# W271 Section 3 Lab 2

*Kiersten Henderson, Jill Zhang, Hoang Phan, Daghan Altas*

*10/8/2017*

```r
library(knitr)
library(vcd)
opts_chunk$set(tidy.opts=list(width.cutoff=75),tidy=TRUE)
library(Hmisc)
library(ggplot2)
library(dplyr)
library(GGally)
library(data.table)
library(stargazer)
library(tidyverse)
library(forcats)
library(scales)
library(gridExtra)
library(reshape)
library(ordinal)
library(car)
```

## 1 Introduction

Our team of data scientists is working with a University foundation to identify alumni who are likely to donate in the future. We were able to model not just whether alumni with particular characteristics would donate, but rather what level of contribution particular alumni were most likely to make.

In order to evaluate what variables to include when we estimated our proportional odds model, we considered fundraising domain knowledge and the results of our explanatory data analysis (EDA). Fundraising professionals typically consider an individual's wealth (ability to give) and enthusiasm for giving when they consider who might be the most likely, loyal, and generous donors to a particular cause. For this reason and due to our EDA, we included Class.Year in the model we estimated because it is a proxy for wealth (in our dataset, the alumni who have graduated the longest ago likely have the highest earning potential). Similarly, we included Next.Degree as another potential indicator of wealth (those individuals with graduate degrees frequently have higher incomes). In addition, we grouped majors together according to level of donations typically made by donors from each major. We also considered Enthusiasm for and support of the university which was evident if an alumnus had attended a university event, and also if they had previously donated to their alma matter. For these reasons and due to our EDA, we included Event.Attendence and all previous years donations to the university in our model (FY12Giving-FY16Giving). In addition, we found in our EDA that Gender and Marital.Status played roles in the philanthropic behavior towards the University and we thus included Gender when we estimated our model.

**describe key results of model below**

SUMMARY

For several reasons supported by our EDA, we modeled donation level as a categorical variable and used a proportional odds model to estimate level of donation as the dependent variable.

Our final model is:

We found that. . . .

# 2 Exploratory Data Analysis

```
givings = read.csv("./lab2data.csv")
str(givings)
```

```
## 'data.frame':    1000 obs. of  12 variables:
##  $ X              : int  761 620 214 373 748 1080 1155 1069 1161 457 ...
##  $ Gender         : Factor w/ 2 levels "F","M": 1 2 1 1 2 1 1 1 1 1 ...
##  $ Class.Year     : int  2002 2002 1982 1992 2002 2012 2012 2012 2012 1992 ...
##  $ Marital.Status : Factor w/ 4 levels "D","M","S","W": 2 3 2 2 3 3 3 3 3 2 ...
##  $ Major          : Factor w/ 45 levels "American Studies",..: 39 25 25 2 30 2 3 26 39 15 ...
##  $ Next.Degree    : Factor w/ 47 levels "AA","BA","BAE",..: 37 39 39 35 39 15 39 35 39 18 ...
##  $ AttendenceEvent: int  1 0 1 1 0 1 0 1 0 0 ...
##  $ FY12Giving     : num  50 0 100 0 0 0 0 5 0 0 ...
##  $ FY13Giving     : num  51 0 0 0 0 0 0 10 0 75 ...
##  $ FY14Giving     : num  51 0 100 0 0 0 0 25 0 0 ...
##  $ FY15Giving     : num  0 0 100 0 0 0 0 25 0 0 ...
##  $ FY16Giving     : num  0 0 100 0 0 0 0 50 0 60 ...
```

```
sum(is.na(givings))
```

```
## [1] 0
```

## 2.1 Observations

When we conducted a cursory examination, we found that there are 1000 observations and twelve variables in the dataset (five variables are donations in years 2012-2016). There were no missing values for any of the variables in the dataset.

Donations made in the 2016 fiscal year is our dependent variable. We were provided donations for 5 years as continuous variables. The minimum donation is $0 and the maximum donation is $161 500 (made in 2013). We were provided Gender and Attendence.Event as a binary variables. Furthermore, we were provided Marital.Status as a factor with four levels (D, M, S, W), which we interpret as divorced, married, single, and windowed. Our data set includes graduating class year with five categories each ten years apart (1972, 1982, 1992, 2002, 2012). We were also provided with educational information for alumni; Alumni Major was provides in 45 categories, and information about their next degree was provided with 47 categories.

## 2.2 Data Cleaning

To facilitate interpretation of categories during EDA and modeling, we renamed some of them. In addition, we assigned categorical variables to factors where appropriate.

```
levels(givings$Gender) = c("Female", "Male")
givings$AttendenceEvent = factor(givings$AttendenceEvent, levels = c(0, 1),
    labels = c("Didn't Attend", "Attended"))
levels(givings$Marital.Status) = c("Divorced", "Married", "Single", "Widowed")
givings$Class.Year = factor(givings$Class.Year)
givings$FY12Giving = as.numeric(givings$FY12Giving)
givings$FY13Giving = as.numeric(givings$FY13Giving)
givings$FY14Giving = as.numeric(givings$FY14Giving)
givings$FY15Giving = as.numeric(givings$FY15Giving)
givings$FY16Giving = as.numeric(givings$FY16Giving)
```

In addition, we created a factor variable from FY16Givings as suggested. We did so by grouping FY2016 donations into 5 buckets and we extended this logic to all other fiscal year donations.

```
givings$FY12Giving.Grouped <- factor(cut(givings$FY12Giving, breaks = c(0, 1,
    100, 250, 500, 2e+05), labels = c("[0-1)", "[1-100)", "[100-250)", "[250-500)",
    "[500-200000)"), right = FALSE))
givings$FY13Giving.Grouped <- factor(cut(givings$FY13Giving, breaks = c(0, 1,
    100, 250, 500, 2e+05), labels = c("[0-1)", "[1-100)", "[100-250)", "[250-500)",
    "[500-200000)"), right = FALSE))
givings$FY14Giving.Grouped <- factor(cut(givings$FY14Giving, breaks = c(0, 1,
    100, 250, 500, 2e+05), labels = c("[0-1)", "[1-100)", "[100-250)", "[250-500)",
    "[500-200000)"), right = FALSE))
givings$FY15Giving.Grouped <- factor(cut(givings$FY15Giving, breaks = c(0, 1,
    100, 250, 500, 2e+05), labels = c("[0-1)", "[1-100)", "[100-250)", "[250-500)",
    "[500-200000)"), right = FALSE))
givings$FY16Giving.Grouped <- factor(cut(givings$FY16Giving, breaks = c(0, 1,
    100, 250, 500, 2e+05), labels = c("[0-1)", "[1-100)", "[100-250)", "[250-500)",
    "[500-200000)"), right = FALSE))
```

## 2.3 Univariate Data Analysis

We conducted univariate analysis of the following variables:

- FY12 though FY16 Giving (numerical, log transformed and Grouped)
- Gender
- Class.Year
- Marital.Status
- Major
- Next.Degree
- AttendenceEvent

### 2.3.1 FY12 though FY16 Giving (Numerical, log transformed, and Grouped)

We note that continuous-scale values of the contributions variable have a very strong positive skew. However, at log scale, we observe a bi-modal distribution, with most of the values centered either around 0 or around the $100 range.Therefore, we deem that a log transformation is not an appropriate way to specify the model.

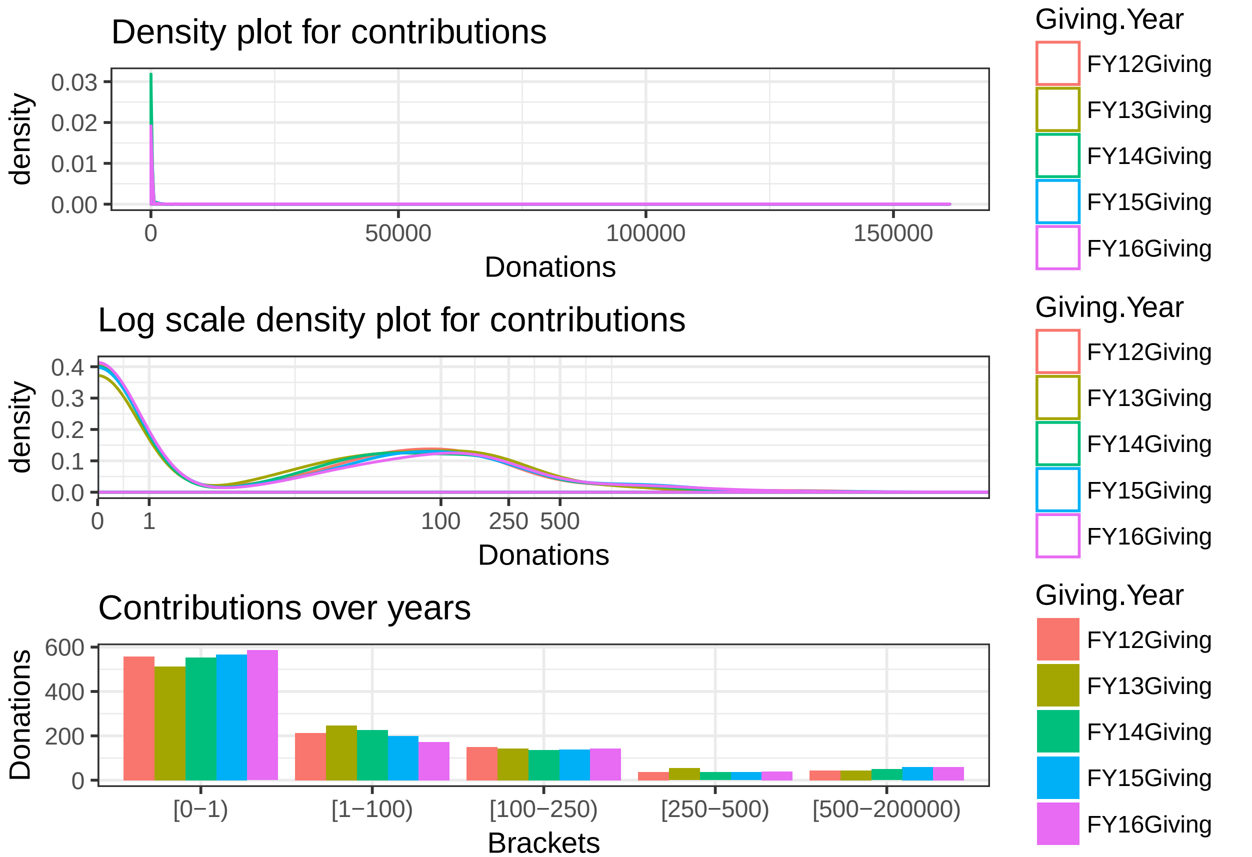
As noted in the data cleaning section, we transformed the all years' giving variables into a categorical variable. We think this approach is justified because all years exhibit a very strong skew and becasue the log transformation was not a good solution and would hinder model interpretation. Binning this variable addressed our concerns and still allows for ease of model interpretation.

From this univariate analysis alone, we can observe that most years follow a similar pattern - ie. donor behavior seems to be consistent over multiple years. We anticipate that there may be a strong correlation between a donor's 2016 donation level and their donation in previous years and we will explore this idea during bivariate analysis. This will be important to explore because it would be important to uniquely identify and model donors who consistently make large contributions because the University can use this information to maximize yearly total donations.

```
givings.tidy.donations <- givings[1:12] %>% gather("Giving.Year", "Donations",
    8:12)
givings.tidy.donations$Giving.Grouped <- factor(cut(givings.tidy.donations$Donations,
    breaks = c(0, 1, 100, 250, 500, 2e+05), labels = c("[0-1)", "[1-100)", "[100-250)",
        "[250-500)", "[500-200000)"), right = FALSE))
givings.tidy.donations.aggregate <- as.data.frame(xtabs(~Giving.Grouped + Giving.Year,
    data = givings.tidy.donations))
p1 <- ggplot(givings.tidy.donations, aes(x = Donations, colour = Giving.Year)) +
    geom_density(alpha = 0.3) + labs(title = "Density plot for contributions") +
    theme_bw()

p2 <- ggplot(givings.tidy.donations, aes(x = Donations, colour = Giving.Year)) +
    geom_density(alpha = 0.3) + scale_x_continuous(breaks = c(0, 1, 100, 250,
    500, 2e+05), trans = "log1p", expand = c(0, 0)) + labs(title = "Log scale density plot for contribu
    scale_y_continuous() + theme_bw()

p3 <- ggplot(givings.tidy.donations.aggregate, aes(x = Giving.Grouped, y = Freq)) +
    geom_bar(aes(fill = Giving.Year), stat = "identity", position = "dodge") +
    theme(axis.text.x = element_text(angle = 90, hjust = 1)) + labs(y = "Donations",
    x = "Brackets", title = "Contributions over years") + theme_bw()
grid.arrange(p1, p2, p3, ncol = 1, nrow = 3)
```



### 2.3.2 Gender

The dataset contains nearly identical numbers of female and male donors. This result is surprising because recently, male and female enrollment is quite different. According to the National Center for Education Statistics, the national average in 2015 was 56% female and 44% for male enrollment in college (https://nces.ed.gov/programs/coe/indicator_cha.asp) and graduation rates are also skewed towards women. However, because college graduation rates were likely skewed toward men in earlier years, they may average out to give an equal distribution of men and women during the period between 1972 and 2012. Furthermore, it is possible the particular University were are studying is unusual in that it has always had equal proportions of male and female alumni.

```
row <- xtabs(~Gender, data = givings)
data.frame(rbind(row, row/dim(givings)[1]), row.names = c("Donor Count", "Ratio"))
```

```
##              Female    Male
## Donor Count 505.000 495.000
## Ratio         0.505   0.495
```

### 2.3.3 Class.Year

We were surprised that only 5 graduation years are provided. Because of this, the individuals sampled are not a random subsample from the entire population of university graduates, but rather a subsample of students with each set spaced apart by 10 years. This is a major caveat and we will be cautious when applying our model to all alumni of the University.

```
row <- xtabs(~Class.Year, data = givings)
data.frame(rbind(row, row/dim(givings)[1]), row.names = c("Class.Year Count",
    "Ratio"))
```

```
##                      X1972   X1982   X1992   X2002   X2012
## Class.Year Count 105.000 176.000 203.000 223.000 293.000
## Ratio              0.105   0.176   0.203   0.223   0.293
```

### 2.3.4 Marital.Status

We were surprised to find that the Divorce to Marriage ratio is very low compared to the the expected ratio of 44% (Wikipedia https://en.wikipedia.org/wiki/Divorce_demography). That said, the measurement methodology is slightly different and we expect rates to change with graduation years (divorce rates are more likely to increase with age). So we are going to assume that Marital.Status data is valid sample for the population.

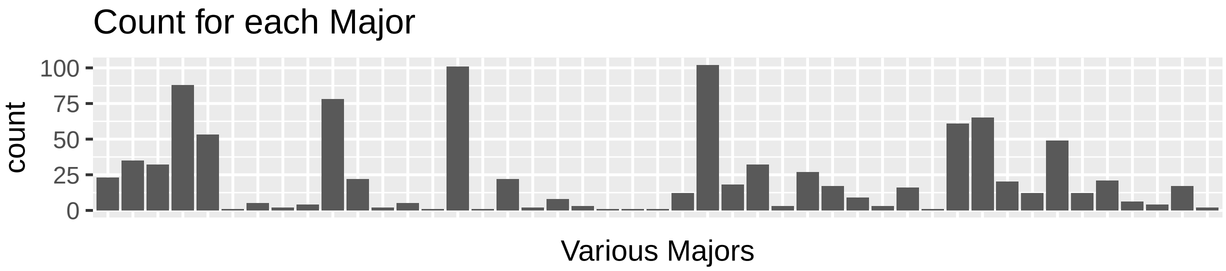
```
row <- xtabs(~Marital.Status, data = givings)
data.frame(rbind(row, row/dim(givings)[1]), row.names = c("Marital.Status Count",
    "Ratio"))
```

```
##                      Divorced Married  Single Widowed
## Marital.Status Count   61.000 584.000 344.000  11.000
## Ratio                   0.061   0.584   0.344   0.011
```

### 2.3.5 Major

Many of the categories of major have very little representation (for example Zoology, Political studies in regional studies) so we don't expect any one of them to make a significant contribution to our model. When we grouped major into larger categories (Liberal Arts, Sciences, etc - 11 categories in total), we also did not gain any insight into donor contributions. We will, however, examine Majors to identify any outlier majors that may have donated disproportionately and use that information to improve our modeling.

```
ggplot(givings, aes(x = Major)) + geom_histogram(stat = "count") + labs(title = "Count for each Major",
    x = "Various Majors") + theme(axis.text.x = element_blank(), axis.ticks.x = element_blank())
```



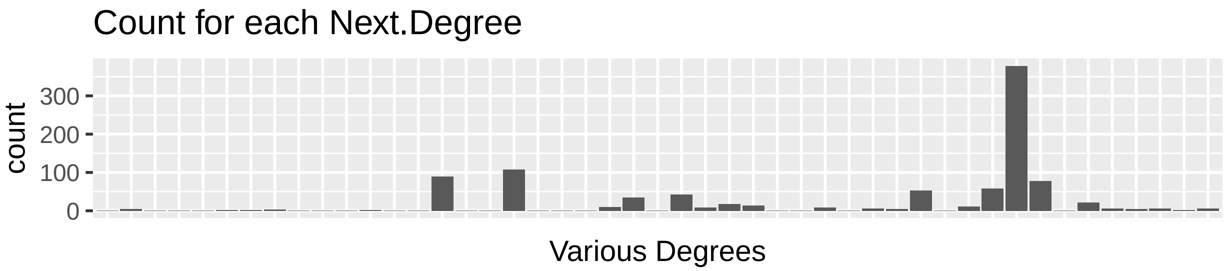
```
head(sort(xtabs(~Major, data = givings)), 3)
```

```
## Major
##             Chinese         Engineering English-Journalism
##                   1                   1                   1
```

### 2.3.6 Next.Degree

The Next.Degree variable has too many categories to be useful either as-is or grouped into subcategories.Many levels only have a single count (ex: MA2, MALS,MSM, BD, etc). We therefore grouped Next.Degree into 3 categories; alumni without a next degree, and those with next degree being a bachelors or above a bachelors (graduate level or professional degree).

```
ggplot(givings, aes(x = Next.Degree)) + geom_histogram(stat = "count") + labs(title = "Count for each N
    x = "Various Degrees") + theme(axis.text.x = element_blank(), axis.ticks.x = element_blank())
```



```
givings$Adv.Deg <- fct_collapse(givings$Next.Degree, bachelor_equivalent = c("AA",
    "BA", "BAE", "BD", "BFA", "BN", "BS", "BSN", "LLB", "LLD", "NDA", "UBDS",
    "UDDS", "UMD", "UMDS", "UNKD", "TC"), above_bachelor = c("DC", "DDS", "DMD",
    "DO", "DO2", "DP", "JD", "PHD", "MA", "MA2", "MAE", "MALS", "MAT", "MBA",
    "MCP", "MD", "MD2", "ME", "MFA", "MHA", "ML", "MLS", "MM", "MPA", "MPH",
    "MS", "MSM", "MSW", "STM"))
```

The Next.Degree as a factor variable is too scathered. Many levels only have a single count (ex: MA2, MALS, MSM, BD, etc). We will group donor into 3 catories; those without a next degree (None), those with a bachelor equivalent and those with a degree higher than bachelor.

### 2.3.7 AttendenceEvent

Surprisingly, 40% of graduates have attended at least one Alumni event organized between 2012 and 2015. Intuitively, we expect a high correlation between this variable and donations (if alumni support, or are enthusiastic about their, they are more likely to donate) so we included this variable in our analysis and modeling.

```
row <- xtabs(~AttendenceEvent, data = givings)
data.frame(rbind(row, row/dim(givings)[1]), row.names = c("AttendenceEvent Count",
    "Ratio"))
```

```
##                       Didn.t.Attend Attended
## AttendenceEvent Count       395.000  605.000
## Ratio                         0.395    0.605
```

## 2.4 Bivariate Data Analysis

We are examined the relationship between the following sets of variables:

- FY16Giving.Grouped vs. (Gender, Class.Year, Marital.Status, Major, Next.Degree, AttendenceEvent)
- FY16Giving.Grouped vs. (FY15Giving.Grouped, …., FY12Giving.Grouped)
- Gender vs. (Class.Year, Marital.Status, Major)
- Major vs. Next.Degree
- Class.Year vs. AttendenceEvent

### 2.4.1 FY16Giving.Grouped vs. Gender

We note two interesting observations in table 1. First, there are more donations in the [$500-$200K) bracket than the [$250-$500) bracket. Furthermore, at $100 donation level or above, men consistently donate more than women do.

```
t1 <- xtabs(~Gender + FY16Giving.Grouped, data = givings)
t1.1 <- round(t1/rowSums(t1), 2)
# kable(list(t1, t1.1), caption = 'Frequency vs. Ratio for
# FY16Giving.Grouped vs. Gender')
t1
```

```
##         FY16Giving.Grouped
## Gender   [0-1) [1-100) [100-250) [250-500) [500-200000)
##   Female   298     106        58        17           26
##   Male     288      67        85        22           33
```

```
t1.1
```

```
##         FY16Giving.Grouped
## Gender   [0-1) [1-100) [100-250) [250-500) [500-200000)
##   Female  0.59    0.21      0.11      0.03         0.05
##   Male    0.58    0.14      0.17      0.04         0.07
```

### 2.4.2 FY16Giving.Grouped vs. Class.Year

There are 3 key insights from table 2 below:

1. Older alumni make disproportionately larger donations (15% of the Class of 72 made $500 + donations).

2. A higher percentage of the older alumni make donations ($0 donations is only 48% for the class of 1972, versus 66% for the class of 2012).

3. But there are more recent graduates, perhaps in part because their current addresses still valid. So even as their ratio is lower, most of the $1-$100 donations come from the class of 2012.

```
t2 <- xtabs(~Class.Year + FY16Giving.Grouped, data = givings)
t2.1 <- round(t2/rowSums(t2), 2)
# kable(list(t2,t2.1), caption = 'Frequency and Ratio for FY16Giving.Grouped
# vs. Class.Year')
t2
```

```
##            FY16Giving.Grouped
## Class.Year [0-1) [1-100) [100-250) [250-500) [500-200000)
##       1972    50       9        23         7           16
##       1982    90      22        35        14           15
##       1992   115      29        38         9           12
##       2002   137      41        25         6           14
##       2012   194      72        22         3            2
```

```
t2.1
```

```
##            FY16Giving.Grouped
## Class.Year [0-1) [1-100) [100-250) [250-500) [500-200000)
##       1972  0.48    0.09      0.22      0.07         0.15
##       1982  0.51    0.12      0.20      0.08         0.09
##       1992  0.57    0.14      0.19      0.04         0.06
##       2002  0.61    0.18      0.11      0.03         0.06
##       2012  0.66    0.25      0.08      0.01         0.01
```

### 2.4.3 FY16Giving.Grouped vs. Marital.Status

Married and single people are biggest source of donations when compared to divorced or widowed alumni. We therefore expect Marital.Status to be a significant explanatory variable in our final model. However, there are many more married or single individuals compared to divorced or widowed alumni.

```
t3 <- xtabs(~Marital.Status + FY16Giving.Grouped, data = givings)
t3.1 <- round(t3/rowSums(t3), 2)
# kable(list(t3,t3.1), caption = 'Frequency and Ratio for FY16Giving.Grouped
# vs. Marital.Status')
t3
```

```
##                FY16Giving.Grouped
## Marital.Status [0-1) [1-100) [100-250) [250-500) [500-200000)
##       Divorced    36       9        11         2            3
##       Married    305      96       109        31           43
##       Single     241      66        23         4           10
##       Widowed      4       2         0         2            3
```

```
t3.1
```

```
##                FY16Giving.Grouped
## Marital.Status [0-1) [1-100) [100-250) [250-500) [500-200000)
##       Divorced  0.59    0.15      0.18      0.03         0.05
##       Married   0.52    0.16      0.19      0.05         0.07
##       Single    0.70    0.19      0.07      0.01         0.03
##       Widowed   0.36    0.18      0.00      0.18         0.27
```

### 2.4.4 Donation level vs. Major

There are 45 majors in the dataset and as mentioned, some majors only have a single record. During model estimation, it would not be appropriate include all 45 majors individually in the model as binary variables for two reasons:

1. It will cause a curse of dimensionality that will reduce predictive power.
2. These binary variables will hold most of their records as zero and we have very little information about them.

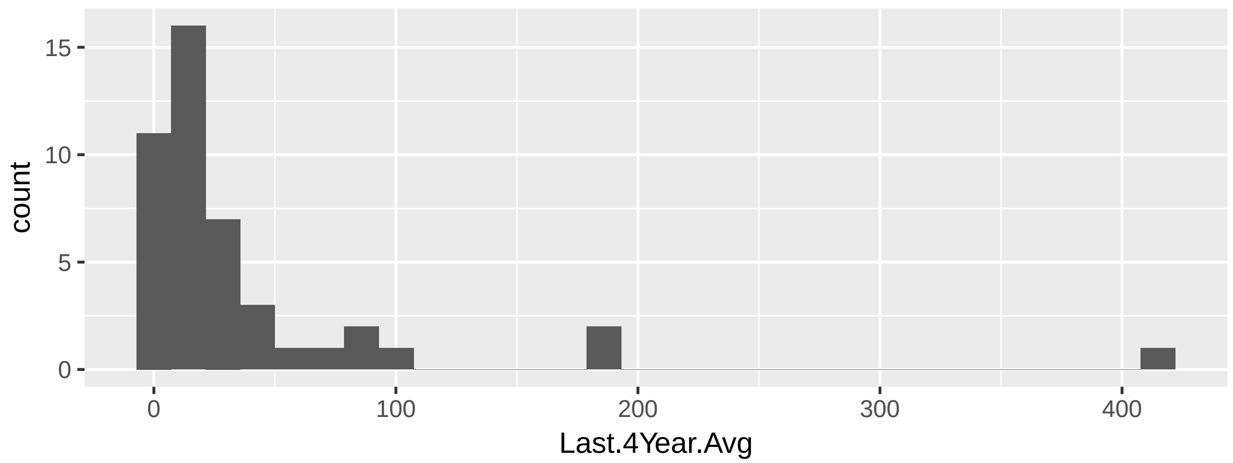
Therefore, we decide to group majors and after considering several ways of doing so, the method we found most appropriate was to group majors by their median donations in the last 4 years (2012-2015). Therefore, we began by calculating the average donation for each person over the last four years, and could then calculate the median donation over the last 4 years for each major.

To optimize grouping of majors, we began with very granular cuts in the median donation (5 dollars increases). Importantly, because it is a median value, it is much less extreme than the original donation amount (see histogram below). For half of the majors, the median donation is less than 13.75 and only 25% of majors make median donations larger than 33.75, so the original cut-offs we assigned [0, 1), [1, 100), [100, 250), [250, 500), [500, 200000) won't work for grouping major because most of the values will be skewed in the [1:100) category. Using the more granular cutoffs, we were able to observe the following patterns to make our final major groupings:

1. There is an extreme value at 400 in the histogram than came from an English Journalism alumnus. The English Jouralism major has a single alumnus in the dataset and this person donated $1500 in 2015 and has donated every year from 2012 to 2015. This is a caveat in our analysis because we are extrapolating from the donations made by a single person to represent the behavior of other alumni with that major. We proceeded by labeling English Jouralism as a high donation level major, but will definitely need more data to justify this point if possible.
2. [0,1) is a natural cut-off point which means no one in this major donates
3. Majors with median donations in [1,10) are showing a similar behavior in 2016 donations (70% if alumni in these majors didn't donate in 2016). [10-35) are showing a similar behavior in 2016 donations, with 50-60% alumni no donations, 20% donating 1-100 and about 10% donating 100-250 and 5% in following two categories . Less than 25% of the major median are larger than 35, we can group them together.
4. We therefore decided to use the following dollar value cut-offs: [0,1),[1,10),[10,35) and 35+ to group the median donation values and we gave them the follwoing donatin category labels: "No","Low","Medium", and "High" donation. We anticipate that that the higher the donation level is for a major, the higher the donation these alumni will make in 2016. This is consistent with our third contingency table in this section. As the major donation_level increase from No to High, the percentage of 2016 donation in [0,1) decreases while the percentage in higher buckets like [250-500),[500-200000) increase.

Based on this analysis, we classified majors into four donation categories: "No","Low","Medium", and "High" donation majors. We will use this variable when we estimate our model.

```
givings$Last.4Year.Avg <- rowMeans(givings[c("FY12Giving", "FY13Giving", "FY14Giving",
    "FY15Giving")])
Major.Index <- data.frame(aggregate(Last.4Year.Avg ~ Major, data = givings,
    median))
ggplot(Major.Index, aes(Last.4Year.Avg)) + geom_histogram(bins = 30)
```



```r
givings <- merge(x = givings, y = Major.Index, by = "Major")

givings$Major.Donation.Level <- factor(cut(givings$Last.4Year.Avg.y, labels = c("NO",
    "Low", "Medium", "High"), breaks = c(0, 1, 10, 30, 2e+05), right = FALSE))
t18 <- xtabs(~Major.Donation.Level + FY16Giving.Grouped, data = givings)
round(t18/rowSums(t18), 2)
```

```
##                     FY16Giving.Grouped
## Major.Donation.Level [0-1) [1-100) [100-250) [250-500) [500-200000)
##               NO      0.72    0.15      0.11      0.02         0.00
##               Low     0.67    0.16      0.08      0.06         0.03
##               Medium  0.58    0.18      0.15      0.03         0.07
##               High    0.48    0.13      0.22      0.10         0.08
```

```r
givings$High.Donor.Major <- ifelse(givings$Major %in% c("History", "Psychology",
    "Biology", "Economics"), TRUE, FALSE)
```

### 2.4.5 FY16Giving.Grouped vs. Next.Degree

A higher proportion of people with above_bachelor degree make top donations ($500 or more), compared to other groups.

```r
t12 <- xtabs(~Adv.Deg + FY16Giving.Grouped, data = givings)
round(t12/rowSums(t12), 2)
```

```
##                      FY16Giving.Grouped
## Adv.Deg                [0-1) [1-100) [100-250) [250-500) [500-200000)
##   bachelor_equivalent  0.57    0.22      0.12      0.03         0.05
##   above_bachelor       0.49    0.21      0.17      0.04         0.09
##   NONE                 0.72    0.11      0.11      0.04         0.02
```

### 2.4.6 FY16Giving.Grouped vs. AttendenceEvent

The data is inline with our expectations. Among the people who donate, there is a strong correlation between attendence and donations. In fact, most of the top donors (52 out of 59, 85%) have attended an Alumni event.

```r
(t4 <- xtabs(~AttendenceEvent + FY16Giving.Grouped, data = givings))
```

```
##                  FY16Giving.Grouped
## AttendenceEvent [0-1) [1-100) [100-250) [250-500) [500-200000)
##   Didn't Attend   286      61        36         5            7
##   Attended        300     112       107        34           52
```

### 2.4.7 FY16Giving.Grouped vs. previous years' Donation levels

To analyze the relationship between FY16 donation and previous years' donations, we plotted previous years donations against the FY16 donations. The x-axis is the previous donation(FY12Giving to FY15 Giving), the y-axis is the probability of 2016 donations in each level [0, 1), [1, 100), etc.

We noticed that the probability that 2016 donations located in [0,1)("red line") decrease as previous years' donation decreases. The probability that 2016 donations fall in the in [1,100),[100,250),[250,500) range first increases with previous years' donation levels and then decreases.

In addition, the probability that the 2016 donation falls in the [500,200000) range is first unchanged with the previous years' donation but increases quickly once the previous years' level exceeds $350. Taken together, these observations suggest that there is a clear relationship between previous years' donation levels and the target variable. However, relationship between them doesn't seem to be linear. The following contingency tables underscore this point, and clearly show how 2016 donations are highly related to previous years' donation level.

Considering the high correlation between the previous years' donation level and 2016 donation level, we wanted to carefully check whether we should include donations from all previous years in the model or whether the most recent year (2015) would suffice. For example, if an alumnus made consecutively donations in 2014 and 2015, is he more likely to donate in 2016 than an alumnus only donate in 2015?

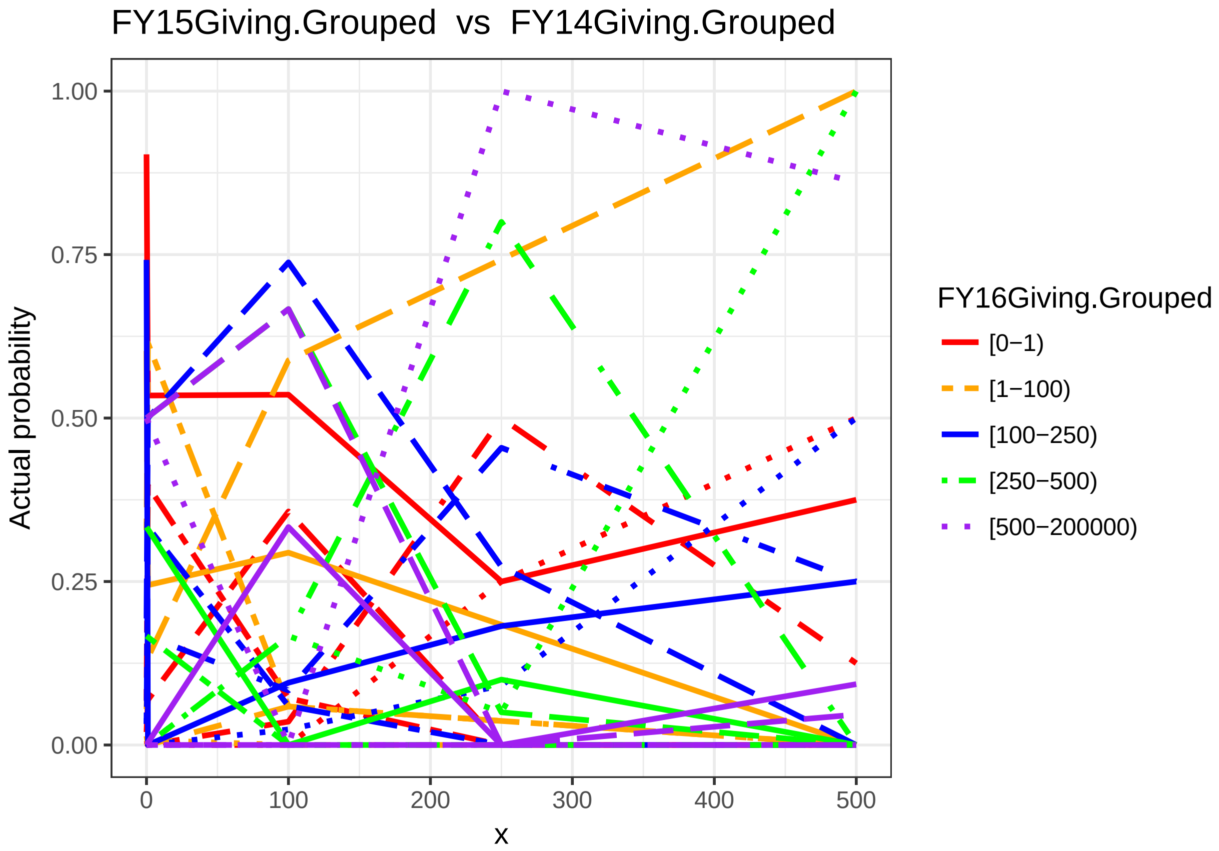
The following plot shows the effect of 2015 donations and 2014 donations on 2016 donations. The x-axis is 2015 donations. The color shows 2014 donation while the line type shows 2016 donations (different donation levels). The contingency table shows the same information. For example, the actual probability that 2016 donation locates in [0,1) is 0.9 if the alumni neither donated in 2014 nor in 2015. This percentage decreased to 0.6 if this person donated in 2014 even if he didn't donate in 2015. The situation is similar for other year combinations. From both the plot and the contingency table, it seems we should include all of the previous years while estimating our model because every year does provide additional information to make our prediction.

**add jill's graph - added - must check if it is what was intended**

```r
FY_analysis <- function(x, y) {
    groups <- aggregate(givings$FY16Giving.Grouped, by = list(givings[, x],
        givings[, y], givings$FY16Giving.Grouped), "length")
    colnames(groups) <- c("x", "y", "FY16Giving.Grouped", "cnt")
    # head(groups15_14)
    groups_cast <- cast(groups, x + y ~ FY16Giving.Grouped, value = "cnt")

    colnames(groups_cast)[c(1, 2)] <- c("x", "y")
    output <- cbind(groups_cast[, c(1, 2)], groups_cast[, c(3, 4, 5, 6, 7)]/rowSums(groups_cast[,
        c(3, 4, 5, 6, 7)], na.rm = TRUE))
    output[is.na(output)] <- 0
    return(output)
}
output <- FY_analysis("FY15Giving.Grouped", "FY14Giving.Grouped")
melt_output <- melt(output, id = c("x", "y"))
levels(melt_output[, 1]) <- c(0, 1, 100, 250, 500)
melt_output[, 1] <- as.numeric(levels(melt_output[, 1]))[melt_output[, 1]]
```

```r
ggplot(melt_output, aes(x, value, colour = y, linetype = variable)) + geom_line(lwd = 1) +
    theme_bw() + ggtitle(paste("FY15Giving.Grouped", " vs ", "FY14Giving.Grouped")) +
    ylab("Actual probability") + scale_linetype_manual(values = c("solid", "twodash",
    "longdash", "12345678", "dotted")) + scale_color_manual(values = c("red",
    "orange", "blue", "green", "purple")) + guides(color = guide_legend(title = "FY16Giving.Grouped"),
    linetype = guide_legend(title = "FY16Giving.Grouped"))
```



```r
colnames(output)[1:2] <- c("FY15Giving.Grouped", "FY14Giving.Grouped")
head(cbind(output[, 1:2], round(output[, 3:7], 2)), 4)
```

```
##   FY15Giving.Grouped FY14Giving.Grouped [0-1) [1-100) [100-250) [250-500)
## 1              [0-1)              [0-1)  0.90    0.07      0.02      0.00
## 2              [0-1)            [1-100)  0.60    0.33      0.05      0.01
## 3              [0-1)          [100-250)  0.74    0.03      0.19      0.03
## 4              [0-1)          [250-500)  0.33    0.17      0.50      0.00
##   [500-200000)
## 1         0.01
## 2         0.00
## 3         0.00
## 4         0.00
```

```r
xtabs(~FY16Giving.Grouped + FY12Giving.Grouped, data = givings)
```

```
##                   FY12Giving.Grouped
## FY16Giving.Grouped [0-1) [1-100) [100-250) [250-500) [500-200000)
##       [0-1)          462      73        38         9            4
##       [1-100)          60      96        16          0             1
##       [100-250)        26      40        69          5             3
##       [250-500)         4       2        16         16             1
##       [500-200000)      6       2        10          7            34
```

```
xtabs(~FY16Giving.Grouped + FY13Giving.Grouped, data = givings)
```

```
##                   FY13Giving.Grouped
## FY16Giving.Grouped [0-1) [1-100) [100-250) [250-500) [500-200000)
##       [0-1)          441      94        40          7             4
##       [1-100)         39     123        10          1             0
##       [100-250)       19      27        73         19             5
##       [250-500)        5       0        13         18             3
##       [500-200000)     9       3         7          9            31
```

```
xtabs(~FY16Giving.Grouped + FY14Giving.Grouped, data = givings)
```

```
##                   FY14Giving.Grouped
## FY16Giving.Grouped [0-1) [1-100) [100-250) [250-500) [500-200000)
##       [0-1)          461      82        34          4             5
##       [1-100)         56     108         8          1             0
##       [100-250)       23      33        74          8             5
##       [250-500)        5       2        15         17             0
##       [500-200000)     8       1         5          6            39
```

```
xtabs(~FY16Giving.Grouped + FY15Giving.Grouped, data = givings)
```

```
##                   FY15Giving.Grouped
## FY16Giving.Grouped [0-1) [1-100) [100-250) [250-500) [500-200000)
##       [0-1)          480      64        29          5             8
##       [1-100)         57     108         8          0             0
##       [100-250)       23      25        88          4             3
##       [250-500)        3       1        10         23             2
##       [500-200000)     4       1         3          4            47
```

#### 2.4.8 Gender vs. Class.Year

As expected, over the years, the gender ratio converges towards a gender neutral 50%, but in the earlier years males were a higher percentage of the sample. It is also worth noting that there is an unexpected change in the ratio for the class of 2002. We will explore the Gender:Class.Year interaction in the next section of the EDA.

```
t6 <- xtabs(~Class.Year + Gender, data = givings)
t(t6)
```

```
##         Class.Year
## Gender   1972 1982 1992 2002 2012
##   Female   38   80  102  133  152
##   Male     67   96  101   90  141
```

```
t(round(t6/rowSums(t6), 2))
```

```
##         Class.Year
## Gender   1972 1982 1992 2002 2012
##   Female 0.36 0.45 0.50 0.60 0.52
##   Male   0.64 0.55 0.50 0.40 0.48
```

### 2.4.9 Gender vs. Marital.Status

We previously observed that that Married and Single alumni were more likely to donate than divorced or widowed alumni, however, the vast majority of sample consists of Married and Single alumni. We anticipate that this will weaken the predictive power of the Marital.Status variable. We note here that strong skew in widow ratio can be explained by the life expectancy differences between men and women.

```
t7 <- xtabs(~Marital.Status + Gender, data = givings)
t(t7)
```

```
##         Marital.Status
## Gender   Divorced Married Single Widowed
##   Female       37     282    178       8
##   Male         24     302    166       3
```

```
t(round(t7/rowSums(t7), 2))
```

```
##         Marital.Status
## Gender   Divorced Married Single Widowed
##   Female     0.61    0.48   0.52    0.73
##   Male       0.39    0.52   0.48    0.27
```

### 2.4.10 Gender vs. Major

Here, we explored the relationship between high/medium/low donation major groups and gender

We had already established that among the high level donors, men had a higher ratio than women. We now conclude that this is also reflected for Majors. Majors that on average had lower previous year donations had a higher percentage of females than males and majors that have the highest donation levels have more males than females.

```
t8 <- xtabs(~Major.Donation.Level + Gender, data = givings)
t(t8)
```

```
##         Major.Donation.Level
## Gender    NO Low Medium High
##   Female  29  79    373   24
##   Male    25  38    364   68
```

```
t(round(t8/rowSums(t8), 2))
```

```
##         Major.Donation.Level
## Gender     NO  Low Medium High
##   Female 0.54 0.68   0.51 0.26
##   Male   0.46 0.32   0.49 0.74
```

### 2.4.11 Major vs. Next.Degree

skipping for now

### 2.4.12 Major vs. AttendenceEvent

skipping for now

```
t9 <- xtabs(~Major.Donation.Level + AttendenceEvent, data = givings)
t(t9)
```

```
##                 Major.Donation.Level
## AttendenceEvent  NO Low Medium High
##   Didn't Attend  31  53    281   30
##   Attended       23  64    456   62
```

```
t(round(t9/rowSums(t9), 2))
```

```
##                 Major.Donation.Level
## AttendenceEvent   NO  Low Medium High
##   Didn't Attend 0.57 0.45   0.38 0.33
##   Attended      0.43 0.55   0.62 0.67
```

### 2.4.13 Class.Year vs. AttendenceEvent

It is remarkable that graduates from 1972 have the same (~60%) attendance rate as the class of 2012. Of note, the Class of 2002 has an usual spike in attendance, but we cannot explain it with the information available to us.

```
t10 <- xtabs(~Class.Year + AttendenceEvent, data = givings)
t(t10)
```

```
##                 Class.Year
## AttendenceEvent 1972 1982 1992 2002 2012
##   Didn't Attend   41   83   86   65  120
##   Attended        64   93  117  158  173
```

```
t(round(t10/rowSums(t10), 2))
```

```
##                 Class.Year
## AttendenceEvent 1972 1982 1992 2002 2012
##   Didn't Attend 0.39 0.47 0.42 0.29 0.41
##   Attended      0.61 0.53 0.58 0.71 0.59
```

## 2.5 Interactions

### 2.5.1 Gender, Class.Year, 2016 Donations

```
t11 <- xtabs(~Class.Year + FY16Giving.Grouped + Gender, data = givings)
t11
```

```
## , , Gender = Female
##
##            FY16Giving.Grouped
## Class.Year [0-1) [1-100) [100-250) [250-500) [500-200000)
##       1972    21       3         4         4            6
##       1982    38      13        18         6            5
##       1992    61      17        15         3            6
##       2002    73      33        15         3            9
##       2012   105      40         6         1            0
##
## , , Gender = Male
##
##            FY16Giving.Grouped
## Class.Year [0-1) [1-100) [100-250) [250-500) [500-200000)
##       1972    29       6        19         3           10
##        1982    52       9         17          8             10
##        1992    54      12         23          6              6
##        2002    64       8         10          3              5
##        2012    89      32         16          2              2
```

The difference in top donations (above $500) can be explained by male / female ratio. For example, 6 women made $500+ donations from the class of 72, vs. 10 man. However, their ratio (6/10) is not too far from the female/male ratio (0.56) for the class of FY72. We are not anticipating a strong interaction between Gender, Class.Year, and 2016 Donation levels.

### 2.5.2 Gender, Major, 2016 donations

```
t11 <- xtabs(~Major.Donation.Level + FY16Giving.Grouped + Gender, data = givings)
t11
```

```
## , , Gender = Female
## 
##                     FY16Giving.Grouped
## Major.Donation.Level [0-1) [1-100) [100-250) [250-500) [500-200000)
##               NO        23       5         1         0            0
##               Low       54      15         2         4            4
##               Medium   212      80        49        10           22
##               High       9       6         6         3            0
## 
## , , Gender = Male
## 
##                     FY16Giving.Grouped
## Major.Donation.Level [0-1) [1-100) [100-250) [250-500) [500-200000)
##               NO        16       3         5         1            0
##               Low       24       4         7         3            0
##               Medium   213      54        59        12           26
##               High      35       6        14         6            7
```

Looking at the top donors from Majors that have usually high donation levels, we see that they are more likely to be male. We know that top donor Majors (Major.Donation.Level == High) have a 3-to-1 Male/Female ratio. But we observe a 7-to-0 ration for Male/Female distribution for donors who have donated above $500 and are from top donor majors. **So we we believe there may be an interaction between Gender, Major, 2016 donations**

## 3 Statistical Modeling

Start the section summarizing the key results - what variables, if any, are the key predictors of the year 2016 contribution? What are the key techniques you have experimented? What method did you use in your final model? How did you choose the final model? What model performance criteria did you use to choose the final model? What statistical infernece did you perform? Explain them. Comment on statistical significance vs. economic significance.

-list variables included

-latex final model

a) Our prefered model is:

$$logit(\hat{P}(Y \leq j)) = 0.000 - 0.000\ age + i'mfinishingthisnow$$

-comparison ordinal nominal

-how chose final model:

-performance criteria

-statistical inference we performed

-stats significance versus economic significance

I don't know what to do with this refactoring of variables below - somehow justify using them in univariate analysis?

```
givings$Married <- factor(ifelse(givings$Marital.Status == "Married", TRUE,
    FALSE))
givings$Class.Year = as.numeric(givings$Class.Year)
```

```
model_B1a <- clm(formula = FY16Giving.Grouped ~ FY15Giving.Grouped, data = givings,
    link = "logit")
Anova(model_B1a)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: FY16Giving.Grouped
##                    Df  Chisq Pr(>Chisq)
## FY15Giving.Grouped  4 628.74  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# add gender
model_B1b <- clm(formula = FY16Giving.Grouped ~ FY15Giving.Grouped + Class.Year,
    data = givings, link = "logit")
Anova(model_B1b)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: FY16Giving.Grouped
##                    Df  Chisq Pr(>Chisq)
## FY15Giving.Grouped  4 459.44  < 2.2e-16 ***
## Class.Year          1 139.48  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(model_B1a, model_B1b)
```

```
## Likelihood ratio tests of cumulative link models:
##
##           formula:                                           link:
## model_B1a FY16Giving.Grouped ~ FY15Giving.Grouped              logit
## model_B1b FY16Giving.Grouped ~ FY15Giving.Grouped + Class.Year logit
##           threshold:
## model_B1a flexible
## model_B1b flexible
##
##           no.par    AIC  logLik LR.stat df Pr(>Chisq)
## model_B1a      8 1705.7 -844.85
## model_B1b      9 1707.7 -844.85  0.0087  1     0.9257
```

```r
# add marital status
model_B1c <- clm(formula = FY16Giving.Grouped ~ FY15Giving.Grouped + Marital.Status,
    data = givings, link = "logit")
Anova(model_B1c)
```

```
## Analysis of Deviance Table (Type II tests)
## 
## Response: FY16Giving.Grouped
##                     Df  Chisq Pr(>Chisq)    
## FY15Giving.Grouped  4 431.40  < 2.2e-16 ***
## Marital.Status      3 352.16  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
anova(model_B1a, model_B1c)
```

```
## Likelihood ratio tests of cumulative link models:
## 
##           formula:                                                link:
## model_B1a FY16Giving.Grouped ~ FY15Giving.Grouped                  logit
## model_B1c FY16Giving.Grouped ~ FY15Giving.Grouped + Marital.Status logit
##           threshold:
## model_B1a flexible  
## model_B1c flexible  
## 
##           no.par    AIC  logLik LR.stat df Pr(>Chisq)
## model_B1a      8 1705.7 -844.85                      
## model_B1c     11 1708.5 -843.26  3.1832  3     0.3642
```

```r
# add attend event
model_B1d <- clm(formula = FY16Giving.Grouped ~ FY15Giving.Grouped + AttendenceEvent,
    data = givings, link = "logit")
Anova(model_B1d)
```

```
## Analysis of Deviance Table (Type II tests)
## 
## Response: FY16Giving.Grouped
##                     Df  Chisq Pr(>Chisq)    
## FY15Giving.Grouped  4 573.37  < 2.2e-16 ***
## AttendenceEvent     1 126.66  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
anova(model_B1a, model_B1d)
```

```
## Likelihood ratio tests of cumulative link models:
## 
##           formula:                                                 link:
## model_B1a FY16Giving.Grouped ~ FY15Giving.Grouped                   logit
## model_B1d FY16Giving.Grouped ~ FY15Giving.Grouped + AttendenceEvent logit
##           threshold:
## model_B1a flexible  
## model_B1d flexible  
## 
##           no.par    AIC  logLik LR.stat df Pr(>Chisq)
## model_B1a      8 1705.7 -844.85                      
## model_B1d      9 1697.7 -839.86  9.9832  1    0.00158 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# add major
model_B1e <- clm(formula = FY16Giving.Grouped ~ FY15Giving.Grouped + Major.Donation.Level,
    data = givings, link = "logit")
Anova(model_B1e)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: FY16Giving.Grouped
##                       Df  Chisq Pr(>Chisq)
## FY15Giving.Grouped    4 433.25  < 2.2e-16 ***
## Major.Donation.Level  3 365.85  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
anova(model_B1a, model_B1e)
```

```
## Likelihood ratio tests of cumulative link models:
##
##           formula:
## model_B1a FY16Giving.Grouped ~ FY15Giving.Grouped
## model_B1e FY16Giving.Grouped ~ FY15Giving.Grouped + Major.Donation.Level
##           link: threshold:
## model_B1a logit flexible
## model_B1e logit flexible
##
##           no.par    AIC  logLik LR.stat df Pr(>Chisq)
## model_B1a      8 1705.7 -844.85
## model_B1e     11 1710.1 -844.04   1.635  3     0.6515
```

```r
# add advance degree
model_B1x <- clm(formula = FY16Giving.Grouped ~ FY15Giving.Grouped + Adv.Deg,
    data = givings, link = "logit")
Anova(model_B1x)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: FY16Giving.Grouped
##                     Df  Chisq Pr(>Chisq)
## FY15Giving.Grouped  4 454.26  < 2.2e-16 ***
## Adv.Deg             2 274.99  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
anova(model_B1a, model_B1x)
```

```
## Likelihood ratio tests of cumulative link models:
##
##           formula:                                          link:
## model_B1a FY16Giving.Grouped ~ FY15Giving.Grouped           logit
## model_B1x FY16Giving.Grouped ~ FY15Giving.Grouped + Adv.Deg logit
##           threshold:
## model_B1a flexible
## model_B1x flexible
```

```
##
##           no.par    AIC  logLik LR.stat df Pr(>Chisq)
## model_B1a      8 1705.7 -844.85
## model_B1x     10 1693.3 -836.67  16.374  2  0.0002783 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# add gender
model_B1y <- clm(formula = FY16Giving.Grouped ~ FY15Giving.Grouped + Gender,
    data = givings, link = "logit")
Anova(model_B1y)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: FY16Giving.Grouped
##                     Df  Chisq Pr(>Chisq)
## FY15Giving.Grouped  4 571.86  < 2.2e-16 ***
## Gender              1 139.66  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(model_B1a, model_B1y)
```

```
## Likelihood ratio tests of cumulative link models:
##
##           formula:                                         link:
## model_B1a FY16Giving.Grouped ~ FY15Giving.Grouped          logit
## model_B1y FY16Giving.Grouped ~ FY15Giving.Grouped + Gender logit
##           threshold:
## model_B1a flexible
## model_B1y flexible
##
##           no.par    AIC  logLik LR.stat df Pr(>Chisq)
## model_B1a      8 1705.7 -844.85
## model_B1y      9 1707.6 -844.81  0.0799  1     0.7774
```

```
# add all above variables but not previous years' donation
model_B1f <- clm(formula = FY16Giving.Grouped ~ Class.Year + Marital.Status +
    AttendenceEvent + FY15Giving.Grouped + Major.Donation.Level + Adv.Deg +
    Gender, data = givings, link = "logit")
Anova(model_B1f)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: FY16Giving.Grouped
##                       Df    Chisq Pr(>Chisq)
## Class.Year            1  12.7142  0.0003629 ***
## Marital.Status        3 233.8733  < 2.2e-16 ***
## AttendenceEvent       1   0.2194  0.6395194
## FY15Giving.Grouped    4  10.2087  0.0370550 *
## Major.Donation.Level  3 312.6503  < 2.2e-16 ***
## Adv.Deg               2 275.6507  < 2.2e-16 ***
## Gender                1   0.0469  0.8284998
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
anova(model_B1a, model_B1f)
```

```
## Likelihood ratio tests of cumulative link models:
##
##           formula:
## model_B1a FY16Giving.Grouped ~ FY15Giving.Grouped
## model_B1f FY16Giving.Grouped ~ Class.Year + Marital.Status + AttendenceEvent + FY15Giving.Grouped + I
##           link: threshold:
## model_B1a logit flexible
## model_B1f logit flexible
##
##           no.par    AIC  logLik LR.stat df Pr(>Chisq)
## model_B1a      8 1705.7 -844.85
## model_B1f     19 1699.2 -830.59  28.529 11   0.002685 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# add previous years' donation
model_B1h <- clm(formula = FY16Giving.Grouped ~ FY15Giving.Grouped + FY14Giving.Grouped +
    FY13Giving.Grouped + FY12Giving.Grouped, data = givings, link = "logit")
Anova(model_B1h)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: FY16Giving.Grouped
##                    Df   Chisq Pr(>Chisq)
## FY15Giving.Grouped  4 627.036  < 2.2e-16 ***
## FY14Giving.Grouped  4 121.631  < 2.2e-16 ***
## FY13Giving.Grouped  4  20.779  0.0003503 ***
## FY12Giving.Grouped  4  31.991  1.921e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
anova(model_B1a, model_B1h)
```

```
## Likelihood ratio tests of cumulative link models:
##
##           formula:
## model_B1a FY16Giving.Grouped ~ FY15Giving.Grouped
## model_B1h FY16Giving.Grouped ~ FY15Giving.Grouped + FY14Giving.Grouped + FY13Giving.Grouped + FY12Gi
##           link: threshold:
## model_B1a logit flexible
## model_B1h logit flexible
##
##           no.par    AIC  logLik LR.stat df Pr(>Chisq)
## model_B1a      8 1705.7 -844.85
## model_B1h     20 1558.0 -758.98  171.74 12  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# add all variables
model_B1g <- clm(formula = FY16Giving.Grouped ~ Class.Year + Marital.Status +
    AttendenceEvent + FY15Giving.Grouped + Major.Donation.Level + Adv.Deg +
    Gender + FY14Giving.Grouped + FY13Giving.Grouped + FY12Giving.Grouped, data = givings,
    link = "logit")
Anova(model_B1g)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: FY16Giving.Grouped
##                     Df    Chisq Pr(>Chisq)
## Class.Year           1  27.3727  1.678e-07 ***
## Marital.Status       3 260.7304  < 2.2e-16 ***
## AttendenceEvent      1   1.9970  0.1576110
## FY15Giving.Grouped   4   1.9664  0.7419301
## Major.Donation.Level  3  89.4522  < 2.2e-16 ***
## Adv.Deg              2  66.4901  3.646e-15 ***
## Gender               1   0.0490  0.8248218
## FY14Giving.Grouped   4   9.6577  0.0466057 *
## FY13Giving.Grouped   4  20.4974  0.0003982 ***
## FY12Giving.Grouped   4  30.5314  3.815e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(model_B1h, model_B1g)
```

```
## Likelihood ratio tests of cumulative link models:
##
##           formula:
## model_B1h FY16Giving.Grouped ~ FY15Giving.Grouped + FY14Giving.Grouped + FY13Giving.Grouped + FY12Gi
## model_B1g FY16Giving.Grouped ~ Class.Year + Marital.Status + AttendenceEvent + FY15Giving.Grouped + 
##           link: threshold:
## model_B1h logit flexible
## model_B1g logit flexible
##
##           no.par  AIC  logLik LR.stat df Pr(>Chisq)
## model_B1h     20 1558 -758.98
## model_B1g     31 1564 -751.02  15.932 11     0.1437
```

```
# add all variables
model_B1j <- clm(formula = FY16Giving.Grouped ~ Married + AttendenceEvent +
    FY15Giving.Grouped + Major.Donation.Level + Adv.Deg + Gender + FY14Giving.Grouped +
    FY13Giving.Grouped + FY12Giving.Grouped, data = givings, link = "logit")
summary(model_B1j)
```

```
## formula:
## FY16Giving.Grouped ~ Married + AttendenceEvent + FY15Giving.Grouped + Major.Donation.Level + Adv.Deg
## data:    givings
##
##  link  threshold nobs logLik  AIC     niter max.grad cond.H
##  logit flexible  1000 -752.37 1560.73 6(0)  3.76e-09 7.0e+02
##
## Coefficients:
##                                Estimate Std. Error z value Pr(>|z|)
## MarriedTRUE                    -0.01833    0.16249  -0.113 0.910196
## AttendenceEventAttended         0.25573    0.16572   1.543 0.122789
## FY15Giving.Grouped[1-100)       1.21163    0.20930   5.789 7.08e-09 ***
## FY15Giving.Grouped[100-250)     2.16673    0.26704   8.114 4.90e-16 ***
## FY15Giving.Grouped[250-500)     3.16498    0.51755   6.115 9.63e-10 ***
## FY15Giving.Grouped[500-200000)  4.79923    0.59942   8.006 1.18e-15 ***
## Major.Donation.LevelLow        -0.06911    0.42464  -0.163 0.870723
## Major.Donation.LevelMedium      0.08056    0.36567   0.220 0.825624
## Major.Donation.LevelHigh        0.32846    0.42965   0.764 0.444579
## Adv.Degabove_bachelor          -0.23019    0.22914  -1.005 0.315095
## Adv.DegNONE                    -0.62602    0.24713  -2.533 0.011306 *
## GenderMale                      0.10951    0.15232   0.719 0.472166
## FY14Giving.Grouped[1-100)       0.59451    0.22153   2.684 0.007282 **
## FY14Giving.Grouped[100-250)     0.83944    0.32007   2.623 0.008723 **
## FY14Giving.Grouped[250-500)     1.83479    0.50119   3.661 0.000251 ***
## FY14Giving.Grouped[500-200000)  1.63541    0.65830   2.484 0.012981 *
## FY13Giving.Grouped[1-100)       0.94894    0.22599   4.199 2.68e-05 ***
## FY13Giving.Grouped[100-250)     1.27372    0.32398   3.931 8.44e-05 ***
## FY13Giving.Grouped[250-500)     1.50705    0.42921   3.511 0.000446 ***
## FY13Giving.Grouped[500-200000)  2.04300    0.65428   3.123 0.001793 **
## FY12Giving.Grouped[1-100)       0.59192    0.21778   2.718 0.006568 **
## FY12Giving.Grouped[100-250)     0.74112    0.28782   2.575 0.010025 *
## FY12Giving.Grouped[250-500)     0.88114    0.46333   1.902 0.057205 .
## FY12Giving.Grouped[500-200000)  1.99772    0.65422   3.054 0.002261 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Threshold coefficients:
##                        Estimate Std. Error z value
## [0-1)|[1-100)            2.4071     0.4353   5.530
## [1-100)|[100-250)        4.0524     0.4513   8.979
## [100-250)|[250-500)      6.4884     0.4904  13.230
## [250-500)|[500-200000)   7.8717     0.5336  14.751
```

```
Anova(model_B1j)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: FY16Giving.Grouped
##                      Df    Chisq Pr(>Chisq)
## Married               1  30.5769  3.209e-08 ***
## AttendenceEvent       1  80.6232  < 2.2e-16 ***
## FY15Giving.Grouped    4 221.0286  < 2.2e-16 ***
## Major.Donation.Level  3  88.2277  < 2.2e-16 ***
## Adv.Deg               2  64.2462  1.120e-14 ***
## Gender                1   0.0485  0.8256239
## FY14Giving.Grouped    4   9.0404  0.0600968 .
## FY13Giving.Grouped    4  20.3539  0.0004251 ***
## FY12Giving.Grouped    4  30.5306  3.816e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(model_B1j, model_B1g)
```

```
## Likelihood ratio tests of cumulative link models:
##
##           formula:
## model_B1j FY16Giving.Grouped ~ Married + AttendenceEvent + FY15Giving.Grouped + Major.Donation.Level
## model_B1g FY16Giving.Grouped ~ Class.Year + Marital.Status + AttendenceEvent + FY15Giving.Grouped +
##           link: threshold:
## model_B1j logit flexible
## model_B1g logit flexible
##
```

```
##           no.par    AIC  logLik LR.stat df Pr(>Chisq)
## model_B1j     28 1560.7 -752.37
## model_B1g     31 1564.0 -751.02  2.7017  3     0.4399
```

```
conf.beta <- confint(object = model_B1j, level = 0.95)
conf.beta <- as.data.frame.matrix(conf.beta)
conf.beta[, 3] <- model_B1j$coefficients[5:28]
colnames(conf.beta)[3] = "estimated"
round(conf.beta, 2)
```

```
##                                2.5 % 97.5 % estimated
## MarriedTRUE                    -0.34   0.30     -0.02
## AttendenceEventAttended        -0.07   0.58      0.26
## FY15Giving.Grouped[1-100)       0.80   1.62      1.21
## FY15Giving.Grouped[100-250)     1.65   2.69      2.17
## FY15Giving.Grouped[250-500)     2.15   4.18      3.16
## FY15Giving.Grouped[500-200000)  3.65   6.01      4.80
## Major.Donation.LevelLow        -0.89   0.78     -0.07
## Major.Donation.LevelMedium     -0.62   0.82      0.08
## Major.Donation.LevelHigh       -0.50   1.19      0.33
## Adv.Degabove_bachelor          -0.68   0.22     -0.23
## Adv.DegNONE                    -1.11  -0.14     -0.63
## GenderMale                     -0.19   0.41      0.11
## FY14Giving.Grouped[1-100)       0.16   1.03      0.59
## FY14Giving.Grouped[100-250)     0.21   1.47      0.84
## FY14Giving.Grouped[250-500)     0.85   2.82      1.83
## FY14Giving.Grouped[500-200000)  0.33   2.93      1.64
## FY13Giving.Grouped[1-100)       0.51   1.39      0.95
## FY13Giving.Grouped[100-250)     0.64   1.91      1.27
## FY13Giving.Grouped[250-500)     0.66   2.35      1.51
## FY13Giving.Grouped[500-200000)  0.75   3.33      2.04
## FY12Giving.Grouped[1-100)       0.16   1.02      0.59
## FY12Giving.Grouped[100-250)     0.17   1.30      0.74
## FY12Giving.Grouped[250-500)    -0.03   1.79      0.88
## FY12Giving.Grouped[500-200000)  0.71   3.29      2.00
```

```
round(exp(conf.beta), 2)
```

```
##                                2.5 % 97.5 % estimated
## MarriedTRUE                     0.71   1.35      0.98
## AttendenceEventAttended         0.93   1.79      1.29
## FY15Giving.Grouped[1-100)       2.23   5.07      3.36
## FY15Giving.Grouped[100-250)     5.18  14.78      8.73
## FY15Giving.Grouped[250-500)     8.56  65.28     23.69
## FY15Giving.Grouped[500-200000) 38.58 407.02    121.42
## Major.Donation.LevelLow         0.41   2.18      0.93
## Major.Donation.LevelMedium      0.54   2.28      1.08
## Major.Donation.LevelHigh        0.61   3.28      1.39
## Adv.Degabove_bachelor           0.51   1.25      0.79
## Adv.DegNONE                     0.33   0.87      0.53
## GenderMale                      0.83   1.50      1.12
## FY14Giving.Grouped[1-100)       1.17   2.80      1.81
## FY14Giving.Grouped[100-250)     1.23   4.33      2.32
## FY14Giving.Grouped[250-500)     2.33  16.69      6.26
## FY14Giving.Grouped[500-200000)  1.40  18.74      5.13
## FY13Giving.Grouped[1-100)         1.66   4.03      2.58
## FY13Giving.Grouped[100-250)       1.89   6.75      3.57
## FY13Giving.Grouped[250-500)       1.94  10.49      4.51
## FY13Giving.Grouped[500-200000)  2.11  27.92      7.71
## FY12Giving.Grouped[1-100)         1.18   2.77      1.81
## FY12Giving.Grouped[100-250)       1.19   3.68      2.10
## FY12Giving.Grouped[250-500)       0.97   5.97      2.41
## FY12Giving.Grouped[500-200000)  2.03  26.93      7.37
```

# 3 Should we chose an Ordinal or Nominal Multinomial model?

```
c1 <- xtabs(~Gender + FY16Giving.Grouped, data = givings)
c2 <- xtabs(~Marital.Status + FY16Giving.Grouped, data = givings)
c3 <- xtabs(~Class.Year + FY16Giving.Grouped, data = givings)
c4 <- xtabs(~Major.Donation.Level + FY16Giving.Grouped, data = givings)
c5 <- xtabs(~AttendenceEvent + FY16Giving.Grouped, data = givings)

odds_ratio <- function(r1) {
    df1 <- as.data.frame.matrix(r1)
    n <- dim(df1)[1]
    len <- dim(df1)[2]
    odds = data.frame(matrix(0, n, len - 1))
    colnames(odds) <- colnames(df1)[1:len - 1]

    for (i in seq(1, len - 1)) {
        if (i == 1) {
            lowerp <- df1[, 1]
        } else {
            lowerp <- rowSums(df1[, 1:i])
        }
        if (i == len - 1) {
            upperp <- df1[, len]
        } else {
            upperp <- rowSums(df1[, (i + 1):len])
        }
        odds[, i] <- lowerp/upperp
    }
    round(odds, 2)

    oratio <- data.frame(matrix(0, n - 1, len - 1), row.names = rownames(df1)[2:n])
    colnames(oratio) <- colnames(odds)
    for (j in seq(1, n - 1)) oratio[j, ] <- odds[j + 1, ]/odds[j, ]
    return(round(oratio, 2))
}
# Gender
odds_ratio(c1)
```

```
##      [0-1) [1-100) [100-250) [250-500)
## Male  0.97    0.63      0.74      0.76
```

```
# Marital Status
odds_ratio(c2)
```

```
##          [0-1) [1-100) [100-250) [250-500)
## Married  0.76    0.78      0.62      0.65
## Single   2.14    3.79      3.42      2.65
## Widowed  0.24    0.14      0.05      0.08
```

```
# Graduating class
odds_ratio(c3)
```

```
##   [0-1) [1-100) [100-250) [250-500)
## 2  1.15    1.36      1.42      1.93
## 3  1.25    1.39      1.71      1.48
## 4  1.22    1.62      1.17      0.94
## 5  1.23    2.49      5.67      9.75
```

```
# Major
odds_ratio(c4)
```

```
##        [0-1) [1-100) [100-250) [250-500)
## Low     0.77    0.72      0.18      0.00
## Medium  0.68    0.65      0.99      0.51
## High    0.67    0.50      0.50      0.85
```

```
# Addent Event
odds_ratio(c5)
```

```
##           [0-1) [1-100) [100-250) [250-500)
## Attended  0.37     0.3      0.19      0.19
```

# ORDINAL REGRESSION

The estimated model is:

$$logit(\hat{P}(Y \le j)) = \hat{\beta}_{j0} + 0.07Married + 0.27Attended + 1.21FY15Giving ... + 2.10FY12Giving.Grouped$$

where $\hat{beta}_{10} = 2.73, \hat{beta}_{20} = 4.38, \hat{beta}_{30} = 6.82, \hat{beta}_{40} = 8.21$

To fit the model, we use the model using the *clm* function below:

```
model_B1j <- clm(formula = FY16Giving.Grouped ~ Married + Adv.Deg + Gender +
    AttendenceEvent + High.Donor.Major + FY15Giving.Grouped + FY14Giving.Grouped +
    FY13Giving.Grouped + FY12Giving.Grouped, data = givings, link = "logit")
```

# INDEPENDENCE

We test for independence among the variables using the Anova function, with:

$$H_0 : \beta_2 = \beta_3 = ...\beta_22 = 0$$
$$H_a : any\ \beta_i \neq 0$$

From the test, we find that there is strong evidence that association exists for all variables except for AttendenceEvent and High.Donor.Major.

```
Anova(model_B1j)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: FY16Giving.Grouped
##                    Df    Chisq Pr(>Chisq)
## Married             1  70.5296  < 2.2e-16 ***
## Adv.Deg             2 326.0237  < 2.2e-16 ***
## Gender              1 350.7112  < 2.2e-16 ***
## AttendenceEvent     1   0.0036  0.9519374
## High.Donor.Major    1   0.9331  0.3340485
## FY15Giving.Grouped  4   9.5705  0.0483182 *
## FY14Giving.Grouped  4 117.3825  < 2.2e-16 ***
## FY13Giving.Grouped  4  19.5494  0.0006128 ***
## FY12Giving.Grouped  4  30.6527  3.604e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

# CONFIDENCE INTERVAL

We derive the confidence intervals for the Odds Ratios to determine the significance and impact of the variables within our model:

```
## model_polr = polr(formula = FY16Giving.Grouped ~ Married + Adv.Deg +
## Gender + AttendenceEvent + High.Donor.Major + FY15Giving.Grouped +
## FY14Giving.Grouped +FY13Giving.Grouped + FY12Giving.Grouped, data
## =givings, method ='logistic')

## conf.beta = confint(object = model_polr, level = 0.95) ci.ord =
## exp(conf.beta) # inverted coefficients to assess responses at or ABOVE a
## certain level round(data.frame(low = ci.ord[,2], up = ci.ord[,1]), 2)

conf.beta <- confint(object = model_B1j, level = 0.95)
conf.beta <- as.data.frame.matrix(conf.beta)
conf.beta[, 3] <- model_B1j$coefficients[5:26]
colnames(conf.beta)[3] = "estimated"
round(conf.beta, 2)
```

```
##                                2.5 % 97.5 % estimated
## MarriedTRUE                    -0.33   0.31     -0.01
## Adv.Degabove_bachelor          -0.67   0.23     -0.22
## Adv.DegNONE                    -1.09  -0.12     -0.61
## GenderMale                     -0.17   0.43      0.13
## AttendenceEventAttended        -0.06   0.59      0.26
## High.Donor.MajorTRUE           -0.30   0.34      0.02
## FY15Giving.Grouped[1-100)       0.80   1.62      1.21
## FY15Giving.Grouped[100-250)     1.66   2.71      2.18
## FY15Giving.Grouped[250-500)     2.19   4.22      3.21
## FY15Giving.Grouped[500-200000)  3.67   6.04      4.82
## FY14Giving.Grouped[1-100)       0.15   1.02      0.58
## FY14Giving.Grouped[100-250)     0.19   1.44      0.82
## FY14Giving.Grouped[250-500)     0.81   2.78      1.80
## FY14Giving.Grouped[500-200000)  0.27   2.89      1.58
## FY13Giving.Grouped[1-100)       0.52   1.40      0.96
## FY13Giving.Grouped[100-250)     0.64   1.91      1.27
## FY13Giving.Grouped[250-500)     0.66   2.35      1.50
## FY13Giving.Grouped[500-200000)  0.75   3.35      2.06
## FY12Giving.Grouped[1-100)       0.17   1.03      0.60
## FY12Giving.Grouped[100-250)     0.21   1.33      0.77
## FY12Giving.Grouped[250-500)    -0.02   1.80      0.90
## FY12Giving.Grouped[500-200000)  0.70   3.30      2.00
```

```
round(exp(conf.beta), 2)
```

```
##                                 2.5 % 97.5 % estimated
## MarriedTRUE                      0.72   1.36      0.99
## Adv.Degabove_bachelor            0.51   1.26      0.80
## Adv.DegNONE                      0.34   0.89      0.55
## GenderMale                       0.85   1.54      1.14
## AttendenceEventAttended          0.94   1.80      1.30
## High.Donor.MajorTRUE             0.74   1.40      1.02
## FY15Giving.Grouped[1-100)        2.22   5.04      3.34
## FY15Giving.Grouped[100-250)      5.26  15.02      8.86
## FY15Giving.Grouped[250-500)      8.98  67.86     24.71
## FY15Giving.Grouped[500-200000) 39.27 417.91    124.21
## FY14Giving.Grouped[1-100)        1.16   2.77      1.79
## FY14Giving.Grouped[100-250)      1.20   4.23      2.26
## FY14Giving.Grouped[250-500)      2.25  16.18      6.05
## FY14Giving.Grouped[500-200000)   1.31  17.97      4.87
## FY13Giving.Grouped[1-100)        1.68   4.06      2.61
## FY13Giving.Grouped[100-250)      1.89   6.75      3.57
## FY13Giving.Grouped[250-500)      1.94  10.45      4.50
## FY13Giving.Grouped[500-200000)   2.13  28.63      7.84
## FY12Giving.Grouped[1-100)        1.19   2.79      1.82
## FY12Giving.Grouped[100-250)      1.23   3.78      2.16
## FY12Giving.Grouped[250-500)      0.98   6.08      2.46
## FY12Giving.Grouped[500-200000)   2.02  27.19      7.40
```

The table above details the range of Odds Ratios for any given variable compared with its base, holding all other variables constant. For example, with 95% confidence* the odds of an alumni donating above a given donation bracket is .72 to 1.36 times when they are married than when they are not. The most notable trend is among the previous Fiscal Year donations, as there are increasing Odds for a donation for alumni who donated in higher donation brackets in previous years. Specifically, we see that if an alumni donated in the highest bracket (500-200000) in FY15, his/her odds of donating above a given bracket are 39 to 417 times compared to if he/she did not donate in FY15. This trend holds true for all previous Fiscal Years, although to a lesser extent the less recent the donation year.

A few addition variable interpretations:

- The odds of donating above a certain bracket are between .51 to 1.26 times as large when the alumni had a bachelor or equivalent degree versus no advanced degree. This range is reduced to .34 to .89 times as large when the alumni had above a bachelor degree compared to no advanced degree.
- The odds of donating above a certain bracket are between .85 to 1.54 times as large when the alumni was a male compared to female.
- The odds of donating above a certain bracket are between .94 and 1.80 times as large when the alumni attended at least one event compared to not attending any events.
- The odds of donating above a certain bracket are between .74 to 1.40 times as large when the alumni majored in what we categorized as a 'High Donor' major compared to other majors.

We notice that many of the variables include 1 within the 95% confidence interval. This suggests that there is not sufficient evidence to indicate that these variables increase the odds of donating. However, through our EDA and logical assessment, many of these variables can be key contributers and also add information to other variables in the model, thus we chose to keep them.

*95% confidence is defined in the frequentist perspective, meaning given 100 samples drawn from the same population, we can expect 95 of the samples to have a 95% confidence interval that contains the true population parameter.

## 4 Final Remarks

After examining the data and using the data to build a predictive model, what are your departing thoughts? What are the strengths and weaknesses in your analysis? Should the administration trust your result? Are there subsample in your sample that your model did a bad job in predicting their contribution behavior? If so, why? Are there other “things”, a wish list, that you think can be used to improve your model? If so, what are they? Perhaps you can make a suggestion to the administration to collect those information in the future.

-departing thoughts -strengths weakness in analysis -admin trust result? -subsample that model did a bad job predicting -wish list of things to improve model

I dumped this here, i’ll synthesize/streamline

For future: data we wished we had:

Wealth indicators: www.bidpal.com/identifying-major-donors-top-strategies-tools

-self-reported household income

-real estate ownership -people who own $2 million dollars or more in real estate are 17 times more likely donate to a nonprofit than the average person. Can use the individuals home address and use website such as Zillow to estimate donor’s property value.

-use information to cultivate relationships with possible future major donors (invite possible donors to attend more high-end University events)

-Apparently spending habits at charity auctions can be used to determine who might be a major donor- because it is an indication of how wealthy someone might be. Therefore if any of the University Events were are auctions, we would request the auction bidding data from the University to help identify potential major donors.

Past political giving: -past charitable giving (for instance political contributions, which is publicly accessible) are good indicators of who could become a major donor to the University. Political gifts indicate a willingness to donate, an interest in philanthropy, and the capacity to give in terms of wealth.