**Assumption Checking LDA vs. QDA - R Tutorial (Pima Indians Data Set)**

In this blog post, we will be discussing how to check the assumptions behind linear and quadratic discriminant analysis for the [Pima Indians data](https://www.kaggle.com/uciml/pima-indians-diabetes-database). We also built a [Shiny app](http://thatdatatho.com/2018/04/09/86/) for this purpose. The code is available [here](http://thatdatatho.com/2018/04/16/r-shiny-code-for-pima-indian-diabetes-data-set/).

What we will be covering:

* Data checking and data cleaning
* Checking assumption of equal variance-covariance matrices
* Checking normality assumption

In the [next blog post](http://thatdatatho.com/2018/04/02/classification-quadratic-vs-linear-discriminant-analysis-pima-indians-data-set/), we will be implementing the linear discriminant algorithms.

Let's get started!

First, we have to load some libraries and then read in the data.

library(biotools)

library(heplots)

library(ggplot2)

library(car)

library(dplyr)

library(gridExtra)

diabetes <- read.csv("diabetes.csv")

head(diabetes, 3)

## Pregnancies Glucose BloodPressure SkinThickness Insulin BMI

## 1 6 148 72 35 0 33.6

## 2 1 85 66 29 0 26.6

## 3 8 183 64 0 0 23.3

## DiabetesPedigreeFunction Age Outcome

## 1 0.627 50 1

## 2 0.351 31 0

## 3 0.672 32 1

dim(diabetes)

## [1] 768 9

diabetes <- rename(diabetes, Pedigree = DiabetesPedigreeFunction) # from dplyr

When examining the data frame, we notice that there are some people with an insulin level of zero. Insulin levels might be very important in determining if someone has diabetes or not and an insulin level of 0 is not possible. Therefore, we are deciding to throw away all observations that have an insulin level of 0.

ins.zero <- subset(diabetes, Insulin == 0)

nrow(ins.zero)

## [1] 374

nrow(diabetes)

## [1] 768

Given that we only have 768 observation, it hurts very much that we have to throw away almost half of our observations. One could do the analysis without insulin or impute the missing insulin level based on the fact that someone has diabetes or not. In addition to the zero values in the insulin column, there are also zero values in columns two to six. We are going to remove these values as well.

diabetes[, 2:6][diabetes[, 2:6] == 0] <- NA # replaces all zero values from column two to six with NA

diabetes <- na.omit(diabetes) # now we omit all NA values

nrow(diabetes) # at the end, we are left with 392 observations for our analysis

## [1] 392

diabetes$Outcome <- ifelse(diabetes$Outcome == 1, "Diabetes", "No Diabetes") # we substitute ones with "Diabetes" and zeros with "No Diabetes"

**Checking Assumption of Equal Variance-Coavariance matrices**

We want to use LDA and QDA in order to classify our observations into diabetes and no diabetes. From this post we know our assumptions of LDA and QDA so let’s check them:  
First, in order to get a sense of our data and if we have equal variances among each class, we can use a boxplots.

plot <- list()

box\_variables <- c("Pregnancies", "Age", "Glucose")

for(i in box\_variables) {

plot[[i]] <- ggplot(diabetes, aes\_string(x = "Outcome", y = i, col = "Outcome", fill = "Outcome")) +

geom\_boxplot(alpha = 0.2) +

theme(legend.position = "none") +

scale\_color\_manual(values = c("blue", "red"))

scale\_fill\_manual(values = c("blue", "red"))

}

do.call(grid.arrange, c(plot, nrow = 1))

[magicthumb id="diab\_box"]

covEllipses(diabetes[,1:8],

diabetes$Outcome,

fill = TRUE,

pooled = FALSE,

col = c("blue", "red"),

variables = c(1:5, 8),

fill.alpha = 0.05)

[magicthumb id="big\_cov"]

ggplot(diabetes, aes(x = Glucose, y = Insulin, col = Outcome)) +

geom\_point() +

stat\_ellipse() +

scale\_color\_manual(values = c("blue", "red"))

[magicthumb id="cov\_ellipses"]

From this scatterplot, we can clearly see that the varaince for the diabetes group is much wider than the variance from the non-diabetes group. This is becuase the blue points have a wider spread. The red points in contrast do not have as wide of a spread as the blue points.  
We are using the BoxM test in order to check our assumption of homogeneity of variance-covariance matrices.  
$latex H\_o$ = Covariance matrices of the outcome variable are equal across all groups  
$latex H\_a$ = Covariance matrices of the outcome variable are different for at least one group

boxm <- boxM(diabetes[, c(1:5, 8)], diabetes$Outcome) # using columns 1 to 5 and 8

boxm

##

## Box's M-test for Homogeneity of Covariance Matrices

##

## data: diabetes[, c(1:5, 8)]

## Chi-Sq (approx.) = 121.31, df = 21, p-value = 4.112e-16

When are choosing our alpha to be 0.05 then from our result we can conclude that we have a problem of heterogeneity of variance-covariance matrices. The plot below gives information of how the groups differ in the components that go into Box’s M test.

plot(boxm)

The log determinants are ordered according to the sizes of the ellipses we saw in the covariance ellipse plots. This plot confirms our visualizations that we have ellipses of different sizes and therefore, no equal variance-covariance matrices. It is worth noting that the Box M test is sensitive and can detect even small departures from homogeneity.  
As an additional check, we can perform a levene test to check for equal variances.

leveneTest(Pregnancies ~ Outcome, diabetes)

## Levene's Test for Homogeneity of Variance (center = median)

## Df F value Pr(>F)

## group 1 28.653 1.481e-07 \*\*\*

## 390

## ---

## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

From this test, we can see how the variances of the groups differ for pregnancies. They also differ for the variables Age, Glucose, and Insulin (not shown).

leveneTest(BloodPressure ~ Outcome, diabetes)

## Levene's Test for Homogeneity of Variance (center = median)

## Df F value Pr(>F)

## group 1 0.5761 0.4483

## 390

For BloodPressure, the variance seems to be equal. This is also the case for SkinThickness, and BMI (not shown).

diab.yes <- subset(diabetes, Outcome == "Diabetes")

diab.no <- subset(diabetes, Outcome == "No Diabetes")

**Checking Assumption of Normality**

With the following qqplots, we are checking that the distribution of the predictors is normally distributed within the diabetes group and non diabetes group.

* **Diabetes Group**

variable\_1 <- c("Pregnancies", "Glucose", "Insulin", "Age")

par(mfrow = c(2, 2))

for(i in variables) {

qqnorm(diab.yes[[i]]); qqline(diab.yes[[i]], col = 2)

}

[magicthumb id="qqplot\_yes"]

When looking at the distribution of Glucose, we can see that it is roughly normally distributed because the dots fall on the red line. We can see that the tails of the distribution are thicker than the tails from a normal distribution because the points deviate from the line at the left and at the right from the plot.

We can see that the distribution of the variable Pregnancies is not normally distributed because the points deviate from the line very much.

Insulin and age are also not normally distributed. The curve pattern in the plot is in the shape of a bow which indicated skewing. The points are above the line, then below it, and then above it again. This indicated that the skewing is to the right.

variable\_2 <- c("BMI", "SkinThickness", "BloodPressure", "Pedigree")

par(mfrow = c(2, 2))

for(i in variables) {

qqnorm(diab.yes[[i]]); qqline(diab.yes[[i]], col = 2)

}

[magicthumb id="qqq"]

The variables skin thickness and blood pressure look like they are normally distributed. The down-swing at the left and up-swing at the right of the plot for the blood pressure variable suggests that the distribution is heavier-tailed than the theoretical normal distribution.

For the BMI variable, the points swing up substantially at the right of the plot. These points might be outliers but we cannot infer that by looking at these plots.

The distribution for the pedigree function variable looks again right skewed and not normally distributed.

If one wants to test for normality, the shapito.test function is an option. Here is an example of how to use it:

shapiro.test(diab.yes$SkinThickness)

##

## Shapiro-Wilk normality test

##

## data: diab.yes$SkinThickness

## W = 0.99547, p-value = 0.9571

The null hypothesis is that the data is normally distributed. Based on the results, we fail to reject the null hypothesis and conclude that the data is normally distributed for skin thickness.

* **Non Diabetes Group**

par(mfrow = c(2, 2))

for(i in variables) {

qqnorm(diab.no[[i]]); qqline(diab.no[[i]], col = 2)

}

[magicthumb id="qqqq"]

All variables do not seem to be normally distributed and all distributions seem to be right skewed based on the bow shape of the points in all four plots

par(mfrow = c(2, 2))

for(i in variable\_2) {

qqnorm(diab.no[[i]]); qqline(diab.no[[i]], col = 2)

}

[magicthumb id="qqplot\_no"]

BMI, skin thickness, and bloodpressure seem to be roughly normally distributed whereas the pedigree function has this bow shape which suggests it is right skewed.  
Another visualization technique is to plot the density of the predictors for each group. Through the plots below, we can detect if the predictors in each group are normally distributed and we can also check for equal variance.

plot <- list()

for(i in names(diabetes)[-9]){

plot[[i]] <- ggplot(diabetes, aes\_string(x = i, y = "..density..", col = "Outcome")) +

geom\_density(aes(y = ..density..)) +

scale\_color\_manual(values = c("blue", "red")) +

theme(legend.position = "none")

}

do.call(grid.arrange, c(plot, nrow = 4))

[magicthumb id="dist"]

From the plots above we can conclude, that a lot of distributions are right skewed and that the variance is often also different.

In our next [post](http://thatdatatho.com/2018/04/02/73/), we are going to implement lda and qda and see, which algorithm gives us a better classification rate.