**Project:**

**1. Introduce your data:**

What is the population of interest?

What kind of variables you have?

Specify the type of variables.

**2. Data cleaning and manipulation:**

Explain any data cleaning

Explain any data manipulation (changing structure, creating new variables,…)

(Final variables)

**3. Research question**

What is the question?

Why is this interesting?

Why logistic regression?

**4. Initial Selection of the variables:**

Response: Binary (We will check within the two groups)

Numerical Variables: Two sample T-test: (To compare the mean of independent variables withing the two groups of response variable)

t.test(group1, group2, var.equal=T)

t.test(group1, group2, var.equal=F)

Variance Test: var.test(group1, group2) Use the p-value to decide what kind of T-test you need to use.

Categorical Variables: Chi-square Test: (To determine if there is an association between explanatory & response)

count <- matrix(c(a,b,c,d), nrow=2)

chisq.test(count)

**Based on these tests we will choose our initial variables.**

**5. Check for Multicollinearity:**

Occurs when an independent variable is highly correlated with one or more of the other independent variables

The results parameter estimates are unstable & the standard errors are large

To detect multicollinearity:

1. Examine the correlation coefficient for each pair of independent variable. A correlation that is greater than 0.8 indicates the problem. Some look at the pairs (), to visually inspect the scatterplot between every pair of quantitative explanatory variables

2. Or/and use Variance Inflation Factor (VIF) and check that all values are under 10. The value that is >10, usually indicates a multicollinearity. (Different cut off)

**6. Interaction terms:**   
Interaction.plot(mydata$cat1, mydata$cat2, mydata$y, …)

It does not mean that is significant (it means there is potential)

**Later:**

Choose one interaction terms, If you found more than one interactions in your model

We will compare models with and without interaction terms

Using Interaction.plot: If you didn’t find any interaction, still consider one interaction term to practice and compare the tow models.

**7. Splitting the dataset:**

Split your data to train and test data (Example 70% vs 30%).

Use training data to fit your model.

set.seed(#) # is used so that each time we get the same data set after splitting

sample\_size<- sample.split(mydata$y, SplitRatio = 7/10) #Splitting the dataset into 70/30 ratio

train<-subset(mydata, sample\_size==T)

test<-subset(mydata, sample\_size==F)

nrow(train)

nrow(test)

**8. Select explanatory variables**: Use **stepwise** **variable selection**s (Backward elimination/Forward selection) to choose your variables for two models (With and without interaction)

**9. Compare the model** with **interaction vs model with no interaction**

Likelihood ratio test (use Anova, in library (car)) to check the goodness of fit

Wald test

To check whether a parameter is significant for a given model.

**The result of step 9 confirms what is your final model.**

**10**. Using the final model on train data, predict the responses and prepare the **Classification report** (Including: Sensitivity, Specificity and Accuracy) to measure the quality of predictions

Note, if your data is unbalanced, means the number of 1 is much less than the number of 0, it will affect the result.

Do the classification in both models: regardless of efficiency of the model.

**11. ROC Curve:**

The receiver operating characteristic (ROC) curve, which is defined as a plot of test sensitivity as the y coordinate versus its 1-specificity or false positive rate (FPR)

Do the Roc curve in both models: regardless of efficiency of the model.

Plot both curves in one graph

**12. Lack of fit test** (Hosmer-Lemshow test- ungrouped data)

Interpret the coefficient