr(as.factor(party) ~ crimespend, data = df,
    method = "logistic", Hess = TRUE)

summary(mod.nominal)
```

```
## Call:
## multinom(formula = party ~ crimespend, data = df)
##
## Coefficients:
##             (Intercept) crimespend
## Independent  -1.0762664  0.2759674
## Republican   -0.9478781  0.1475354
##
## Std. Errors:
##             (Intercept) crimespend
## Independent   0.1752312 0.04984528
## Republican    0.1848616 0.05424163
##
## Residual Deviance: 2380.54
## AIC: 2388.54
```

```
# Insert code here to determine significance

summary(mod.ordinal)
```

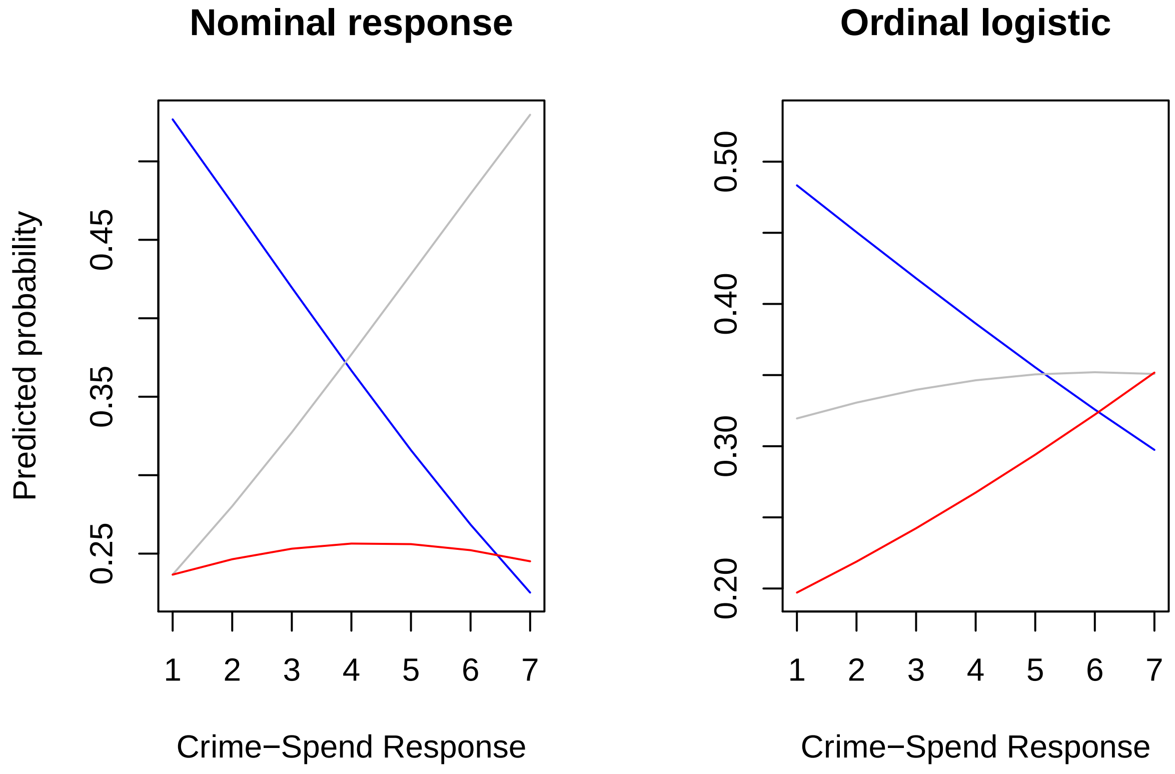
```
## Call:
## polr(formula = as.factor(party) ~ crimespend, data = df, Hess = TRUE,
##     method = "logistic")
##
## Coefficients:
##             Value Std. Error t value
## crimespend 0.1321    0.03744    3.53
##
## Intercepts:
##                        Value   Std. Error t value
## Democrat|Independent    0.0655  0.1354     0.4836
## Independent|Republican  1.5364  0.1436    10.7003
##
## Residual Deviance: 2400.021
## AIC: 2406.021
## (82 observations deleted due to missingness)
```

```
# Insert code here to determine signifiance
```

Breakout Session 4: Predicted probabilities and comparing models (Time remaining)

We created two different models using the same data, but do they give us the same results? We explore that issue here by examining two predicted probability parts, generated from each model. The blue lines correspond to the probability a respondent is a Democrat, gray corresponds to Independent, and red corresponds to Republican.

What does each chart tell you? Which one do you think is correct and why?



Nominal response

Ordinal logistic

Take home

1. Repeat this exercise but code *crimespend* as a factor variable. What do you notice?
2. Repeat this exercise with a contiuous independent variable.
3. Build a model and try to answer any question you find interesting. Interpret coefficients, conduct statistical tests, and calculate confidence intervals of the coefficients.