W271 Section 3 Lab 2

Kiersten Henderson, Jill Zhang, Hoang Phan, Daghan Altas

10/22/2017

```
library(easypackages)
packages("knitr","Hmisc","ggplot2","tidyverse","forcats", "gridExtra","reshape", "stargazer", "ordinal"
opts_chunk$set(tidy.opts=list(width.cutoff=70),tidy=TRUE)
```

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 model, we considered fundraising domain knowledge and the results of our explanatory data analysis (EDA).

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

1. We elected to use an Ordinal model as the donation brackets are clearly ordered and the odds ratios for each independent variable are similar across the various response levels.
2. After completing multiple levels of EDA, we found key trends and nuances in the data that we used for modeling
3. We compared 3 models on their alignment with our EDA, goodness of fit, and logical real-world structure, and ultimately chose a final model that exceled in all three of these aspects.

The final model we estimated is:

$$logit(\hat{P}(Y \leq j)) = \hat{\beta}_{j0} + \hat{\beta}_1 Married + \hat{\beta}_2 Attended + \hat{\beta}_3 Major.Donation.Level + \hat{\beta}_4 Adv.Deg + \hat{\beta}_5 Gender + \hat{\beta}_6 FY15Giving.Grou$$

where $\hat{beta}_{10} = 2.41, \hat{beta}_{20} = 4.05, \hat{beta}_{30} = 6.49, \hat{beta}_{40} = 7.87$

This model primarily consists of previous years' donations, as we found they were the best predictors of donating in the current year. 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, we included Advanced Degree as a 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, thus we included Event Attendence as well. In addition, we found in our EDA that Gender and Marital Status contributed to identifying the philanthropic behavior towards the University and we thus included Gender and Marital status when we estimated our model.

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$Class.Year = as.factor(givings$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

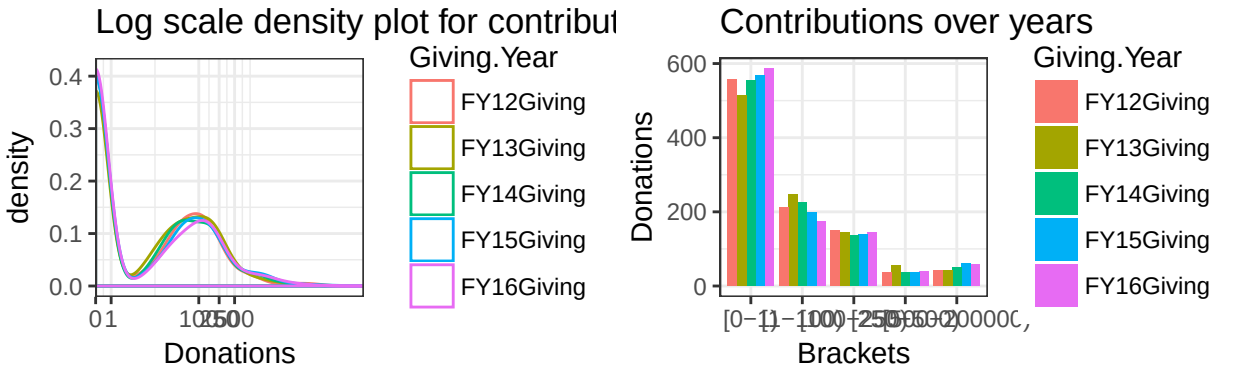
We conducted univariate analysis for every variable.

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

We notice that continuous-scale values of the contributions variable have a very strong positive skew (we didn't inlcude this plot due to page limitations). However, at log scale, we observe a bi-modal distribution, with most of the values centered either around 0 or around the $100 range. 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 is more difficult to interpret. 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))
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 con
    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(p2, p3, ncol = 2)
```



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.

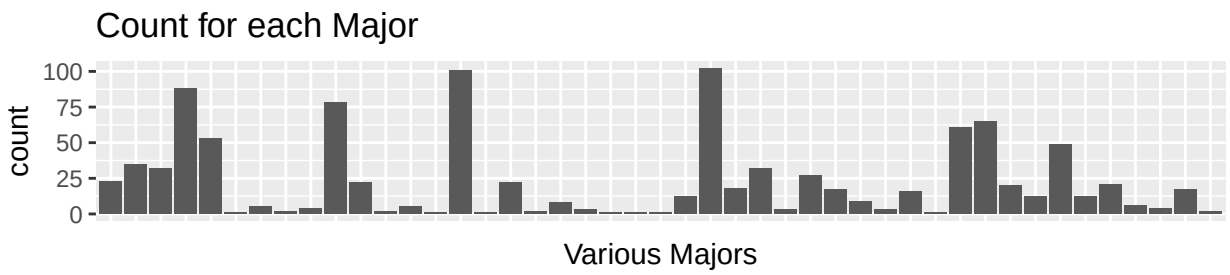
```
row <- xtabs(~Marital.Status, data = givings)
data.frame(rbind(row, row/dim(givings)[1]), row.names = c("Marital.Status Count",
    "Ratio"))
```

```
##                       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())
```



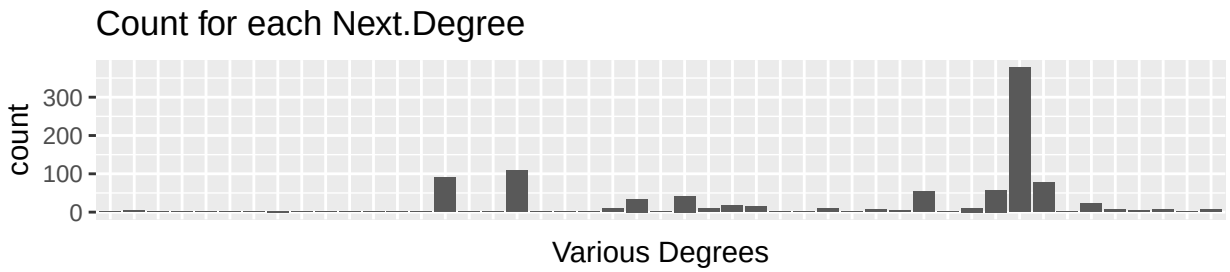

```
head(sort(xtabs(~Major, data = givings)), 3)
```

```
## 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 Next.Degree", 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))
```

```
##         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 <- round(t1/rowSums(t1), 2))
```

```
##         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))
```

```
##           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 <- round(t2/rowSums(t2), 2))
```

```
##           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))
```

```
##               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 <- round(t3/rowSums(t3), 2))
```

```
##                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:

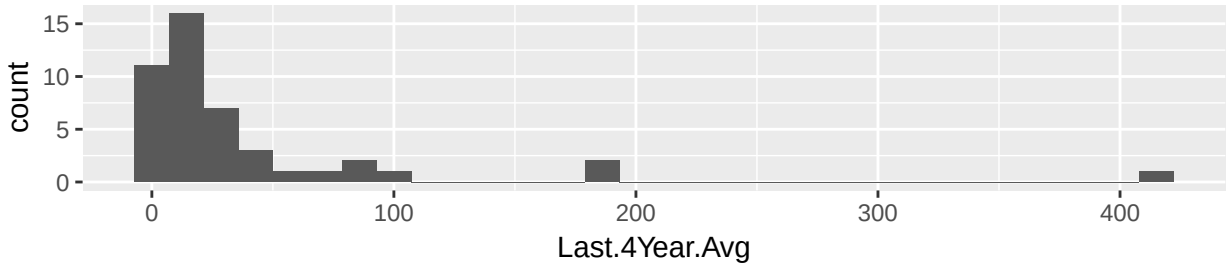
1. It will cause a curse of dimensionality that will reduce predictive power.

2. These binary variables will hold most of their records as zero and we have very little information about them.

Therefore, we've decided 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, then calculated the median donation for the last 4 years for each major.

We were very careful when deciding the median donation cutoffs used to create the new donation categories of "No","Low","Medium", and "High" donation majors. We made very granular cuts in the median donation (5 dollars increases) in order to optimize our categories. A chart of the distribution of median donations is below. 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)
```




```
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
```

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.

```
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.

```
(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

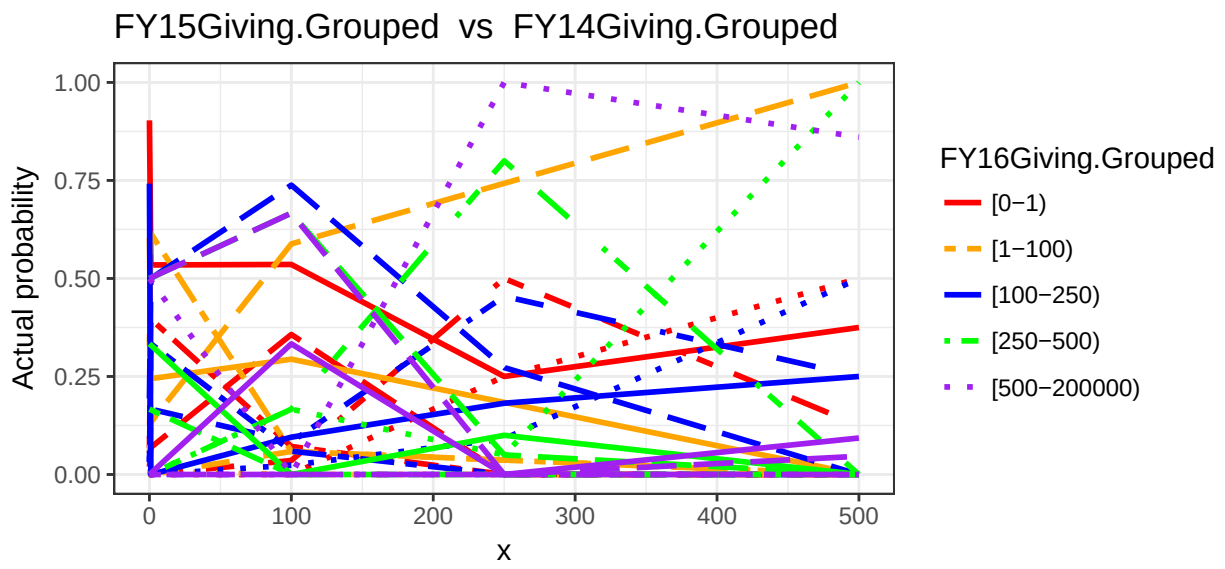
To analyze the relationship between FY16 donation and previous years' donations, we reviewed previous years donations against the FY16 donations. We noticed that the probability that 2016 donations located in [0,1) decrease as previous years' donation decreases. In addition, the probability that the 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.

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.

```
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")
    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]]
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"))
```




```
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
```

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 will create a simpler variable (Married? yes or no) to improve the parsimony of the final model. We also 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
```

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

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 (Adv.Deg)

We examined the distribution of alumni with or without next degrees based on the donation level of their major. We found that alumni who did majors characterized by high donation levels most frequently either did not have a second degree, or had a next degree that was above the bachelors level. Surprisingly, alumni who did a second bachelor's degree were the least likely to give at any donation level.

```
t99 <- xtabs(~Major.Donation.Level + Adv.Deg, data = givings)
t(t99)
```

```
##                      Major.Donation.Level
## Adv.Deg                NO Low Medium High
##   bachelor_equivalent   6  13     97    5
##   above_bachelor       18  47    390   46
##   NONE                 30  57    250   41
```

```
t(round(t99/rowSums(t99), 2))
```

```
##                      Major.Donation.Level
## Adv.Deg                 NO  Low Medium High
##   bachelor_equivalent 0.11 0.11   0.13 0.05
##   above_bachelor      0.33 0.40   0.53 0.50
##   NONE                0.56 0.49   0.34 0.45
```

2.4.12 Major vs. AttendenceEvent

Based on our segmentation of each Major's donation level, we find that there is a linear trend between Attendance and Major Donation Level, namely that as the donation level increased, so did the percentage of event attendees. This is not surprising considering we saw earlier that Attendance is correlated with donation level.

```
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

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

```
## , , 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

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

```
## , , 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

3.1 Summary

The final model we estimated is:

$$logit(\hat{P}(Y \leq j)) = \hat{\beta}_{j0} + \hat{\beta}_1 Married + \hat{\beta}_2 Attended + \hat{\beta}_3 Major.Donation.Level + \hat{\beta}_4 Adv.Deg + \hat{\beta}_5 Gender + \hat{\beta}_6 FY15Giving.Grou$$

We found that past years giving is the key predictor for 2016 contributions. There are increasing Odds of donation for alumni who previously donated in high donation brackets. This trend applies to all previous Fiscal Years, but more so to the more recent donation years. However, to as lesser extent Gender, being in high donation Major category, and attendance of events increased the odds of 2016 contribution.

3.2 Modeling Process

Based on our exploratory data analysis we decided to include most of the variables provided in the dataset as explanatory variables while estimating our model. These included all years of FYGiving.Grouped (2012-2015), our categorical variable of Married, AttendenceEvent, Major.Donation.Level, Adv.Deg,Gender, and as described below we originally included Class.Year, but removed it from the model due to poor model fit.

3.3 Ordinal versus Nominal Multinomial model selection

We noted that there was a natural order to donation levels so using a nominal model will include additional (order) information that the multinominal model will overlook. Further, from a interpretability standpoint, there is no value of estimating distinct parameters between each donation levels. Lastly, when reviewing each Independent Variable, we found that their odds ratios were relatively similar across the different levels of the response variable FY16Giving.Grouped. This provided additional justification for estimating an Ordinal 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)
## 1982  1.15    1.36      1.42      1.93
## 1992  1.25    1.39      1.71      1.48
## 2002  1.22    1.62      1.17      0.94
## 2012  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
```

3.4 Model Selection

In our analysis, which we do not show for brevity, we identified that FY15Giving.Grouped had the highest in-model explanatory power of all explanatory variables by themselves (the lowest AIC and it was statistically significant). Thus, we chose the following "base", (H_0) model for hypothesis testing.

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

```
## 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
```

All the models we estimated are summarized in table below (*NB that the table displays Wald SE and p values based on the Wald Test, however our Likelihood Ratio tests come to the same conclusion as theWald Tests).

```
model1_Ha <- clm(formula = FY16Giving.Grouped ~ Class.Year + Married +
    AttendenceEvent + FY15Giving.Grouped + Major.Donation.Level + Adv.Deg +
    Gender + FY14Giving.Grouped + FY13Giving.Grouped + FY12Giving.Grouped,
    data = givings, link = "logit")
anova(model_Ho, model1_Ha)
```

```
## Likelihood ratio tests of cumulative link models:
##
##           formula:
## model_Ho  FY16Giving.Grouped ~ FY15Giving.Grouped
## model1_Ha FY16Giving.Grouped ~ Class.Year + Married + AttendenceEvent + FY15Giving.Grouped + Major.D
##           link: threshold:
```

```
## model_Ho  logit flexible
## model1_Ha logit flexible
##
##           no.par    AIC  logLik LR.stat df Pr(>Chisq)
## model_Ho       8 1705.7 -844.85
## model1_Ha     32 1562.9 -749.47  190.76 24  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
AIC(model_Ho, model1_Ha)
```

```
##           df      AIC
## model_Ho   8 1705.709
## model1_Ha 32 1562.950
```

Our null hypothesis, Ho, is that:

$\beta_{Class.Year} = \beta_{Married} = \beta_{AttendenceEvent} = \beta_{Major.Donation.Level} = \beta_{Adv.Deg} = \beta_{Gender} = \beta_{FY14Giving.Grouped} = \beta_{FY13Giving.Grouped} = \beta_{FY12Giving.Grouped} = 0$

Our alternative hypothesis is that:

$\beta_{Class.Year} \neq \beta_{Married} \neq \beta_{AttendenceEvent} \neq \beta_{Major.Donation.Level} \neq \beta_{Adv.Deg} \neq \beta_{Gender} \neq \beta_{FY14Giving.Grouped} \neq \beta_{FY13Giving.Grouped} \neq \beta_{FY12Giving.Grouped} \neq 0$

By comparing the FY15Giving.Grouped-only model (model1.Ho) to the model with all years of FYGiving.Grouped (2012-2015), our categorical variable of Married, AttendenceEvent, Major.Donation.Level, Adv.Deg,Gender, and Class.Year (model1.Ha) using the Likelihood Ratio test, we found that adding the additional variables besides FY15Giving.Grouped contributed to the model - ie. there is strong support to reject the null hypothesis.

Because there is good evidence that our addition of 9 explanatory variables (model.Ha) to our base model (model.Ho) made a singificant contribution to the model, and because the AIC decreased when we added the explanatory variables as suggested by our EDA (model.Ho = 1705.7, model.Ha = 1560.5), we proceeded with evaluating model fit.

We are aware of the strong correlation between donation levels from 2012 to 2016. However, we've decided to keep them as well for two reasons: 1. They are contributing to the model prediction significnatly(shown by the ANOVA test). 2 The predicted coefficient looks consistent with our observation in EDA. 3 Although a lot of category variables are not signficant in Wald test because of inflated standard error, it doesn't necessary negatively effect the final prediction.

We found that the Class.Year is neither significant in the model nor reports a correct coefficient. On the top of that, in our EDA it is clear that earlier (specially for 1972) graduates is more likely to donate high while the coefficient is showing the opposite, which is very misleading for our model interpretation. The reason for the problem is that Class.Year highly correlated with all the previous donation variables, which increases the standard error (leading to unsatisfactory significance test and incorrect coefficient) and doesn't provide additional information (the model with or without this variable is not significantly different from each other). **Considering the poor fit of model.Ha, we removed Class.Year and re-evaluated the model we estimated with the remaining explanatory variables.**

Our null hypothesis now was that:

$\beta_{Married} = \beta_{AttendenceEvent} = \beta_{Major.Donation.Level} = \beta_{Adv.Deg} = \beta_{Gender} = \beta_{FY14Giving.Grouped} = \beta_{FY13Giving.Grouped} = \beta_{FY12Giving.Grouped} = 0$

```
model2_Ha <- 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(model2_Ha)
```

```
## 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_Ho, model2_Ha)
```

```
## Likelihood ratio tests of cumulative link models:
##
##           formula:
## model_Ho  FY16Giving.Grouped ~ FY15Giving.Grouped
## model2_Ha FY16Giving.Grouped ~ Married + AttendenceEvent + FY15Giving.Grouped + Major.Donation.Level
##           link: threshold:
## model_Ho  logit flexible
## model2_Ha logit flexible
##
##           no.par    AIC  logLik LR.stat df Pr(>Chisq)
```

```
## model_Ho       8 1705.7 -844.85
## model2_Ha     28 1560.7 -752.37  184.98 20  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
AIC(model_Ho, model2_Ha)
```

```
##           df      AIC
## model_Ho   8 1705.709
## model2_Ha 28 1560.734
```

```
Anova(model2_Ha)
```

```
## 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
```

From the likelihood ratio test, we find that there is strong evidence that association exists for all variables except for AttendenceEvent and Major.Donor.Level. The estimated model then becomes:

$$logit(\hat{P}(Y \leq j)) = \hat{\beta}_{j0} + \hat{\beta}_1 Married + \hat{\beta}_2 Attended + \hat{\beta}_3 Major.Donation.Level + \hat{\beta}_4 Adv.Deg + \hat{\beta}_5 Gender + \hat{\beta}_6 FY15Giving.Grou$$

where $\hat{beta}_{10} = 2.41, \hat{beta}_{20} = 4.05, \hat{beta}_{30} = 6.49, \hat{beta}_{40} = 7.87$ We test for independence among the variables using the Anova function, with:

$$H_0 : \beta_2 = \beta_3 = ...\beta_p = 0$$

$$H_a : any\ \beta_i \neq 0$$

Finally, we want to investigate the interaction between Gender and Major.Donation.Level that we have identified in section 2.5.2

```
model3_Ha <- clm(formula = FY16Giving.Grouped ~ Married + AttendenceEvent +
    FY15Giving.Grouped + Major.Donation.Level + Adv.Deg + Gender + FY14Giving.Grouped +
    FY13Giving.Grouped + FY12Giving.Grouped + Gender:Major.Donation.Level,
    data = givings, link = "logit")
anova(model_Ho, model2_Ha, model3_Ha)
```

```
## Likelihood ratio tests of cumulative link models:
##
##           formula:
## model_Ho  FY16Giving.Grouped ~ FY15Giving.Grouped
## model2_Ha FY16Giving.Grouped ~ Married + AttendenceEvent + FY15Giving.Grouped + Major.Donation.Level
## model3_Ha FY16Giving.Grouped ~ Married + AttendenceEvent + FY15Giving.Grouped + Major.Donation.Level
##           link: threshold:
```

```
## model_Ho  logit flexible
## model2_Ha logit flexible
## model3_Ha logit flexible
##
##           no.par    AIC  logLik  LR.stat df Pr(>Chisq)
## model_Ho       8 1705.7 -844.85
## model2_Ha     28 1560.7 -752.37 184.9751 20     <2e-16 ***
## model3_Ha     31 1562.7 -750.34   4.0534  3     0.2558
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
AIC(model_Ho, model2_Ha, model3_Ha)
```

```
##           df      AIC
## model_Ho   8 1705.709
## model2_Ha 28 1560.734
## model3_Ha 31 1562.680
```

We note that the interaction term increases AIC and is not supported by the LR independence test. We choose model2_Ha as our final model

```
stargazer(model_Ho, model1_Ha, model2_Ha, model3_Ha, star.cutoffs = c(0.05,
    0.01, 0.001))
```

% Table created by stargazer v.5.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Mon, Oct 23, 2017 - 02:21:10

3.5 Odds Ratios and their Confidence Intervals

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

```
conf.beta <- confint(object = model2_Ha, level = 0.95)
conf.beta <- as.data.frame.matrix(conf.beta)
conf.beta[, 3] <- model2_Ha$coefficients[5:28]
colnames(conf.beta)[3] = "estimated"
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
```

Table 1:

	Dependent variable:			
	FY16Giving.Grouped			
	(1)	(2)	(3)	(4)
Class.Year1982		0.289		
		(0.305)		
Class.Year1992		0.189		
		(0.304)		
Class.Year2002		0.094		
		(0.299)		
Class.Year2012		0.565		
		(0.303)		
Married		0.073	−0.018	−0.023
		(0.168)	(0.162)	(0.163)
AttendenceEventAttended		0.248	0.256	0.259
		(0.172)	(0.166)	(0.166)
FY15Giving.Grouped[1-100)	2.027***	1.224***	1.212***	1.234***
	(0.172)	(0.209)	(0.209)	(0.210)
FY15Giving.Grouped[100-250)	3.641***	2.224***	2.167***	2.157***
	(0.222)	(0.269)	(0.267)	(0.268)
FY15Giving.Grouped[250-500)	5.588***	3.257***	3.165***	3.112***
	(0.375)	(0.520)	(0.518)	(0.522)
FY15Giving.Grouped[500-200000)	7.685***	4.960***	4.799***	4.785***
	(0.439)	(0.610)	(0.599)	(0.597)
Major.Donation.LevelLow		−0.108	−0.069	0.295
		(0.424)	(0.425)	(0.609)
Major.Donation.LevelMedium		0.031	0.081	0.654
		(0.367)	(0.366)	(0.550)
Major.Donation.LevelHigh		0.354	0.328	0.547
		(0.439)	(0.430)	(0.677)
Adv.Degabove_bachelor		−0.153	−0.230	−0.240
		(0.234)	(0.229)	(0.229)
Adv.DegNONE		−0.550*	−0.626*	−0.625*
		(0.250)	(0.247)	(0.248)
GenderMale		0.093	0.110	1.075
		(0.154)	(0.152)	(0.725)
FY14Giving.Grouped[1-100)		0.560*	0.595**	0.586**
		(0.223)	(0.222)	(0.222)
FY14Giving.Grouped[100-250)		0.760*	0.839**	0.862**
		(0.325)	(0.320)	(0.320)

```
## 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
```

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 .71 to 1.35 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 38 to 407 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 1.25 times as large when the alumni had above a bachelor degree versus having a bachelor degree. This range is reduced to .33 to .87 times as large when the alumni had no additional degree compared to having a bachelor degree.
- The odds of donating above a certain bracket are between .83 to 1.50 times as large when the alumni was a male compared to female.
- The odds of donating above a certain bracket are between .93 and 1.79 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 .61 to 3.28 times as large when the alumni majored in what we categorized as a 'High Donation' major compared the no donation major category.

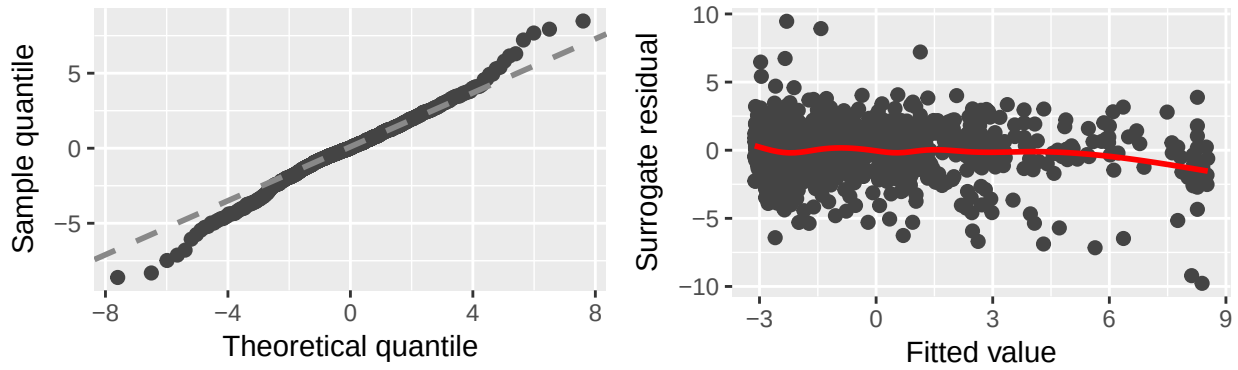
We notice that many of the variables include 1 within the 95% confidence interval. This suggests that there is not sufficient evidence to indicate that these variables increase the odds of donating. However, through our EDA and logical assessment, many of these variables can be key contributers and also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

3.6 Model Evaluation

We are going to use a special library to get the residuals (since clm() is not a glm model).

```
p1 <- autoplot(model2_Ha)
p2 <- autoplot(model2_Ha, what = "fitted")
grid.arrange(p1, p2, ncol = 2)
```

```
## `geom_smooth()` using method = 'gam'
```



We note that the residuals deviate from the Normal at the extremities but remain Normal at the center. While the variance is not fully constant, the shape and spline of the residuals vs. fitted values appear acceptable for our model.

3.7 Model Fit

We correctly predict 731 out of 1000 samples, so our model is 73% accurate. We are more accurate (90% !) for predicting alumni who won't donate and we are also reasonably accurate at predicting who will donate $500 or more. However, our accuracy for the mid-range of donations is not great.

```
# explanatory variables
var.list <- c("Class.Year", "Married", "AttendenceEvent", "FY15Giving.Grouped",
    "Major.Donation.Level", "Adv.Deg", "Gender", "FY14Giving.Grouped",
    "FY13Giving.Grouped", "FY12Giving.Grouped")
pred.list <- c("[0-1)", "[1-100)", "[100-250)", "[250-500)", "[500-200000)")
newData <- unique(givings[, var.list])
predicts <- cbind(newData, predict(model1_Ha, newdata = newData, type = "class"))
predicts <- merge(predicts, givings, by = var.list)
(results.t <- xtabs(~FY16Giving.Grouped, data = predicts[(predicts$fit ==
    predicts$FY16Giving.Grouped), ]))
```

```
## FY16Giving.Grouped
##        [0-1)      [1-100)     [100-250)     [250-500) [500-200000)
##          530           57           93             3           45
```

```
round(results.t/xtabs(~FY16Giving.Grouped, data = givings), 2)
```

```
## FY16Giving.Grouped
##        [0-1)      [1-100)     [100-250)     [250-500) [500-200000)
##         0.90         0.33         0.65          0.08         0.76
```

4 Final Remarks

Through our analysis we've found multiple variables in the data that correlate with, and help predict, donation levels from alumni. Although our model fits the data well, we've identified a few areas that should be noted:

1. Our model leverages previous years' donations as a primary predictor of the level of donation for the current year. For the majority of alumni, this works well as we have data from previous years. For new alumni, however, this variable will be missing for a certain number of years. We have found, however, that the further back the donation data goes, the less predictive power it provides, thus this is well mitigated. It would still be beneficial to ensure we have as much historical data on alumni donation as possible, so as to make full use of our model.

2. There are certain subcategories of variables (e.g. Widowed in the Marital Status variable, certain majors in the Major variable, etc.) that have very little data. For our analysis, we've elected to group these variables (as in the Major.Grouped specification), which works well, but it is likely that having additional observations in these subcategories would help fit a better model. We understand this is not a controllable input, but we are highlighting it here for awareness.

3. Comparing the predictions from our model with the actual donation levels in FY16, we find that our predictions are far more accurate for the lowest and highest tiers of donations. This is likely because there is greater variability in the intermediate level, whereas the relationships are much stronger as you look at the extremes. In our scenario, this actually works well because a large percentage of the total donations are weighted toward the higher donation tiers, which mean that accurately predicting alumni

in these tiers are more important than other tiers. Additionally, accurately predicting those who will not donate is also helpful so the administration can allocate their outreach accordingly.

4. We approached this analysis with an in-sample EDA and model assessment as we wanted to leverage as much of the data in our analysis as possible, knowing that additional data from next year would be relatively easy to attain and test on. Given a larger sample, we would be more comfortable splitting the data in to a training and testing sample, especially given the small subcategories from the variables mentioned in the previous point.

As with most analyses, more data can lead to better models, but we've analyzed the existing data set and developed a model which we think fits the data and will predict the level of donations well. We recommend the administration leverage this model to predict FY17 Donation levels and check it against the actuals to confirm the model's accuracy, particularly on the highest level donations.

We do, however, note that fundraising professionals typically make use of datasets that include some measure of wealth. Rather than use proxies for wealth (like Next Degree and high donor Major category for example), we recommend that the Univeristy obtain actual income level information for alumni. If this is not possible, there are other ways fundraisers commonly use to estimate wealth that include estimating potential donors' home values using donor's home addresses - people who own $2 million dollars or more in real estate are 17 times more likely donate to a nonprofit than the average person (www.bidpal.com/identifying-major-donors-top-strategies-tools). By identifying wealthy alumni, the University could invite them to "VIP University Events" and cultivate relationships with these donors that tend to result in larger donations.