Statistical Consulting Project

Quinn Sweetnam

March 27, 2018

Abstract

Though Involutional Ptosis treatment is well-documented and thorough, the causes of this eye condition are still obscure. Existing research links the presence of Ptosis to certain medical conditions and procedures but in the Winter of 2018 doctors from the Stein Eye Institute, in association with Ronald Reagan Medical Center in Los Angeles, asked Senior Statisitcs students at UCLA to analyze data to better their understanding of what causes Ptosis. This study aims to further analyze the relationship between Ptosis and a number of risk factors through statistical analysis. The risk factors under study are Hypertension, Ischemic Heart Disease, Heart Failure, Peripheral Vascular Disease, Chronic Kidney Disease, Hyperthyroidism, Type I Diabetes, Type II Diabetes, Hyperlipidemia, Obesity, (history of) Alcohol Abuse, (history of) Tobacco use, and Peripheral Neuropathy. The study also examines Ocular surgery as a risk factor for developing ptosis. Charged with measuring the contribution of certain risk factors to developing Ptosis, the study was limited to running logistic models due to the ease of interpretability and the primary focus was the main effect of a given risk factor so variable transformations were limited. To conduct the analysis, five logistic regression models were ran on selected subgroups: patients with at least one risk factor (RF) listed, and patients who underwent ocular surgery. Regression models that included patients with ocular surgery added one more variable (ocular surgery) to the model alongside the other RF variables. The model with the greatest predictive power included patients who had ocular surgery and at least one RF; this logistic regression model includes second-order interaction effects between RF. The AUC for this model is 68.06% and identified Eye Surgery, Hyperthyroid Disease, Type II Diabetes, Kidney Disease, Alcohol Abuse, and Hypetension as statistically significant predictors of Ptosis at a 5% significance level.

Explanation of the Data

The data for the analysis was gathered through a case control study and was provided by the Stein Institute. The control group consists of 13,128 patients without ptosis, and the cohort consists of 8,297 patients with ptosis. The original experimental design had two different control groups selected, which were age and gender matched in a 4 to 1 ratio; patients under 18 were excluded from the study. All of the patient records came from major medical centers located in California and patients could only be identified via randomly generated study ID's as to protect patient identity and data.

The variables of this study are all binary and are as follows:

Variables (called risk factors)

- Ptosis (response)
- Eye Surgery
- Hypertension
- Ischemic Heart Disease
- Heart Failure
- Perivascular Disease
- Kidney Disease
- Hyperthyroid
- Hypothyroid

- Type I Diabetes
- Type II Diabetes
- Hyperlipidemia
- Obesity
- Alcohol Abuse
- Tobacco Use
- Peripheral Neuropathy

Study Goal

This study was tasked with finding and measuring the affect of the different risk factors on causing Ptosis in the general population. It was the above risk factors that they were interested in and so this study does not look to examine or incorporate alternate diseases, demographic, or environmental data. The doctors were interested in drawing inferences from models and explroatory analysis, not finding the most accurate or powerful predictive model. According to the doctors at the Stein Institute, a previous third party had analyzed the data and their conclusions had been confusing and contrary to established medical knowledge so easy interpretation and attention to detail were key.

Study Assumptions

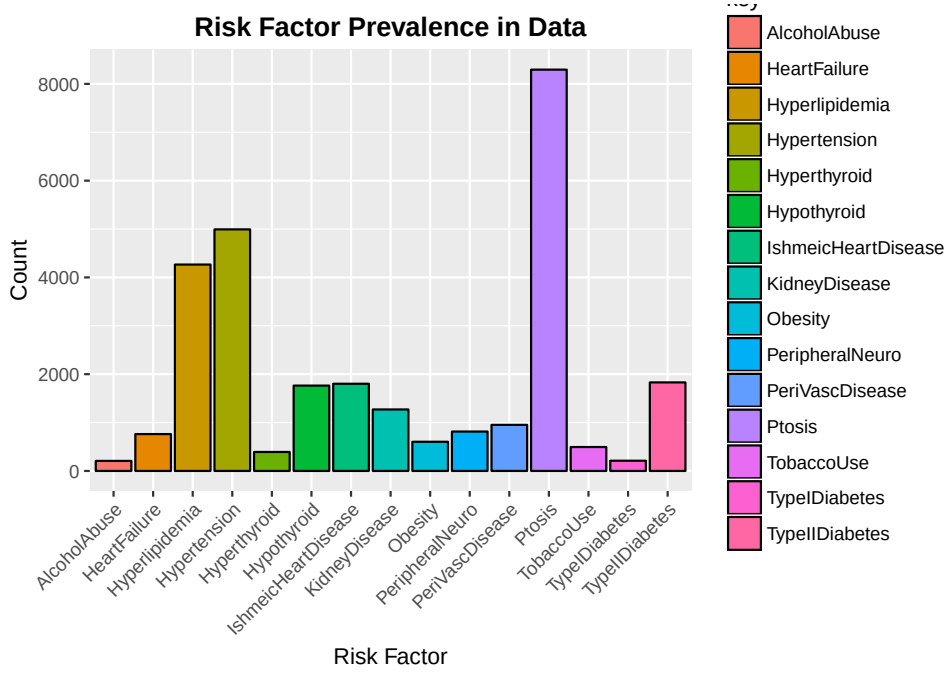
1. High levels of dependence are often found in medical data and most of the variables were found to be dependent on another using Chi-Square Tests of independence. This dependence is ignored during logistic regression.
2. The data was collected in a statistically appropriate manner, according to the doctors interested in the analysis.
3. Ocular surgery was conducted on the same eye that Ptosis was present, *despite having no way to ensure this from the data.* This concern was repeatedly raised and reported to the doctors interested in the study but they instructed the study to proceed anyway and to make this assumption. It remained a concern and should be duely noted before preceeding with understanding the results.

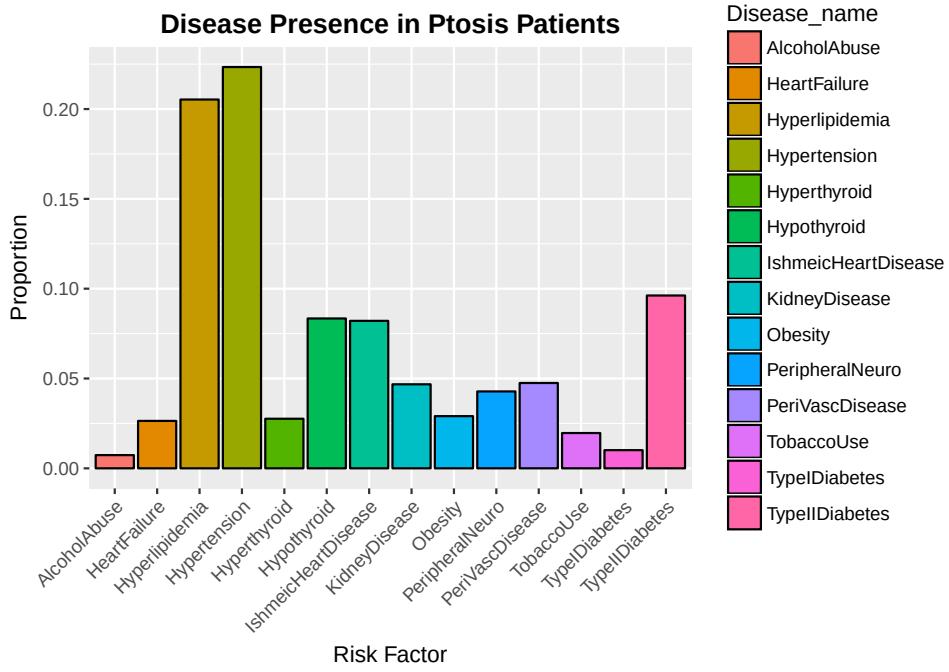
Loading in Data

```
## [1] 13130    15
```

```
## [1] 8295   15
```

Data Exploration





We can see that Ptosis is the most prevelant condition in the data, followed by Hyperlipidemia, Hypertension, Hypothyroid, Ischemic Heart Disease and Type II Diabetes. These same risk factors are frequent in just Ptosis patients with some 20% of Ptosis cases having Hypertension or Hyperlipidemia. This is the first clue that these risk factors may be associated with Ptosis. We can also notice low levels of Tobacco Use and Obesity in the data overall and these are not actually reflective of national obesity and tobacco use rates in the U.S. but of rates in California.

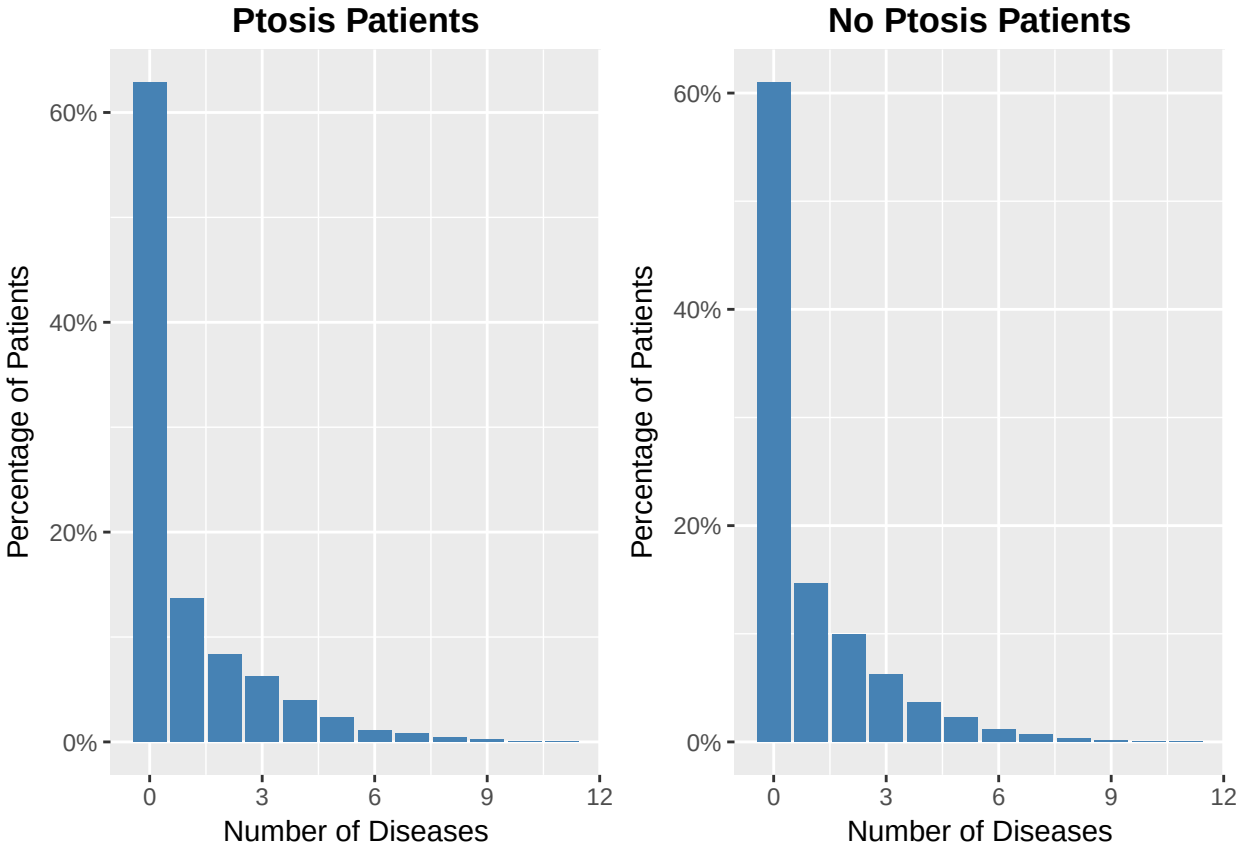
The initial stages of data exploration revealed an important aspect about the structure of the data in the study.

```r
suppressMessages(library(scales))
# How many patients have no instance of a risk factor or Ptosis?
data <- data %>%
  mutate(sum = rowSums(.))

sum(data$sum == 0)/dim(data)[1] # proportion of perfectly "healthy" patients
```

```
## [1] 0.3736756
```

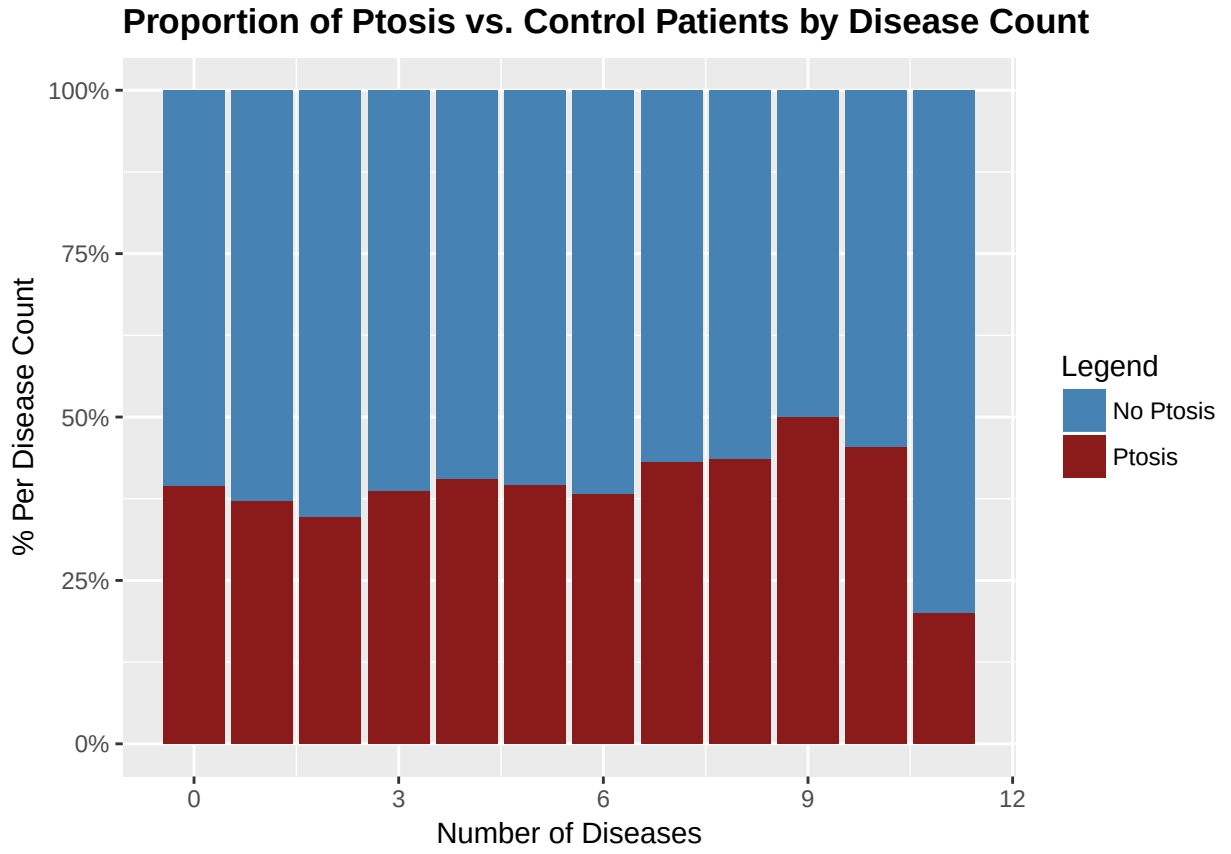
Some 38% of all patients from the data are completely free of any of the risk factors in the study and have no instance of disease. These individuals can be considered perfectly "healthy" by the standards of this study, and make up a large section of the overall sample. This posed the question, just how many patients in the study have none of the risk factors of interest?



Looking at these graphs it becoes immediately apparent that the suspicion that the majority of the data is actually devoid of information is true. You can see that just over 60% of all patients with or without Ptosis have none of the risk factors and this could seriously distort analysis. It is true that in that the lack of information in itself is informative but we did not expect that such a substantial number of the patients were so "healthy." As I will show later, incorporating these "Null" patients in any models will actually cause the model to associate a given risk factor with lower probability of Ptosis.

How does overall health of a patient contribute to Ptosis?

While considering factors that could lead to Ptosis, we thought that one potential contributor of Ptosis could be the overall health of a given patient. While we did not have a measure of "overall" health, we associated the more risk factors a given patient head with decreasing overall health. Thus, Disease Total acted as a pseudo score for patient health.



However, you can see that increasing numbers of disease in a patient do not dramatically increase the proportion of Ptosis cases. There seems to be a slight increase in the proportion of Ptosis cases when patients have 7-9 diseases but overal patient health appears inconsistent with Ptosis.

What do to with the empty cases?

The discovery that a considerable amount of our data did not contain information on the risk factors of the study presented a dilemna. On the one hand, it comfirmed that other factors not in the study likely contribute to Ptosis as 5,214 of the 8,295 Ptosis Patients in the study had none of the risk factors of interest. However, subsequent analysis would likely be distorted by such a large presence of "empty" observations so after consulting with the doctors promoting the study, we made the decision to remove these empty cases and proceed with analysis. This takes the original number of individuals of the study from 21,425 down to 8,205, which is a significant drop in number. However, I believe that by focusing on individuals with at least one risk factor the study will better be alined with the original objective of finding how these particular factors contribute to Ptosis

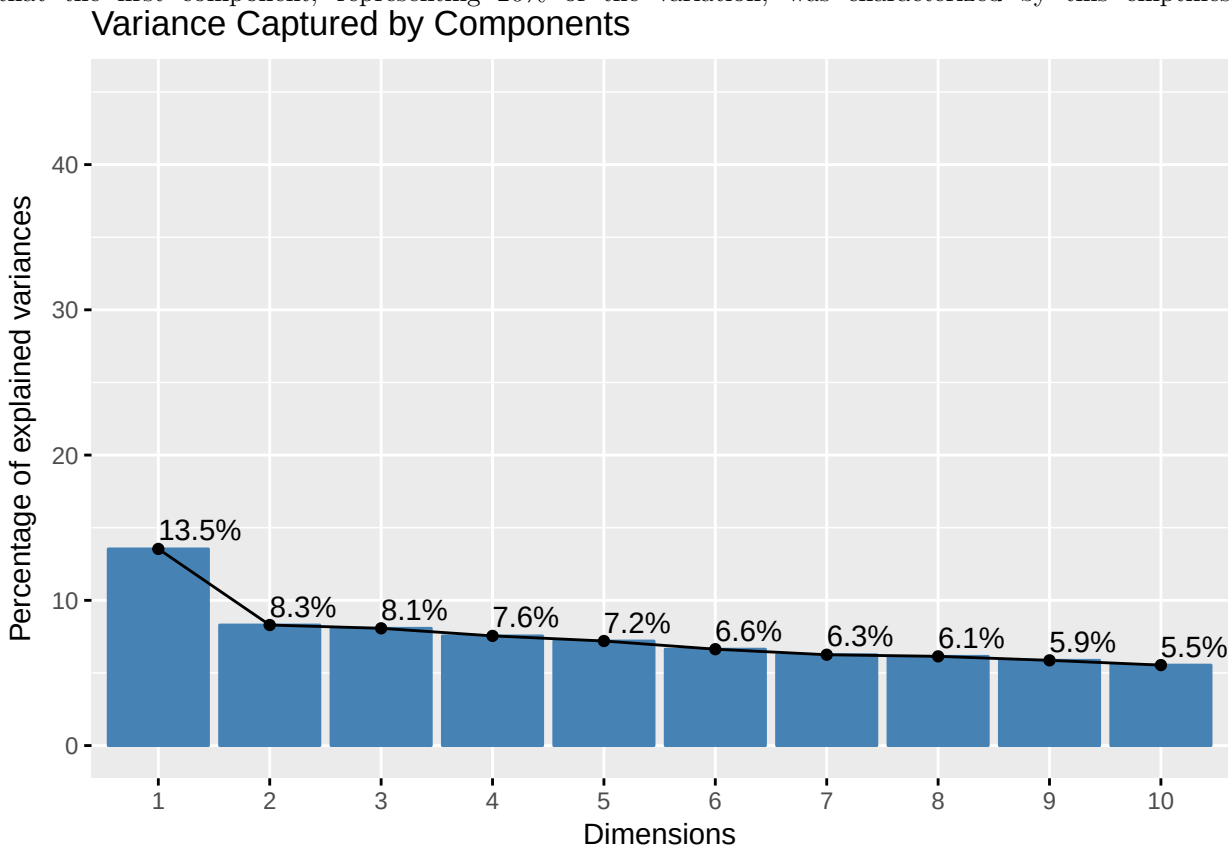
```
## [1] 8205   16
```

Looking for latent patterns amongst the risk factors

To look for underlying groupings within the risk factors and the response, I used a technique called Multiple Correspondence Analysis (MCA). The goal of MCA is to find the "hidden structure" of the categorical variables and we use it to come to a general understanding of how the variables/response are related. I choose to use the Indicator Matrix, which is just a representation of the data as it is, with rows being observations and columns being our risk factors. Associations between the factors are uncovered by calculating the Chi-Square distance between rows and columns, and these distances are maximized as to find the largest sections of

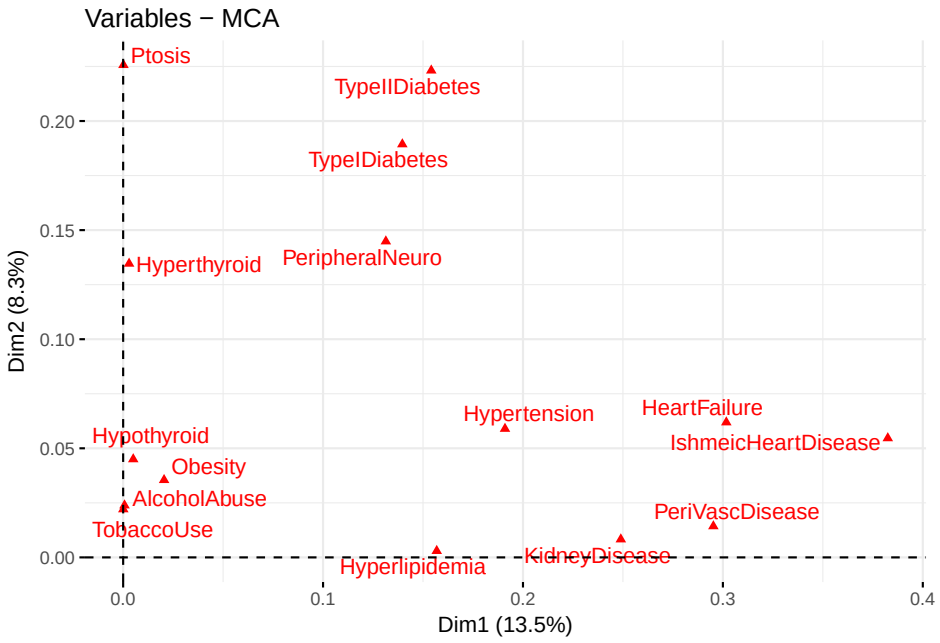
variance within the data. MCA is useful in that it can capture both linear and non-linear patterns, and will help us reduce the dimension of the data set from 15 down to 3-4 (hopefully).

I used the FactoMineR statistical package for this analysis, and uncharacteristccally included the response (Ptosis) in the analysis to see if it could be grouped with any of the risk factors of interest. The analysis was run on the data with the null observations removed because including them meant that the first component, representing 20% of the variation, was characterized by this emptiness.



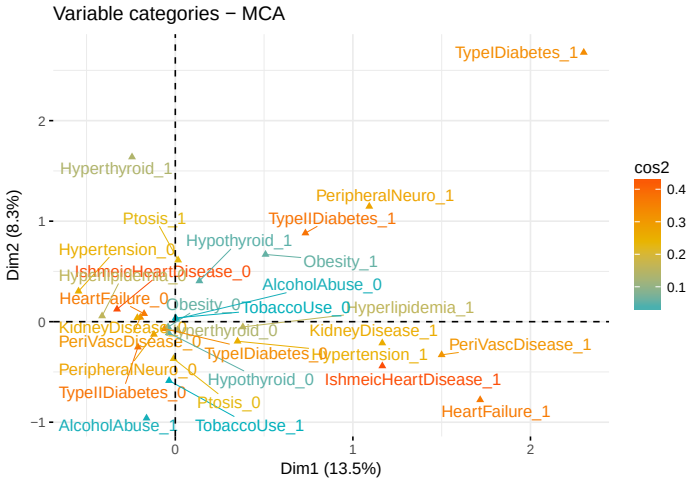

```
##        eigenvalue variance.percent cumulative.variance.percent
## Dim.1 0.13538798        13.538798                    13.53880
## Dim.2 0.08300180         8.300180                    21.83898
## Dim.3 0.08072765         8.072765                    29.91174
## Dim.4 0.07551849         7.551849                    37.46359
## Dim.5 0.07199354         7.199354                    44.66295
## Dim.6 0.06636581         6.636581                    51.29953
```

Initial results of the MCA highlight that the percentage of variances captured by the components does not meet our initial goal of capturing significant variation within the first 3 components. In fact, it takes the first 6 components to capture 50% of the variation in the data revealing the complexity of the underlying relationships between the predictors.



The above plot shows the how correlated each variable is with the first two components, who collectively captured 21.8% of the data. We can begin to see some of the groupings and relations bteween the variables. Diabetes Type I & II are clustered with Peripheral Neuropathy and Heart Failure, Ishemic Heart Disease and Perivascular Disease are grouped. Ptosis is not correlated with the first dimension at call, but is the most correlated with dimension 2, and thus with Diabetes and Peripheral Neuropathy.

Now I will look at the squared cosine of the risk factors, which will indicate how well each variable and level is captured by the first two components. Blue indicates low capture, yellow mid capture, and red high capture.



Both levels of Isemeic Heart Disease are well captured by the first two components, and we can see that Ptosis 1 (patient having Ptosis) is also mildy captured. Thus, this is further confirmation that Ishemic Heart disease, diabetes and heart failure have underlying relationships with Ptosis and each other - at least in the first two components. Another clue for that these risk factors are related to Ptosis.

Logistic Regression to predict Ptosis

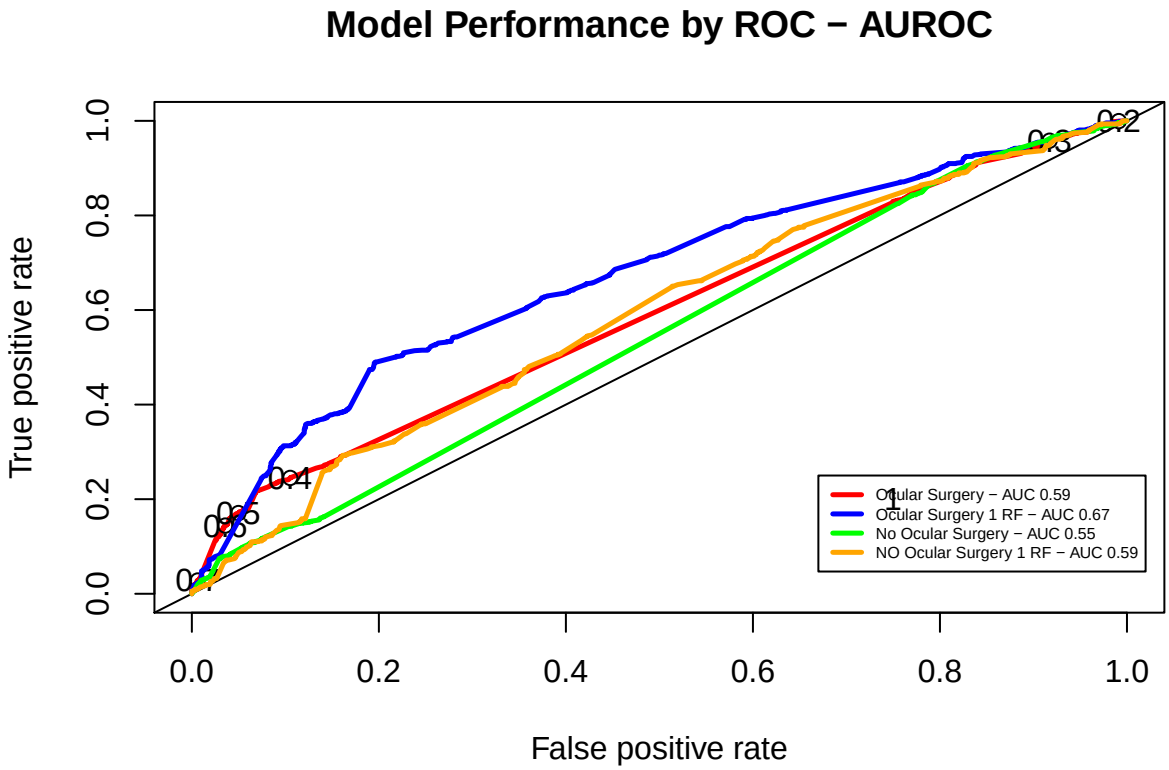
This study choose to model Ptosis using the logistic regression algorithm because of the ease of interpretation and rather short training time. To ensure the stability of the coefficients produced by the model and the

accuracy, the model was trained on a random split of 75% training and 25% testing, and during the training phase 10-Fold cross vaildation was used to further stability. Cross-Validation is the process of dividing the data into separate batches and training the data on 9 of the 10 batches, and validating it on the 10th "hold-out" batch. The batches are rotated through so each one is treated as the hold-out set, and the misclassification rate of the models is averaged.

Including Surgery

During this stage of the analysis, I include the ocular surgery data as an additional risk factor for the model to consider. Ocular surgery is known in medical literature to cause Ptosis and was not included in exploratory analysis because the goal was to understand the relationships of diseases with each other and not a medical procedure. Ocular surgery is binary encoded. Models were trained on 4 different subgroups of the data for comparison; the full original data set, the full data set with surgery included, the reduced data set of patients with at least one risk factor, the reduced data set of at least one risk factor and ocular surgery.

```
## Warning in predict.lm(object, newdata, se.fit, scale = 1, type =
## ifelse(type == : prediction from a rank-deficient fit may be misleading
```



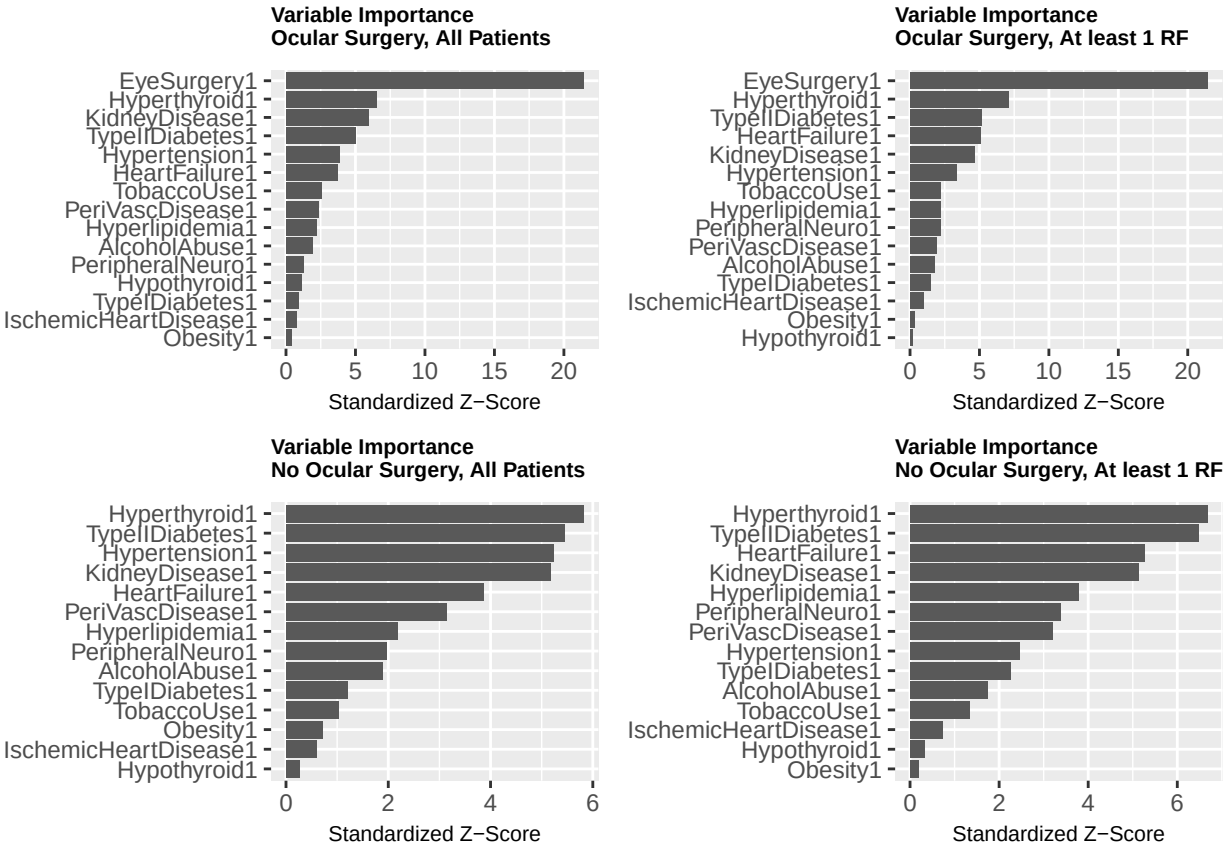
After training the models on the various subgroups of the data, we can see that the most effective model was trained on at least one risk factor, ocular surgery included group. ROC curve analysis shows us that this model had the best sensitivity and specificity values. I also analyzed the two level interactions in a spearate moel and included but we can see that we only get a marginal improvement in AUC score, and after confering with the doctor team we determined that the primary interest was the model main effects as it can be heard to interpret what the odds-ratios of interaction effects truly mean and the doctors emphasis on main effect of risk factos. For simplicity sake, I only include the confusion matrix analysis of the best model below. Considering our focus of finding strong associations between risk factors and the disease we mainly focused on maximizing sensitivity and specificity.

```
## Confusion Matrix and Statistics
##
##
```

```
##        0    1
##   0 1194  581
##   1  162  341
##
##                Accuracy : 0.6738
##                  95% CI : (0.6542, 0.6931)
##     No Information Rate : 0.5953
##     P-Value [Acc > NIR] : 0.000000000000006004
##
##                   Kappa : 0.27
##  Mcnemar's Test P-Value : < 0.00000000000000022
##
##             Sensitivity : 0.8805
##             Specificity : 0.3698
##          Pos Pred Value : 0.6727
##          Neg Pred Value : 0.6779
##              Prevalence : 0.5953
##          Detection Rate : 0.5241
##    Detection Prevalence : 0.7792
##       Balanced Accuracy : 0.6252
##
##        'Positive' Class : 0
##
```

A couple of important metrics jump out in the confusion matrix that indicate that the logistic regression model run on the at least one Risk Factor, surgery included subgroup would be the best for the study. This model had the best trade-off between sensitivity and specificity, and by the far highest sensitivity score. The model choose a "0" score for Ptosis (no Ptosis) as the positive case, so the output above is actually needs to switch sensitivity and specificity. Thus the true sensitivity is 37%, which is the rate we correctly predict true cases of ptosis and while is bad by many standards, this was the best predictive score of ptosis we were able to produce. One drawback of this model is that it tends to over-predict cases of No Ptosis, which we theorize come from non-linear relations between the disease. Other algorithms such as K-Nearest-Neighbors, Discriminant Analysis, Random Forest and Support Vector Machines could more easily capture these non-linear patterns, but the ease of interpretation and scope of our study led us to stay with Logistic Regression.

Variable Importance



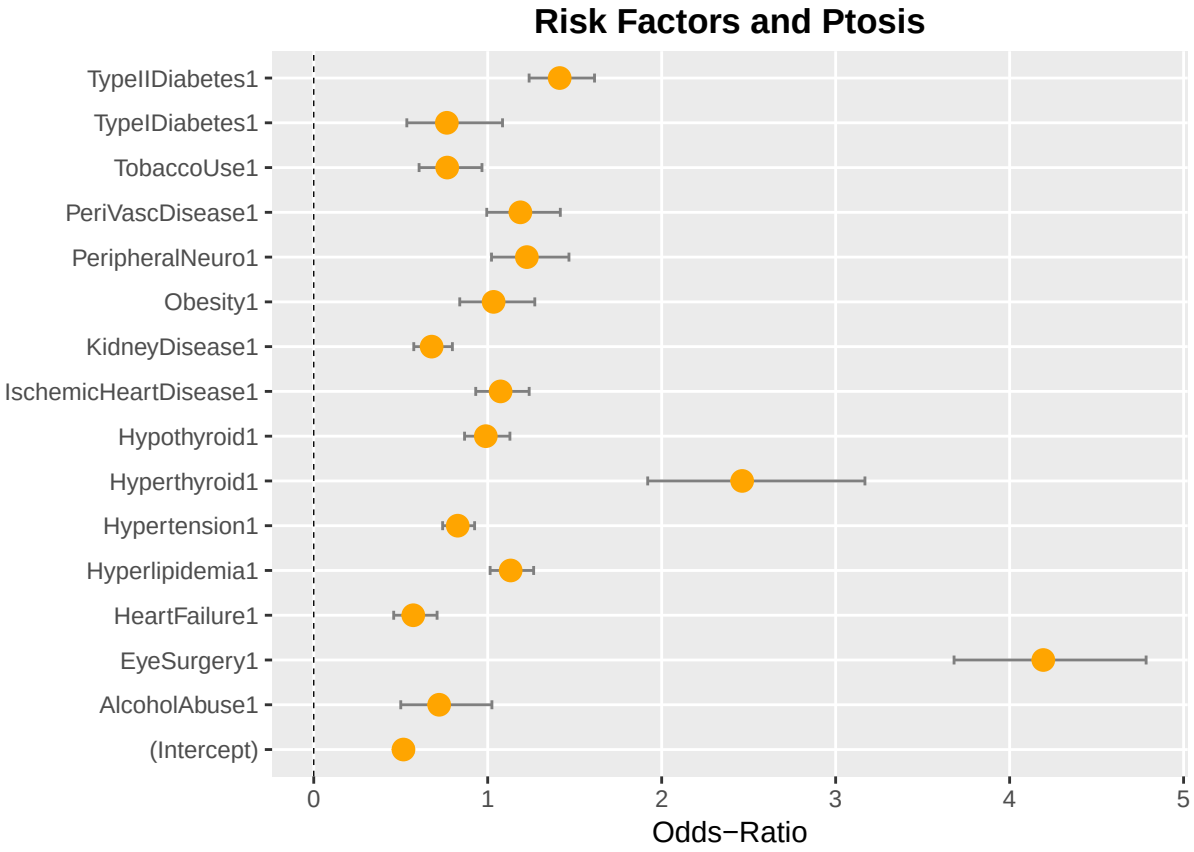
One benefit of running the models on multiple subgroups is that we get consistent results for which risk factors are important. If Ocular Surgery is present it far out-ranks other risk factors in importance, but even after controling for surgery the next most important risk factors are some combination of Hyperthyroidism, Type II Diabetes, Kidney Disease, and Heart Failure. This asserts the relative significance of these risk factors as they continually show up in various subgroups and consistently contribute the most to predicting Ptosis. Furthermore, Hyprerthyroidism is ranked as most important in all the models (surgery controled). Thus, these becamse our *primary risk factors* and visually you can see that they make up the first one or two "levels" in the bar charts, followed by an importance drop off and our *secondary risk factors.* These are Peripheral Nueropathy, Hyperlipidemia, Hypertension and Perivascular Disease.

Hypertension, an interaction Story

While considering interaction effects was ultimatey ruled out because measuring the meaning of the effect of an interaction can be diffult, it is interesting to not that 7 out of the 20 significant interactions had hypertension as 1 pair. This gives some context behind the prevelance of hypertension in the patient population seen before and how common it is in America as a whole (1 out of 3 adults have high blood pressure). The disease has been linked to a number of other health complications such as heart disease, and now quite possibly Ptosis.

Interpretation of Results

```
## Waiting for profiling to be done...
```



Most of the risk factors Odds Ratio (OR) hover around 1 suggesting that they have very little real effect on causing Ptosis. However, having eye surgery makes you 4.2 times more likely to have Ptosis and Hyperthyroidism makes you 2.5 times more likely to have Ptosis. Heart Failure, Kidney Disease and Hypertension are the only signficicant factors that decrease your chance of having Ptosis - 43%, 33% and 18% less likely.

Conclusion

After exploring the issue of sparsity in the data and running logistic regression models on variou subgroups, we determine that the most significant risk factors for the onset of Ptosis are Eye Surgery and Hyperthyroidism, followed by Type II Diabetes. These diseases consistently emerged as important across subgroups and have the strongest and most significant coefficients in the models. Eye Surgery's link to Ptosis is more clear as it is likely a result of damage done to eyelid muscles during a surgery but the statistical team was unsure of the medical link between Hyperthyroidism and Type II Diabetes. It was referred to medical professionals for deliberation and discussion.

Heart Failure, Kidney Disease, and Hypertension all decreased the porbability of Ptosis which is interesting because the diseases themselves are linked. Furthermore, Hypertension was the most significant interaction effect and these interaction effects pointed toward increasing the risk of Ptosis. While hard to isolate individual meaning, it seems to indicate taht Hypertension on its own is a detractor, but once combined with further medical ailments the combination helps lead to Ptosis.

We believe that analytical performance could be improved if demographic and environmental data as well as more diseases are included in analysis. It is our recommendation that these features be gathered in the next round of research and analyzed for importance.

Acknowledgments

This document was comprised of code and analysis done by myself. The original assignment was with a team of two doctors, Dr. Daniel Rootman and Dr. Ben Campbell, and 5 Senior Statistics students, Donjo Lau, Danny Stapleton, Sophie Ringle, Ignat Kulinka, Aida Ylanan and myself. All of these individuals deserves credit but the above is my personal take on the project after our team finished creating a Powerpoint presentation and reflects my work.