Thesis Project

Naby Diop

05/22/2020

Contents

Contents

1 Introduction

In 1987 the Indonesia National Contraceptive Prevalence Survey (NICPS) and Demographic and Health Survey (DHS) conducted a national survey for several reasons. One of the reason was to understand the decline of the natality rate at the time. In fact, according to the summary report of the NICPS team, the natality rate drop from 5.5 children per woman to 3.2 over a decade. The investigators was wondering whether or not the use of contraceptives can explain the decline in natality rate. However, the purpose of this paper is not to investigate the reason why the decline, but to use this data to explore the factors affecting the choice of contraceptive methods. We seek to examine the data using the models multinormal logistic, linear discriminant analysis, classification tree and k nearest neighbor. We will compare these models in terms of accuracy and finaly derive our conclusions based on the model that best fit the data.

2 Preliminary Study

2.1 Numerical Analysis of the Data

The data *contraceptives* is a subset of the Indonesia National Contraceptive Prevalence Survey (NICPS) and Demographic and Health Survey (DHS) data. All the information related to the methods of collection and the different questions of interest that motivates the survey can be found on the World Bank site. It has 1472 observations and 10 variables. All the subjects of the survey were married women who were not pregnant or do not know whether or not they were pregnant at the time of the survey. We seek to predict the contraceptive method used (CMU 1=no-use, 2=long-term, 3=short-term) given the nine explanatory variables (Age, education, husband education, number of children, religion, employement status, husband occupation, standard-of-living and media exposure). To begin our work, we first examine the stucture and summary of the data. The first thing that we realized is that the women are between 16 and 49 years old and they are majoritary religious (muslims) 1252 muslim vs 220 non-muslim. Most of these women do not work (1103 do not work vs 369 work). We also notice that most of them are exposed to media (1363 exposed vs 109 non-exposed).

Stucture of the data

```
## 'data.frame':    1472 obs. of  10 variables:
##  $ WAge     : int  45 43 42 36 19 38 21 27 45 38 ...
##  $ WEdu     : Factor w/ 4 levels "1","2","3","4": 1 2 3 3 4 2 3 2 1 1 ...
##  $ HEdu     : Factor w/ 4 levels "1","2","3","4": 3 3 2 3 4 3 3 3 1 3 ...
##  $ NumChild : int  10 7 9 8 0 6 1 3 8 2 ...
##  $ WReligion: Factor w/ 2 levels "0","1": 2 2 2 2 2 2 2 2 2 2 ...
##  $ WnWorking: Factor w/ 2 levels "0","1": 2 2 2 2 2 2 1 2 2 1 ...
##  $ HOcpt    : Factor w/ 4 levels "1","2","3","4": 3 3 3 3 3 3 3 3 2 3 ...
##  $ SLI      : Factor w/ 4 levels "1","2","3","4": 4 4 3 2 3 2 2 4 2 3 ...
##  $ MdExpo   : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 2 2 ...
##  $ CMU      : Factor w/ 3 levels "1","2","3": 1 1 1 1 1 1 1 1 1 1 ...
```

Summary of the data

```
##       WAge       WEdu    HEdu       NumChild      WReligion WnWorking HOcpt
##  Min.   :16.00   1:152   1: 44   Min.   : 0.000   0: 220    0: 369    1:436
##  1st Qu.:26.00   2:333   2:178   1st Qu.: 1.000   1:1252    1:1103    2:424
```

```
##  Median :32.00   3:410   3:351   Median : 3.000                      3:585
##  Mean   :32.54   4:577   4:899   Mean   : 3.262                      4: 27
##  3rd Qu.:39.00                   3rd Qu.: 4.250
##  Max.   :49.00                   Max.   :16.000
##  SLI     MdExpo   CMU
##  1:129   0:1363   1:628
##  2:229   1: 109   2:333
##  3:430            3:511
##  4:684
##
##
```

2.2 Graphical Analysis of the Data

The purpose of the the graphical analysis here,is to understand the relationship between the reponse and the explanatory variables. We're only going to explore the graphical relationship between CMU and the variables education, religion and age. We start with eduction variable. We want to understand how the choice of contraceptive method is associated with education level. Since the CMU is given at the indivial level, we group the data by education level and contraceptive methed used, we count the number of women in each education level and we count the number of women in each class in CMU variable. We use these numbers to compute the porpotion of choosing each class of CMU for each education level. All of these manupilations are facilitated by the *dplyr*, which has some intersting functions, such as *group_by()* function, that help for groupping the data. The resulting plot is shown below on Figure 1. This plot shows that, the proportion of women who do not use any contractive method falls with education level. The more a woman is educated the more likly she is to use a long or short term contraceptive method. Whereas the proportion of woman that uses a lon-term contraceptive method increases with the education level. The proportion of sort-term users seem to be constant over education level.

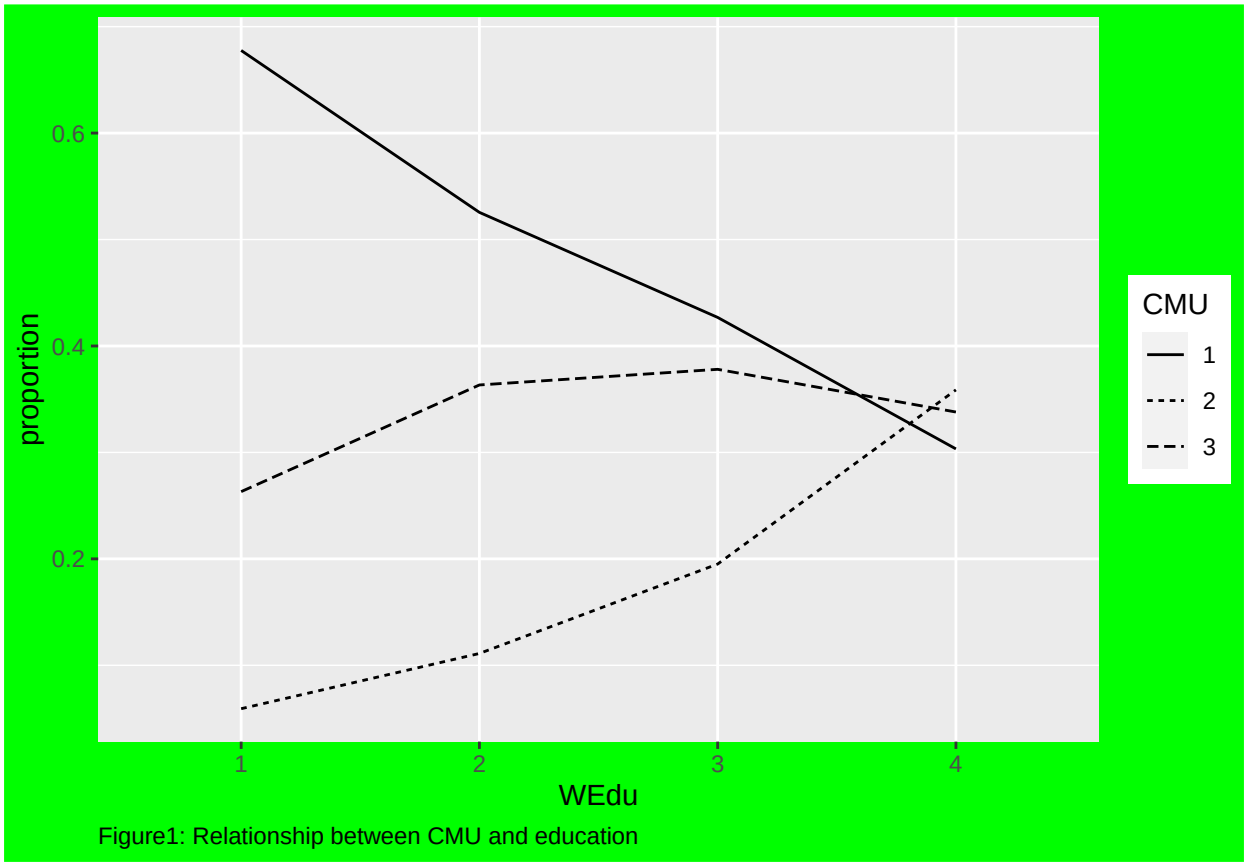


Figure1: Relationship between CMU and education

We use the same manupilation technique for religion variable. The plot is shown below on Figure 2. Muslim women tend to not use any contraceptive method, the short-term use seem to be the prefered method because the proportion of non-islam women is less then the proportion of islam women that use short-term CMU.

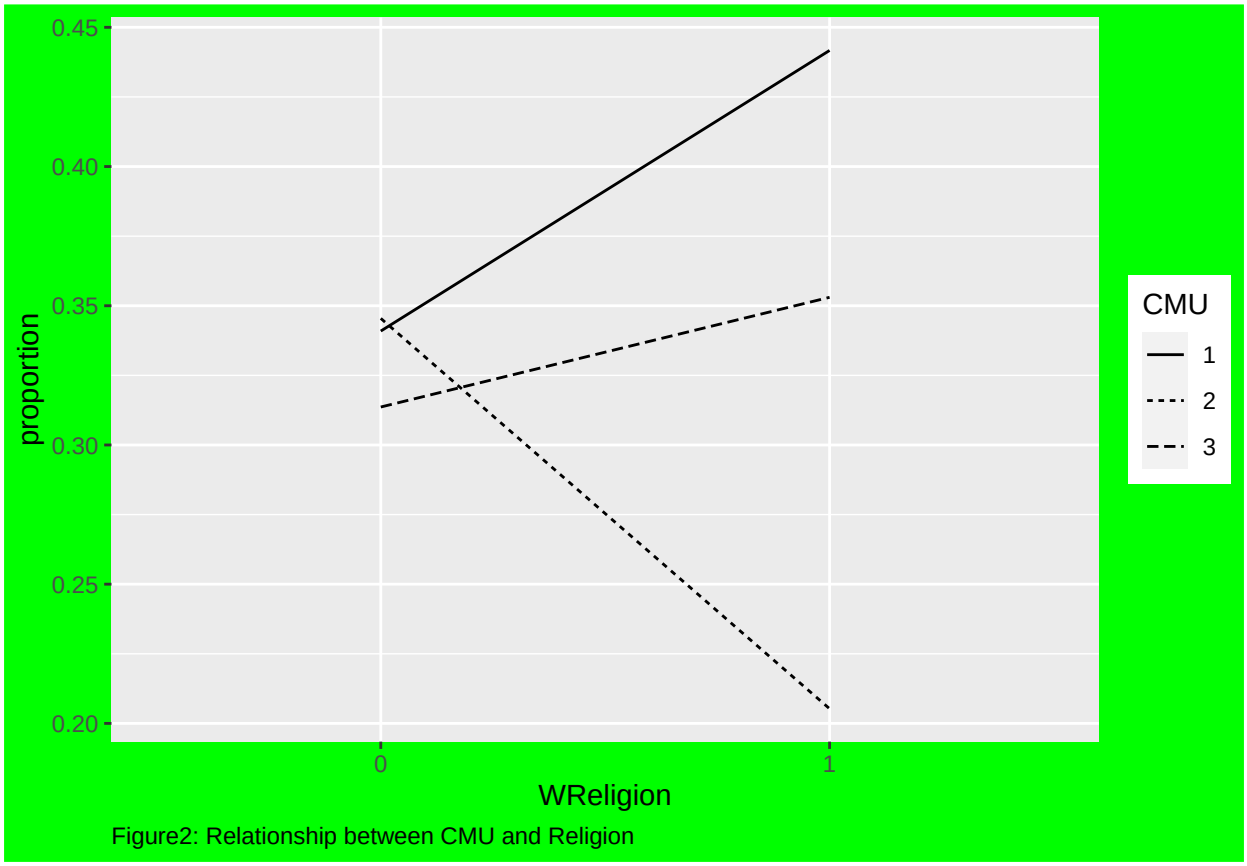
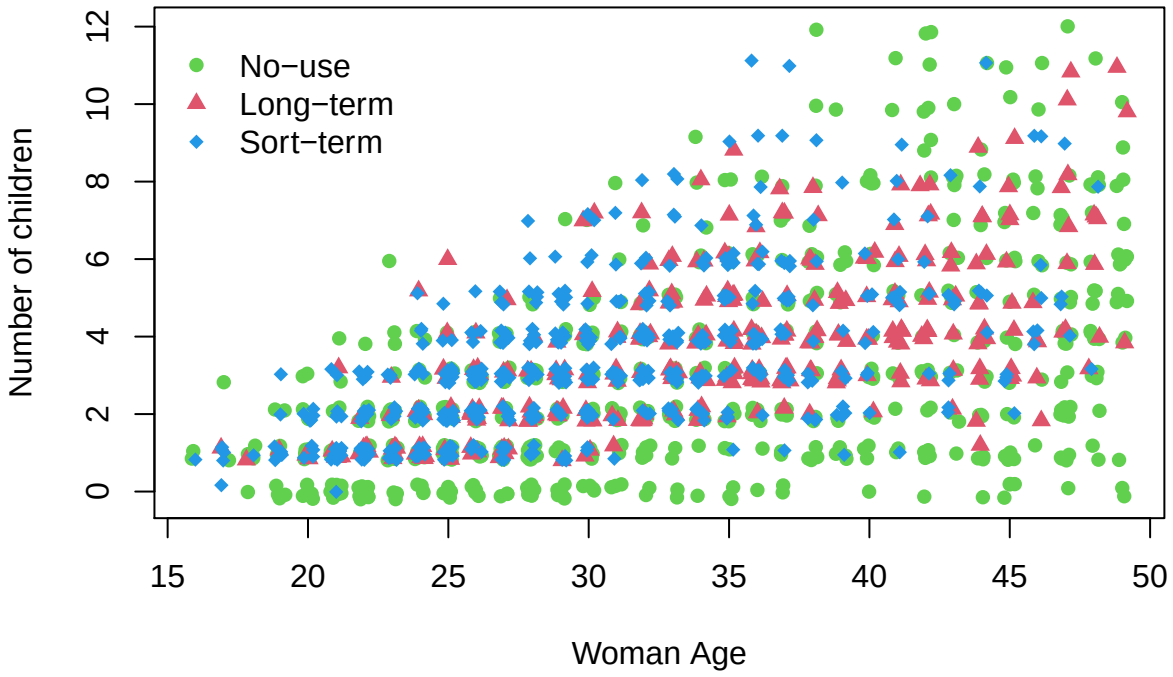


Figure2: Relationship between CMU and Religion

The plot Number of Children vs Women Age, on Figure 3, shows that mosly women over 35 years that have at least one child are the ones who use contraceptive methods. We also notice that all the women who have used a long or short-term contraceptive method have already at least a child. Young women (under 25 year old) mostly use short-term or do not use at all any contraceptive method.

Figure3: Number of children vs Woman Age



Splitting the dataset

For forwarder analysis, we split the data into training and test set. We use 75% of the data as trainning set and 25% as test set.

3 Multinomial model

3.1 Presentation of the model (Julian J. Faraway , 2013).

Multinomial distribution is an extention of the binomial. Its only difference with bionomial distribution is that it integretes the response variables that has more than two possiple outcomes. In this project, for example, the target variable is the contraceptive method used in Indonesia (CMU). It has three different outcomes: long-term, no-use and short-term. For simplification purpose we use the codding 1 for no-use, 2 for long-term and 3 for short-term. Let Y_i be the CMU, so Y_i falls into one of the categories 1, 2, 3 of CMU. Let $p_{ij} = P(Y_i = j$ where j is in C = {1,2,3} the set of the three diferent possible outcome of the random variable CMU. Let Y_{ij} be the number of observations falling into category j in the sample, and let $n_i = \sum_j^3 Y_{ij}$. Using the trainning set we have:

To find the probabilities p_{ij} we fisrt look at the summary of the train set to determine how many women have used each method 1, 2 and 3.

$p_{i1} = 0.423913$, $p_{i2} = 0.2400362$ and $p_{i3} = 0.3360507$

$$p_{ij} = P(Y_i = j),\ \sum_{j=1}^{3} p_{ij} = 1$$

To find Y_{ij} we sum all the individual in category j, in other words we Y_{ij} represent the number of women using contraceptive method j.

$Y_{i1} = \sum(Y_{i1}) = 468$, $Y_{i2} = \sum(Y_{i2}) = 265$ and $Y_{i3} = \sum(Y_{i3}) = 371$

Finally, n_i represent the sample size since it is sum of the Y_{ij}'s, $n_i = 1104$.

```
##       WAge       WEdu    HEdu       NumChild      WReligion WnWorking HOcpt
##  Min.   :16.00   1:114   1: 31   Min.   : 0.000   0:164     0:290     1:319
##  1st Qu.:26.00   2:249   2:134   1st Qu.: 2.000   1:940     1:814     2:321
##  Median :32.00   3:309   3:271   Median : 3.000                       3:446
##  Mean   :32.63   4:432   4:668   Mean   : 3.297                       4: 18
##  3rd Qu.:39.00                   3rd Qu.: 5.000
##  Max.   :49.00                   Max.   :16.000
##  SLI     MdExpo   CMU
##  1: 92   0:1026   1:468
##  2:172   1:  78   2:265
##  3:319            3:371
##  4:521
##
##
```

In general, the multinomial distribution is given as follow:

$$P(Y_{i1} = y_{i1}, Y_{i2} = y_{i2}, Y_{i3} = y_{i3}) = \frac{n_i!}{y_{i1}!y_{i2}!y_{i3}!} p_{i1}^{y_{i1}} p_{i2}^{y_{i2}} p_{i3}^{y_{i3}}$$

Thus, in this paticular cas we have the following:

$$P(Y_{i1} = 468, Y_{i2} = 265, Y_{i3} = 371) = \frac{1104!}{468!265!371!} p_{i1}^{468} p_{i2}^{265} p_{i3}^{371}$$

Now we have the probabilities p_{ij} we must link them to the nine predictors in the data set. Because we are fitting multinomial logistic model we choose logit function as link function. This allow as to keep the probalities p_{ij} between 0 and 1. To fully determine the formula of the link function we must declare a one class as baseline. The choice of the class does not matter, here we will be choosing 1. We denote by x_i the predictors and η_{ij} represents the link function. Thus, we have the formula:

$$\eta_{ij} = x_i^T \beta_j = log \frac{p_{ij}}{pi1}$$

Since $\sum_1^3 p_{ij} = 1$ so we have $p_{i1} = 1 - \sum_2^3 p_{ij}$. Therefore:

$$p_{ij} = \frac{exp(\eta_{ij})}{1 + \sum_2^3 exp(\eta_{ij})}$$

Multinomial logistic regression does necessitate careful consideration of the sample size and examination for outlying cases. It also does not assume normality of the variables, linearity, or homoscedasticity. Therefore we can fit it without having to transform the categorical variables to numerical.

3.2 Fitting the multinomial logit model

The function *multinorm* from the *nnet* package is the used to fit multinormal logit model in R. Although, multinorm function is part of the neural networks package in it doesn't have a deep connection to neural net. The program use the optimisation method in neural net to comput the maximun likelihood. For more detail see Faraway (2016)

```
## Call:
## multinom(formula = CMU ~ ., data = Train_set)
##
## Coefficients:
##   (Intercept)        WAge     WEdu2     WEdu3     WEdu4      HEdu2     HEdu3
## 2   -1.288588 -0.04686179  1.138590 2.1041660 3.1236423 -1.9648094 -1.624228
## 3    0.204446 -0.11185221 -0.202898 0.1786879 0.8780549  0.9655539  1.173986
##       HEdu4  NumChild WReligion1 WnWorking1     HOcpt2     HOcpt3    HOcpt4
## 2 -1.488860 0.3209690 -0.4392366  0.1218311 -0.3448079 -0.1202911 0.2755218
## 3  1.156418 0.3331207 -0.3392766  0.1415519  0.1680364  0.4405653 0.3363384
##        SLI2      SLI3      SLI4    MdExpo1
## 2 0.6736114 1.0416120 1.0920604 -0.7288695
## 3 0.3918101 0.7037839 0.7999449 -0.5099031
##
## Std. Errors:
##   (Intercept)       WAge     WEdu2     WEdu3     WEdu4     HEdu2     HEdu3
## 2   0.9174878 0.01397862 0.5597474 0.5641651 0.5832660 0.6182577 0.5652179
## 3   0.9040067 0.01336142 0.3016847 0.3133581 0.3446905 0.6883582 0.6811676
##       HEdu4   NumChild WReligion1 WnWorking1    HOcpt2    HOcpt3    HOcpt4
## 2 0.5737593 0.05000788  0.2384939  0.1934631 0.2374827 0.2343516 0.7762434
## 3 0.6902654 0.04484951  0.2369451  0.1726470 0.2248902 0.2189322 0.6191709
##        SLI2      SLI3      SLI4   MdExpo1
## 2 0.4919445 0.4673269 0.4694193 0.5073495
## 3 0.3117192 0.2959485 0.2978777 0.3395335
##
## Residual Deviance: 2053.788
## AIC: 2125.788
```

The summary of the multinormal logistic model shows two blocks. The first block correspond coefficients of model equation and the second one correspond to the standard errors. Here, the baseline of the contraceptive method use is 1 corresponding to "no-use" class. The first row compares the CMU="long-term" to our baseline CMU = "no-use. The second row compares the CMU ="sort-term" to CMU = "No-use". Thus, considering the coefficients of the output we can write the model equations. Let's consider α_1 and α_2 to be the intercept in respectively row 1 and row 2. β_{1j} and β_{2j} the coefficients under each outcome of the fariables respectively on row 1 and row 2. In total we have 17 coefficiences. Thus, the equations of the model are given bellow (see detailed equation in the next paragraph for the reduice model).

$$log(\frac{P(CMU=2)}{P(CMU=1)}) = \alpha_1 + \sum_{j=1}^{17} \beta_{1j} X$$

$$log(\frac{P(CMU=2)}{P(CMU=1)}) = \alpha_2 + \sum_{j=1}^{17} \beta_{2j} X$$

As stated above, the β_{ij}'s represent the model coefficients. For example β_{11} represent the coefficients for Wage variable it can be interpreted as the follow: one-unit increase in woman age is associated with a 0.04686179 decrease in the relative log odds of a woman using long-term contraceptive method program vs. not using any method at all. If we takes the example of education, the relative log odds of being a woman using long-term CMU vs. no-use any CMU increases by 1.138590 if moving from education level 1 to the education level 2 (level 1 is the reference group). We can calculate the accuratie of the model from the table bellow.

```
##                         Test_set$CMU
```

```
## predict(mmod, Test_set)   1   2   3
##                        1 104  18  57
##                        2  18  28  27
##                        3  38  22  56
```

The test error is given bellow:

```
## [1] 0.4891304
```

The misclassification rate is 49%, only 51% of the CMU were well classified. This rate is not so bad when we compare it to the misclassification rate of ramdom guessing of 67% ($\frac{2}{3} \times 100$). We can even do better by reduicing this rate using variable selection based on ACI criterion using stepwise method. The model with the smallest ACI gives the best prediction.

```
## Call:
## multinom(formula = CMU ~ WAge + WEdu + HEdu + NumChild + SLI,
##     data = Train_set)
##
## Coefficients:
##   (Intercept)        WAge      WEdu2     WEdu3     WEdu4      HEdu2     HEdu3
## 2  -2.2099359 -0.04314132  1.2908778 2.2721651 3.3461381 -1.9472842 -1.558545
## 3   0.2307249 -0.11397374 -0.0817927 0.2832326 0.9256335  0.9484896  1.212580
##       HEdu4  NumChild      SLI2      SLI3      SLI4
## 2 -1.345214 0.3068511 0.6807813 1.1361146 1.2413023
## 3  1.147633 0.3263993 0.3723338 0.7187738 0.8073303
##
## Std. Errors:
##   (Intercept)       WAge     WEdu2     WEdu3     WEdu4     HEdu2     HEdu3
## 2   0.7626407 0.01352449 0.5447145 0.5477746 0.5603900 0.6128048 0.5589724
## 3   0.7692971 0.01293584 0.2891406 0.2991110 0.3223799 0.6817296 0.6751724
##       HEdu4   NumChild      SLI2      SLI3      SLI4
## 2 0.5634250 0.04868812 0.4898026 0.4626392 0.4622064
## 3 0.6830575 0.04384811 0.3083219 0.2918635 0.2911822
##
## Residual Deviance: 2071.255
## AIC: 2119.255
```

Using the sepwise selection, we see that only the WAge, WEdu, HEd, NumChild and SLI variables are significant for predicting the contraceptive method used.

The equation of this reduice model is:

$$\begin{aligned} log(\frac{P(CMU=2)}{P(CMU=1)}) =& -2.209 - 0.043 \times WAge + 1.291 \times (WEdu=2) + 2.272 \times (WEdu=3) \\ &+ 3.346 \times (WEdu=4) - 1.345 \times (HEdu=2) - 1.558 \times (HEdu=3) - 1.345 \times (HEdu=4) \\ &+ 0.307 \times NumChild + 0.680 \times (SLI=2) + 1.136 \times (SLI=3) + 1.241 \times (SLI=4) \end{aligned}$$

$$\begin{aligned} log(\frac{P(CMU=3)}{P(CMU=1)}) =& 0.230 - -0.113 \times WAge - 0.289 \times (WEdu=2) + 0.283 \times (WEdu=3) \\ &+ 0.925 \times (WEdu=4) + 0.948 \times (HEdu=2) + 1.212 \times (HEdu=3) + 1.147 \times (HEdu=4) \\ &+ 0.326 \times NumChild + 0.372 \times (SLI=2) + 0.718 \times (SLI=3) + 0.807 \times (SLI=4) \end{aligned}$$

NB: All the coefficients are round to three decimale place.

The cross table of predicted vs. oberved for the reduice model is shown bellow.

```
##                          Test_set$CMU
## predict(mmodi, Test_set)   1   2   3
##                        1 102  17  52
##                        2  17  32  24
##                        3  41  19  64
```

Misclassification rate:

```
## [1] 0.4619565
```

The spetwise selection reduice the misclassification rate (46%), 54% of the data is correctly classified by the multinormal logistic model.

4 Linear Discriminant Analysis (LDA).

4.1 Presentation of LDA model.

Linear Discriminant Analysis (*LDA*) is a statistical mmethod that aims to predict the appartenance of an individual to a class given the predictor variables. *LDA* unlike many other statistical models doesn't modelized directly the probability P(Y=c|X=x) (where Y is the response variable and X the preditors, c is the class). It modelized instead, at first, the probability P(X=x|Y=c) then, we use *bayes' theorem* to finally determine P(Y=c|X=x).

Bayes' theorem:

$$P(Y = y|X = x) = \frac{P(X = x|Y = c)}{\sum_{j=1}^{J} P(Y = j)P(X = x|Y = j)}$$

C is the number of class and $p_{ic} = P(Y_i = c)$ the probality of an individual i belonging to class c. In our context for example Y is the contracetive method used, J=3. If we consider the trainning set, we have $p_{i1} = 0.423913$, $p_{i2} = 0.2400362$ and $p_{i3} = 0.3360507$. X = (WReligion, WEdu, HEdu, HOcpt, SLP, WnWorking, MdExpo, WAge, NumChild). LDA model have several assumptions. Among these assumptions the most important once are:

- The groups must be mutually exclusive. For example the choice of one womon must be only one contraceptive methode. In other word a woman can not chose at the same time long and short term method.
- LDA is also very sensitive to outliers. Each group should have the same variance for any predictor. In general a log transformation a variable solve this problem.
- The independent variables should be multivariate normal. In this case we have many variables that are categorical, which leads to a violation of the assumtion. An approach to over come this contraint we can transform the categorical variables to numerical variables using the *optimal scaling* methodology. Optimal scaling process turns qualitative variables into quantitative ones. Optimality is a relative notion, because it is always obtained with respect to the particular data set that is analyzed and the particular criterion that is optimized. Optimal scaling of a data vector is obtained through a least-squares transformation subject to appropriate measurement. We will not develop in detail this methodologie. However more information can be found in the paper of Jacqueline J. Meulman, Ph.D., Data Theory Group in the Faculty of Social and Behavioral Sciences at Leiden University. This method can be implimanted in R using Optiscal package. Thus we will transform the categorical variable first than fit the LDA using the transform dataset.

- The number of cases for each group must not be greatly different. In other words p_{ic} must be approximatively equal. This is not a big deal in this case because all classes have reasonable number of observation when we compare them to each other.

4.2 How LDA model works (Robert Gentleman, Kurt Hornik, and Giovanni Parmigiani)?

Given k groups, linear classification rule consist of assigning an observation to one group among the k group. This means that when we have y that we don't know to which group it belongs to we comput the distance between y and y_i from the i^th group and assign y to the group for which the distance is smallest. This distance is called the *linear discriminant function*. The formula of such distance is given below:

$$D_i^2(X) = (X - \overline{X_i})' S_{pl}^{-1} (X - \overline{X_i})$$

Where S_{pl} is the pooled sample covariance given by the below formula.

$$S_{pl} = \frac{1}{N - C} \sum_{i=1}^{C} (n_i - 1) S_i$$

S_i and n_i are the simple covariance matrix and sample size in the i^{th}, N the Sample size (1472) and C number of classes (C=3). We dirive the classification rule as stated above (smallest distance).

$$\begin{aligned} D^2(X) &= (X - \overline{X_i})' S_{pl}^{-1} (X - \overline{X_i}) \\ &= X' S_{pl}^{-1} X - X' S_{pl}^{-1} X_i - X_i' S_{pl}^{-1} X + X_i' S_{pl}^{-1} X_i \\ &= X' S_{pl}^{-1} X - 2 X_i' S_{pl}^{-1} X - X_i' S_{pl}^{-1} X_i \end{aligned}$$

Since the term $X' S_{pl}^{-1} X$ does not depend on a given group we can delete it and multiply $D_i^2(x)$ by $-\frac{1}{2}$. We then denote $L_i(X) = -\frac{1}{2} D_i^2(X)$ so, for $D_i^2(X)$ mininum, $L_i^2(X)$ will be maximized. The classification rule become: We assign y to a group for which $L_i(y)$ is maximal.

Let denote $c_i' = X' S_{pl}^{-1}$ and $c_{i0} = \frac{1}{2} X_i' S_{pl}^{-1} X_i$

$$\begin{aligned} L_i(X) &= X_i' S_{pl}^{-1} X - \frac{1}{2} X_i' S_{pl}^{-1} X_i \\ &= c_i' X + c_{i0} \end{aligned}$$

Now if we assume the distribution of y given a group G_i, $f(X|G_i)$ to be normaly distributed with same covarience and the prorpotions of observation p_i's equal to each other for all groups. The classification rule is then equivalent to maximizing $p_i f(y|G_i)$

$$f(X|G_i) = N(\mu_i, \Sigma) = \frac{1}{|\Sigma|\sqrt{2\pi}} e^{-\frac{1}{2}(X - \mu_i)' \Sigma^{-1} (X - \mu_i)}$$

To maximize $p_i f(X|G_i)$ we maximize $ln(p_i f(X|G_i))$

$$\begin{aligned} ln(p_i f(X|G_i)) &= ln(p_i) + ln(f(X|G_i)) \\ &= ln(p_i) - \frac{1}{2} ln(2\pi) - \frac{1}{2} |\Sigma| - \frac{1}{2} (X - \mu_i)' \Sigma^{-1} (X - \mu_i) \end{aligned}$$

Since S_{pl} is an estimator of Σ and $\overline{X_i}$ estimator of μ_i we can replace them in the equation above.

$$
\begin{aligned}
ln(p_i f(X|G_i)) &= ln(p_i) + ln(f(X|G_i)) \\
&= ln(p_i) - \frac{1}{2}ln(2\pi) - \frac{1}{2}|S_{pl}| - \frac{1}{2}(X - \overline{X_i})' S_{pl}^{-1}(X - \overline{X_i}) \\
&= ln(p_i) - \frac{1}{2}ln(2\pi) - X' S_{pl}^{-1} X + X_i' S_{pl}^{-1} X - \frac{1}{2} X_i' S_{pl}^{-1} X_i
\end{aligned}
$$

As we have done above the terms $X' S_{pl}^{-1} X$ and $\frac{1}{2}ln(2\pi)$ will be negligated they are constant regading groups. Therefore the rest of the above formula will be denoted by $L_i'(X)$ and it is

$$
L_i'(X) = ln(p_i) + X_i' S_{pl}^{-1} X - \frac{1}{2} X_i' S_{pl}^{-1} X_i
$$

We assign y to the group for which $L_i'(X)$ is the smallest this ullistrates the classification rule.

4.3 Fitting LDA model (Julian J. Faraway , 2013)

As stated above we will fit the LDA model on the transform data set. In pratice the implimantation of *LDA* model is litle bit different from the theory explained above. *LDA* algorithm first computes a measure of total variation in the data using the total sum of square S.

$$
S = \sum_{i=1}^{n} (X_i - \overline{X})(X_i - \overline{X})^T
$$

Than we difine the within-group covariance W as: $W = \sum_{c=1}^{C} \sum_{i=1}^{n_c} (X_{ic} - \overline{X}_c)(X_{ic} - \overline{X}_c)^T$ Where n_c is the total number of indivial in categoriry c, n the sample size. We also define the between-groups covariance B as: $B = \sum_{c=1}^{C} (X_c - \overline{X})(X_c - \overline{X})^T$ Thus S=W+B.The goal is to form a linear combinations a of the predictors to maximize the separation betweem the groups. Therefore, we choose a to maximize

$$
\frac{a^T B a}{a^T W a}
$$

The solution can be found using the eigendecomposition of $W^{-1}B$

The implimentation can be performed using *MASS* library in R.

```
## Call:
## lda(New_Train_set$CMU ~ ., data = New_Train_set)
##
## Prior probabilities of groups:
##         1         2         3
## 0.4375000 0.2219203 0.3405797
##
## Group means:
##       WAge     WEdu     HEdu NumChild WReligion WnWorking    HOcpt      SLI
## 1 33.63354 2.693699 3.297384 2.937888  1.825311  1.726708 2.171811 2.975580
## 2 34.48571 3.435610 3.637856 3.722449  1.954148  1.726531 1.879964 3.450922
## 3 30.11170 3.003171 3.481834 3.316489  1.839385  1.797872 2.211256 3.156525
##     MdExpo
## 1 1.023874
## 2 1.127690
## 3 1.097566
##
## Coefficients of linear discriminants:
```

```
##                   LD1         LD2
## WAge      -0.07052404 -0.11707132
## WEdu       0.56470688 -0.30575135
## HEdu       0.08827405  0.13818448
## NumChild   0.34191105  0.14872625
## WReligion  0.69826206 -0.36498916
## WnWorking  0.08108917  0.36139131
## HOcpt     -0.18490104  0.29972510
## SLI        0.29917191  0.02728084
## MdExpo     0.74609092  0.09450928
##
## Proportion of trace:
##    LD1    LD2
## 0.7038 0.2962
```

The output of LDA shows that the mean in the groups do not have too much difference for all the variables in the data set. For example the mean in each class range in between 30.11170 and 34.34.48571. The porportion oftrace is calculated using the eigenvalues of the decomposition. For this analysis the first componant (LD1 = 0.7038) is strongly dominant therefore, the clasification will mostly depend on this. The coefficients reprepents the a. The first combination, which count for 70.38% of the variation in the data, is dominated respectively by Mdexpo, Wreligion and WEdu variables. The second combinatio, with count for only 29.62% is dominated by WnWorking variable.

Now we determined the the most likly outcome from each case and compare it against the observe class.

```
##
## lda.class  1  2  3
##         1 93 40 49
##         2  9 22 21
##         3 43 26 65
```

```
## [1] 0.4565217
```

Finaly we compute test error rate (misclasssification rate.) which is 45.65%. This means that only only 54.35% of the data are correctly classified.

5 Classification Tree

5.1 Model presentation

Decision tree can be use in both regression and classification problems. In this paper we will be using classification tree methodology because the target variable is categorical. Classification tree provides information on how to make decisions at critical nodes to classify data cases into the known groups, or classes. It starts with the root node, extends to branches that connect a collection of nodes, and ends in leaf nodes. The analysis begins at the root node that includes all data points, tests the decision nodes by assigning each possible outcome in a branch, and decides whether each branch leads to another decision node or terminates in a leaf node. It uses input characteristics to create a model to sort cases into categories of different values on a target variable.

Classification tree only works with categorical variables, when working with continuous variables, we need to decide how to convert them into discrete classes to facilitate decision tree classification. For example number of children in the dataset can be categorized as 1 representing women that have up to 5 children, 2

representing women with more than 5 children but less than or equal to 10 and 3 women with more than 10 children. However, if we do not categorized coutinuous variables the algorithm will automaticly dicide by dividing the variable into two distinct classes. For example the algorithm may dicide to split on the mean of a continuous variable.

5.2 How the classification tree works (Shu, Xiaoling 2020)?

Here the target variable is CMU (contraceptive method used) that is repesented with 3 classes no-use coded as 1 long-term coded as 2 and sort-term coded as 3. The tree induction approach is elegant and attractive because the divide-and- conquer method allows us to reduce the large data into smaller sets and recursively apply the same method to partition the subsets. The goal of each step is to select an attribute that can partition the current group into subgroups that are as pure as possible with respect to the target variable. At first we take the whole data and we partion it into at least 2 subgroups, the subgroups are smaller versions of the same classification. We then take each data subset and apply attribute selection recursively to find the best attribute to partition these subsets. We choose the splitting attributes by testing all of them and selecting whichever ones yield the purest subgroups. We stop when the nodes are pure or when we run out of variables to split. We may also decide to stop earlier.

The bigest challenge is to decide where to split the data. Classificatio trees use multiple algorithms to decide how to split a decision node into two or more subnodes. These subnodes contain instances with higher purity than their parent nodes with respect to CMU. Constructing these subnodes thus increases homogeneity as we move downstream from parent nodes to child nodes. Decision trees split the nodes on all available variables and then select the splits that result in the most homogeneous subnodes. There are four most commonly used algorithms in decision trees: Chi-squared, variance reduction, Gini index, and information gain. For this project we will be using information gain to make a choice of where to split at first. For simplicity, we will only compare two variables, WReligion and WEdu variables, just to show how the algorithm will choose which variable to split on first between these two variables. Information theory uses entropy as a measure to define the degree of disorganization in a system. Purer nodes require less information to describe them, while more impure nodes require more information. When the sample is completely homogeneous, the entropy is zero; if the sample is equally divided (50% – 50%), it has an entropy of one. The formula of the entropy is given as follows.

$$Entropy = -\sum_{j=1}^{c} p_j log_2 p_j$$

where p_j is proportion of women in classe j, c the number of classes in the target variable, in this case 3 different classes.

And the information gain is:

Informatin gain = Entropy(of the total simple)-Information taken from a predictor

Where the Information taken from a predictor is given by the formula:

$$\sum_i WeightedAverage \times Entropy_i$$

The WeightedAverage represents the quotient of the sum of indivials in class j with total sample size.

Lets compute the information gain for the 2 variables Education and religion for more clarity.

Cross table of CMU with Wreligion

```
##    WReligion
## CMU   0   1
##   1  75 553
```

```
##   2  76 257
##   3  69 442
```

Proportion of women in WReligion = 0 (no Islam)

$$p_{01} = \frac{75}{75+76+69} = 0.3409091$$
$$p_{02} = \frac{76}{75+76+69} = 0.3454545$$
$$p_{03} = \frac{69}{75+76+69} = 0.3136364$$
$$E_{rel0} = -p_{01} \times log(p_{01}) - p_{02} \times log(p_{02}) - p_{03} \times log(p_{03}) = 1.583668$$

proportion of women in WReligion= 1 (Islam)

$$p_{11} = \frac{533}{553+257+442} = 0.4416933$$
$$p_{12} = \frac{257}{553+257+442} = 0.2052716$$
$$p_{13} = \frac{442}{553+257+442} = 0.3530351$$
$$E_{rel1} = -p_{11} \times log(p_{11}) - p_{12} \times log(p_{12}) - p_{13} \times log(p_{13}) = 1.519926$$

E_{rel0} and E_{rel1} represent the entropies for non muslims and muslims respectively.

The entropy of the total sample (E_s) is:

$$p_1 = \frac{628}{1472} = 0.4266304$$
$$p_2 = \frac{333}{1472} = 0.2262228$$
$$p_3 = \frac{511}{1472} = 0.3471467$$
$$E_s = -p_1 \times log(p_1) - p_2 \times log(p_2) - p_{13} \times log(p_3) = 1.539246$$

$Information\ taken\ from\ WReligion = \frac{75+76+69}{1472} \times E_{rel0} + \frac{553+257+442}{1472)} \times E_{rel1} = 1.529453$

$Information\ gain\ from\ WReligion = E_s - Information\ taken\ from\ WReligion = 0.009793119$

The Same calculation is carred over for Wedu variable

Cross table of CMU with WEdu variable

```
##    WEdu
## CMU   1   2   3   4
##   1 103 175 175 175
##   2   9  37  80 207
##   3  40 121 155 195
```

$E_{WEdu1} = 1.128743,\ E_{WEdu2} = 1.370687,\ E_{WEdu3} = 1.514808\ and\ E_{WEdu4} = 1.581538$

E_{WEdu1}, E_{WEdu1}, E_{WEdu1} and E_{WEdu1}represent respectivelly the entropy for WEdu level from 1to 4

$$
\begin{aligned}
Information\ taken\ from\ WEdu &= \frac{103+9+40}{1472} \times E_{WEdu1} + \frac{175+37+121}{1472} \times E_{WEdu2} \\
&+ \frac{175+80+155}{1472} \times E_{WEdu3} + \frac{175+207+195}{1472} \times E_{WEdu4} \\
&= 1.468496
\end{aligned}
$$

$Information\ gain\ from\ WEdu = E_s - Information\ taken\ from\ WEdu = 0.0707493$

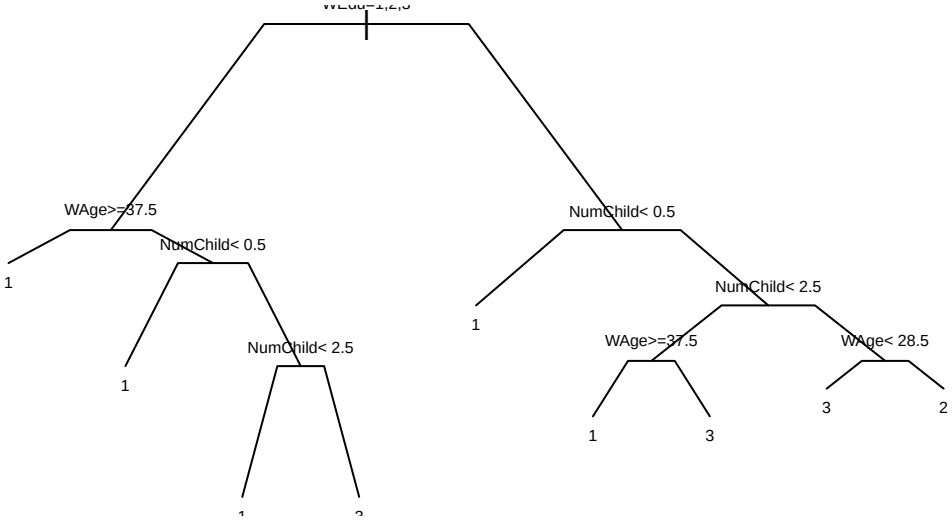
Since information gain from women relegion is less then infortion gain from women education, we will first split on education variable before religion. The same process of decision making is used at every node to determine the split variable.

NB: It is important to highlight that classification tree algorithm first use variable selection. This mean it may find some variables statisticaly unsignicant, so these variables won't be use in model.

Although classification trees are very easy to interprets, it is not great in prediction. One way to overcome this predictibily problems is to build a lot of trees and aggregate the result. This method is known as random forest.

5.3 Fitting the classification tree model

```
##
## Classification tree:
## rpart(formula = Train_set$CMU ~ ., data = Train_set)
##
## Variables actually used in tree construction:
## [1] NumChild WAge     WEdu
##
## Root node error: 636/1104 = 0.57609
##
## n= 1104
##
##         CP nsplit rel error  xerror     xstd
## 1 0.081761      0   1.00000 1.00000 0.025817
## 2 0.030922      1   0.91824 0.93396 0.026046
## 3 0.029874      4   0.82547 0.90409 0.026099
## 4 0.022013      5   0.79560 0.83648 0.026104
## 5 0.011006      7   0.75157 0.78459 0.026001
## 6 0.010000      8   0.74057 0.78616 0.026005
```



We can access to the variable importance in the tree model. Numchild, WAge, WEdu are the three most important variables.

```
##    NumChild        WAge        WEdu        HEdu       HOcpt         SLI
## 54.56748104 42.69830622 35.21736059 11.86853006 11.31742145  5.73567887
##   WReligion      MdExpo   WnWorking
##  4.25907102  0.26240776  0.07796465
```

Misclassification rate

```
##       predicted
## actual   1   2   3
##      1 104  13  43
##      2  18  27  23
##      3  38  19  83
```

```
## [1] 0.4184783
```

We have approximatively 42% of misclassification rate, which is large, We must do better if we hope to see more accurate analysis of the data. Therefore we prune the tree and compute the misclassification error.

```
## [1] 0.4184783
```

Prunning the tree does not change the misclassification rate. However, the error rate for decision tree is better than the one for LDA and multinomial logistic model.

6 K-nearest neighbors

6.1 How it works?

K-nearest neighbors (*KNN*) classifier is a supervised learning method. It is easy to impliment and not hard to interpret. As for decision tree, *KNN* can be use in both regression and classification problems. Here we wish to predict the CMU. We have a data set 1472 women, 628 do not use any contraceptive method, 333 women use the long-term and 511 use the sort-term method. For each women we have 9 variables that

discribes their life, which might affect their choice of CMU. Therefore, we can identify each women in a 9 dimensional space where we determine their nearest neighbores in term of having sumilar values on the 9 variables. Finaly we can assign women to the classe to which most of its neighbors belong.

However KNN classifier presents many challenges when inplimenting it. The first challenge that we encounter is the choice of the number of neighbor k. The choice of k is very important. A good choice raise the accuratie of the model whereas a bad choice increase the misclassification rate. In an early formulation, Cover and Hart (1967), two eminant statisticians, argued that using just one neighbor can be sufficient and even preferable. However, Hastie and Tibshirani (1996), suggested that whether a single neighbor is optimal is highly dependent on the number of features used to determine distance. A solution to the problem of how many neighbors to use involves, some other techniques. One can use cross-validation to select the best value of k or we can fit *KNN* for multiple k and choose the more accurate one. This methods will be explain more closly when fitting the model in the below sexion.

A second preliminary issue is determining the criterion for deciding which data points are closest, that is, what type of distance one will use to determine which points are nearest. Most commonly, KNN techniques use Euclidian, Manhattan (city-block), or Minkowski distance, though other types of distance (Mahalanobis, for instance) could be employed. Third, and related to the prior two matters, is the matter of "vote counting." That is, after we choose k, and determine how to measure distance, we will get, for each woman, a set of k other women which provide information for class prediction. These k women are in essence "voting" on the membership or class of the target case. But since these k women might not agree, how should we tally these votes? Ought we to count them all equally? Or ought we to regard the votes of the closer data points as more informative? In other words, when we increase k, we are increasing the size of the space around the data point where we are searching for information about class membership, and in doing so we increase the chance that we will make an error because we might "cross over" from a space in which one class is dominant to that in which the other is dominant. This is particularly important for boundary cases (that is, the cases in one class which are most similar to cases in the other class). But weighting by inverse distance downplays the importance of cases which are further away and increases the influence of cases which are closer (Attewell, Monaghan and Kwong 2015). The final challenge is the number of predictors which are made use of in determining distance. It would seem, intuitively, that choosing as many predictor variables as possible would be ideal since it would give us more information about which cases are "really" similar rather than just similar on a small number of rather arbitrarily chosen characteristics. However, it turns out that having too much information can be a problem. Increasing the number of features or predictors increases the dimensionality of the search space and therefore the search space's overall size (think about moving from a circle surrounding a point to a sphere with the same radius around that point). By doing so, we end up increasing the number of "neighbors" which are equidistant from the point in question (the one we want to classify). With a large enough number of features, we end up with a search space described by an n-sphere (that is, a sphere in n dimensions, where n is the number of features, in this case n=9), the surface of which is occupied by a large number of data points which "tie" in terms of distance from the point at the center. In this situation the KNN method is fatally compromised by the curse of dimensionality (Hastie and Tibshirani 1996). To solve this problem we can do some preliminary variable selection using stepwise selection, LASSO etc (Attewell, Monaghan and Kwong 2015).

6.2 Fitting the model

As in LDA, KNN model assumes that all the predictors are numerical, therefore we will be using the transform data set used in LDA. Before we fit the model we must first scaled the data so that the output remains unbiased. The porpuse of sclaing the data is to remove all the unit that might put the data in different scales for example Age variable could be in much larger scale then number of children. We use *scale()* function in **R** to perform this task which scales the data points to there z-score metric. We apply scale function to both training and test set than we fit the model. The choice of k is very important, for a small data set we use the squart root of the trainning set size. Here we be using at first k=10. We use the class packege in r to fit knn model.

We access the cross table of predicted class vs. observed class than we examine the model accuracy.

```
##    m1
##      1  2  3
##   1 77 18 50
##   2 31 30 27
##   3 39 26 70
```

Test error rate is:

```
## [1] 0.5190217
```

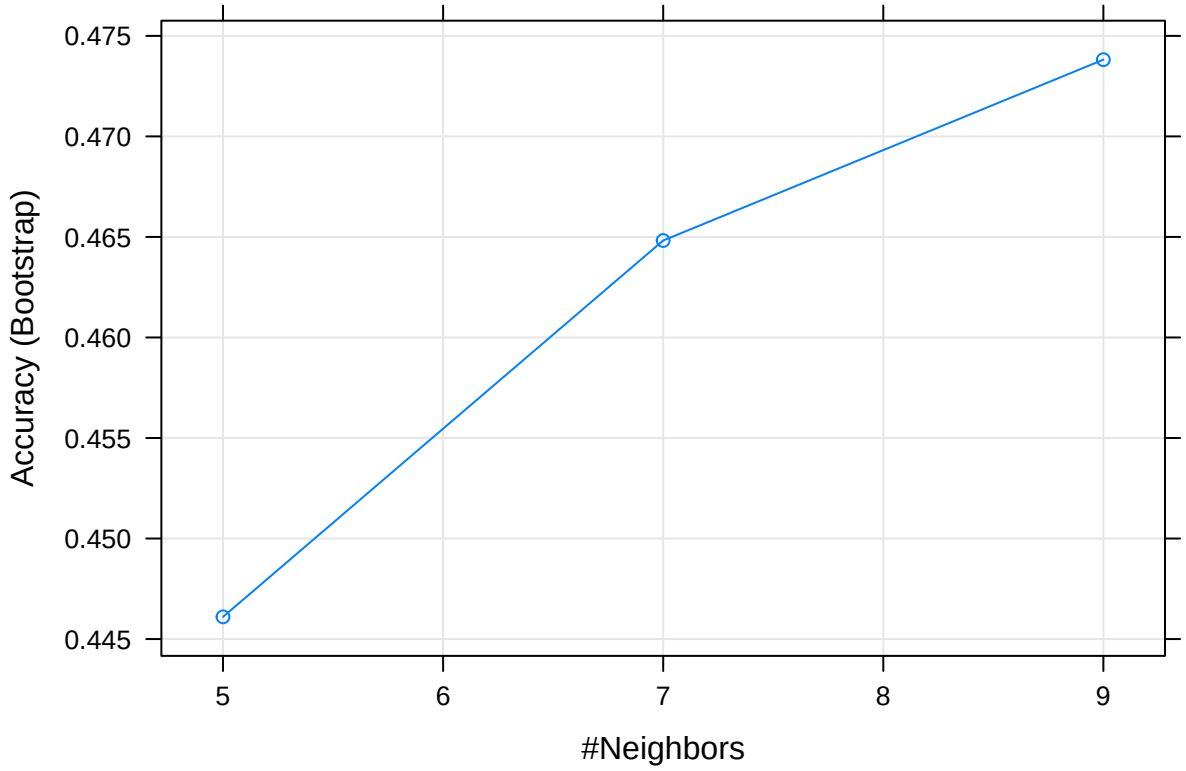
We fit knn model for k in 1 to 100 and we see that k=23 gives the best prediction possible with 46.19% or misclassification rate in this case.

```
##   [1] 0.5163043 0.5923913 0.4891304 0.4891304 0.5000000 0.5000000 0.5054348
##   [8] 0.4891304 0.4945652 0.5054348 0.4918478 0.5135870 0.4836957 0.4782609
##  [15] 0.4918478 0.4755435 0.4728261 0.4701087 0.4809783 0.4755435 0.5000000
##  [22] 0.4836957 0.4619565 0.4755435 0.4619565 0.4782609 0.4728261 0.4701087
##  [29] 0.4864130 0.4782609 0.4809783 0.4945652 0.4728261 0.4809783 0.4755435
##  [36] 0.4809783 0.4836957 0.4809783 0.4755435 0.4809783 0.4864130 0.4782609
##  [43] 0.4864130 0.4809783 0.4755435 0.4809783 0.4728261 0.4918478 0.4891304
##  [50] 0.4918478 0.4945652 0.5027174 0.4836957 0.4918478 0.4891304 0.5000000
##  [57] 0.5190217 0.5108696 0.4945652 0.4864130 0.4972826 0.5054348 0.5081522
##  [64] 0.5108696 0.5027174 0.5081522 0.5054348 0.5108696 0.5135870 0.5108696
##  [71] 0.5027174 0.5000000 0.4972826 0.4864130 0.4782609 0.4864130 0.4782609
##  [78] 0.4891304 0.4836957 0.4972826 0.4945652 0.4918478 0.4891304 0.5000000
##  [85] 0.5027174 0.5000000 0.5000000 0.5027174 0.5217391 0.5244565 0.5027174
##  [92] 0.5108696 0.4972826 0.5108696 0.5027174 0.4891304 0.4864130 0.4945652
##  [99] 0.4918478 0.4918478
```

We also can fit the knn model using the carret package. the particularity of this package is that we do not have to put in k. The model automaticly picks the optimal number of neighbors (k).

```
## k-Nearest Neighbors
##
## 1104 samples
##    9 predictor
##    3 classes: '1', '2', '3'
##
## Pre-processing: centered (9), scaled (9)
## Resampling: Bootstrapped (25 reps)
## Summary of sample sizes: 1104, 1104, 1104, 1104, 1104, 1104, ...
## Resampling results across tuning parameters:
##
##   k  Accuracy   Kappa
##   5  0.4461037  0.142080
##   7  0.4648219  0.169063
##   9  0.4738159  0.181820
##
## Accuracy was used to select the optimal model using the largest value.
## The final value used for the model was k = 9.
```

The summary of this model shows that the optimal k is 9. We also plot accuracy of the model vs. k bellow and we remarque the accuracy peak at k=9.



The model accuraty is 43.22%, in other words 57.88% of miss classification this means the model with class package perform better for this data set than the model in caret package.

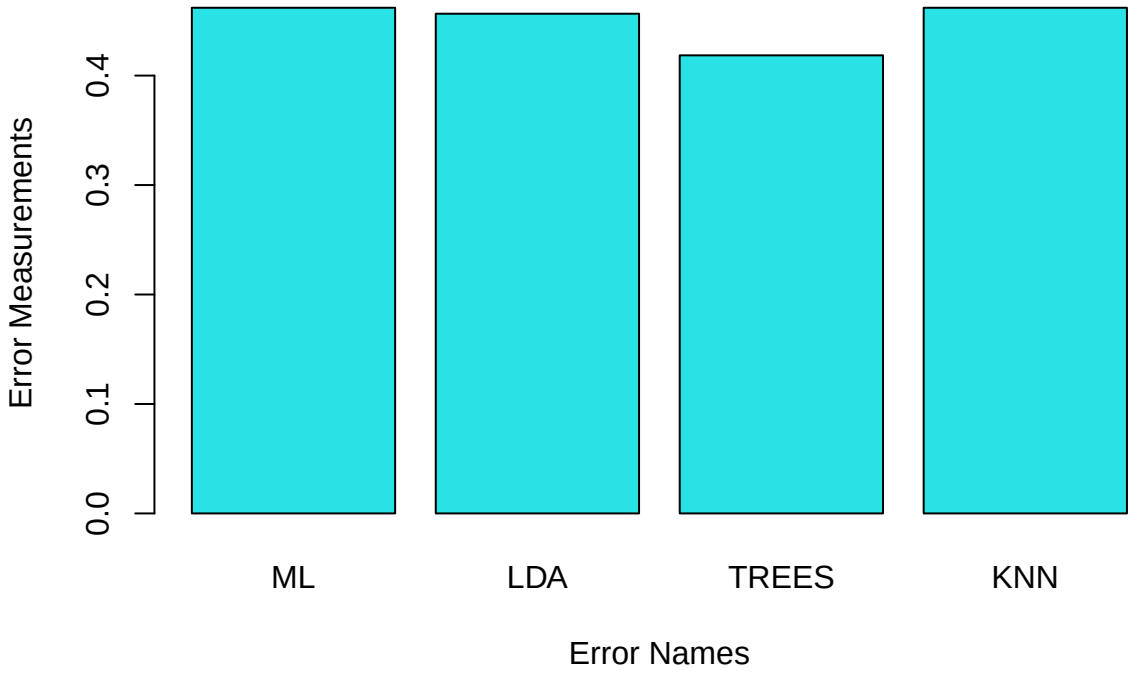
```
## [1] 0.6222826
```

For forwarder analysis with knn model we are going the model in class package and use k=23. This model provive 53.81% of accuracy.

```
## [1] 0.4619565
```

7 Conclusion

The multinomial model is very easy to use. It doesn't requier many variables transformation. It also doen't assume normality of the variables, linearity, or homoscedasticity (Julian J. Faraway 2013). LDA model is very sensitive to outlier alghouth there was no an important outliers in this data. LDA also have a lot of contraints that makes it difficult to use it with this data. For example, it assum that the predictors are gaussian normal therefore, using categorical variables LDA is an violation of the model assumption (Holland, Steven,). However, there are several technics that we can use to transform categorical to numeracal variable such as optimal scaling method (G. Jacoby, William, 2020). Classification tree is a machine learning algorithm it present many problems when it comes to fit big data. It is very important when it comes to check the variables importance but it doesn't do well in predicting. An alternative for better performance is to fit the random forest model which aggregate the output of many tree (Shu, Xiaoling, 2020). Finally knn is very sensitive to the scale of the variables. It is very important to use to sclale all the data points to have a well performing model (Gareth James, Daniela Witten, 2013). The bar blot of the classification error of the four model used to predict the contraceptive model use is shown bellow.



The tree models KNN, LDA and multinormal logistic model have approchimativly the same misclassification rate. Among the the four models classification performs better with an accuracy of 58%.

8 R code

R Packages

```
library("PerformanceAnalytics")
library(ggplot2)
library(pheatmap)
library(faraway)
library(dplyr)
library(tree)
library(rpart)
library(class)
library(MASS)
library(optiscale)
library(caret)
library(e1071)
library(bookdown)
```

Read the data and remane the variables

```
Contraceptive = read.csv("cmc.data")
names(Contraceptive) = c("WAge", "WEdu", "HEdu", "NumChild", "WReligion", "WnWorking", "HOcpt", "SLI",
Contraceptive$WEdu = as.factor(Contraceptive$WEdu)
Contraceptive$HEdu = as.factor(Contraceptive$HEdu)
Contraceptive$WReligion = as.factor(Contraceptive$WReligion)
Contraceptive$WnWorking = as.factor(Contraceptive$WnWorking)
Contraceptive$HOcpt = as.factor(Contraceptive$HOcpt)
```

```
Contraceptive$SLI = as.factor(Contraceptive$SLI)
Contraceptive$MdExpo = as.factor(Contraceptive$MdExpo)
Contraceptive$CMU = as.factor(Contraceptive$CMU)
```

Structure and summary of the data

```
str(Contraceptive)
summary(Contraceptive)
```

Spitting the data into trainning set and test set

```
set.seed(200)
Train_index = sample(dim(Contraceptive)[1], dim(Contraceptive)*0.75)
Train_set = Contraceptive[Train_index, ]
Test_set = Contraceptive[-Train_index, ]
```

Transformation of the categorical variables

```
os.WEdu = opscale(x.qual =Contraceptive$WEdu, level=2, process=1)
os.HEdu = opscale(x.qual =Contraceptive$HEdu, level=2, process=1)
os.WReligion = opscale(x.qual =Contraceptive$WReligion, level=1, process=1)
os.WnWorking = opscale(x.qual =Contraceptive$WnWorking, level=1, process=1)
os.HOcpt = opscale(x.qual =Contraceptive$HOcpt, level=1, process=1)
os.SLI = opscale(x.qual =Contraceptive$SLI, level=1, process=1)
os.MdExpo = opscale(x.qual =Contraceptive$MdExpo, level=1, process=1)
```

```
new.cpt = Contraceptive
new.cpt$WEdu = os.WEdu$os
new.cpt$HEdu = os.HEdu$os
new.cpt$WReligion = os.WReligion$os
new.cpt$WnWorking = os.WnWorking$os
new.cpt$HOcpt = os.HOcpt$os
new.cpt$SLI = os.SLI$os
new.cpt$MdExpo = os.MdExpo$os
```

Splitting the transformed dataset into trainning set and test set

```
set.seed(201)
New_Train_index = sample(dim(new.cpt)[1], dim(new.cpt)*0.75)
New_Train_set = new.cpt[New_Train_index, ]
New_Test_set = new.cpt[-Train_index, ]
```

Graphs

```
egp = group_by(Contraceptive, WEdu, CMU) %>% summarise(count = n()) %>% group_by(WEdu) %>%
  mutate(etotal=sum(count), proportion=count/etotal)
ggplot(egp, aes(x=WEdu, y=proportion, group=CMU, linetype=CMU))+geom_line()
```

```
rgp = group_by(Contraceptive, WReligion, CMU) %>% summarise(count = n()) %>% group_by(WReligion) %>%
  mutate(etotal=sum(count), proportion=count/etotal)
ggplot(rgp, aes(x=WReligion, y=proportion, group=CMU, linetype=CMU))+geom_line()
```

```
plot(jitter(WAge[CMU==1]), jitter(NumChild[CMU==1]), xlab="Woman Age", ylab = "Number of children", col=
points(jitter(WAge[CMU==2]), jitter(NumChild[CMU==2]), col=2, pch=17)
points(jitter(WAge[CMU==3]), jitter(NumChild[CMU==3]), col=4, pch = 18)
legend(x=15, y=12, legend = c("No-use", "Long-term", "Sort-term"), col = c(3,2,4), pch = c(16,17,18), b
```

Multinormal logistic model

```
summary(mmod)
```

```
xtabs(~predict(mmod, Test_set) + Test_set$CMU)
```

Stepwise selection

```
summary(mmodi)
```

```
xtabs(~predict(mmodi, Test_set) + Test_set$CMU)
```

```
1-(102+32+64)/nrow(Test_set)
```

Linear discriminant analysis

```
lda.mod = lda(New_Train_set$CMU ~., data = New_Train_set)
lda.mod
```

```
lda.pred=predict(lda.mod, New_Test_set)
lda.class=lda.pred$class
table(lda.class ,New_Test_set$CMU)
1-(109+22+69)/nrow(New_Test_set)
```

Information calculation for the decision tree

```
xtabs(~CMU+WReligion)
```

```
p1=75/(75+76+69)
p2=76/(75+76+69)
p3=69/(75+76+69)
E_rel0 = -p1*log2(p1)-p2*log2(p2)-p3*log2(p3)
```

```
p1=553/(553+257+442)
p2=257/(553+257+442)
p3=442/(553+257+442)
E_rel1 = -p1*log2(p1)-p2*log2(p2)-p3*log2(p3)
```

```
p1=628/1472
p2=333/1472
p3=511/1472
E_s = -p1*log2(p1)-p2*log2(p2)-p3*log2(p3)
```

```
Information_taken_from_WReligion = ((75+76+69)/1472)*E_rel0 + ((553+257+442)/1472)*E_rel1
```

```
Information_gain_From_WReligion = E_s - Information_taken_from_WReligion
```

Classification tree model

```
set.seed(1111)
tree_mod = rpart(Train_set$CMU ~., Train_set)
printcp(tree_mod)
```

```
plot(tree_mod,branch = 0.4)
text(tree_mod, pretty = 0, cex=0.5)
```

```
tree_mod$variable.importance
```

```
(tt = table(actual=Test_set$CMU, predicted=predict(tree_mod, Test_set, type="class")))
```

```
1-sum(diag(tt))/sum(tt)
```

```
(prune_tree = prune(tree_mod, cp=0.01))
```

```
plot(prune_tree, branch = 0.4)
text(prune_tree, cex=0.6)
```

```
(tt1 = table(actual=Test_set$CMU, predicted=predict(prune_tree, Test_set, type="class")))
```

```
1-sum(diag(tt1))/sum(tt1)
```

Knn model

```
ntrain_set = as.data.frame(lapply(New_Train_set[,1:9], scale))
ntest_set = as.data.frame(lapply(New_Test_set[,1:9], scale))
```

```
set.seed(301)
m1 = knn(train = ntrain_set, test = ntest_set, cl = New_Train_set$CMU, k = 10)
```

```
knn_t = table(New_Test_set$CMU, m1)
knn_t
```

```
1-sum(diag(knn_t))/nrow(Test_set)
```

```
set.seed(301)
tab=numeric(100)
for (i in 1:100) {
  knn_m1 = knn(train = ntrain_set, test = ntest_set, cl = New_Train_set$CMU, k = i)
  knn_tab = table(New_Test_set$CMU, knn_m1)
  error = 1-sum(diag(knn_tab))/nrow(Test_set)
  tab[i]=error
}
tab
```

```
knn_caret = train(ntrain_set, New_Train_set$CMU, method = "knn", preProcess = c("center","scale"))
knn_caret
```

Bar plot

```
error.all = c(mn_error, lda_error, tree_error, knn_error)
names(error.all) = c("ML", "LDA", "TREE", "KNN")
barplot(error.all, col=1 ,xlab="Error Names", ylab="Error Measurements", beside = T)
```

9 References

Robert Gentleman, Kurt Hornik, and Giovanni Parmigiani, *An Introduction to Applied Multivariate Analysis with R*, Springer, 2011.

Gareth James, Daniela Witten, and Trevor Hastie *An Introduction to Statistical Learning*, Springer+Business Media, 2013.

Jay L. Devore, and Kenneth N. Berk, *Modern Mathematical Statistics With Applications*, Springer, Second Edition, 2012.

Julian J. Faraway *Extending the Linear Model with R*: Generalized Linear, Mixed Effect and Nonparametric Regression Models, CRC Press, Second edition, 2016.

Kutner, Nachtsheim, Neter and Li, *Applied Linear Statistical Models*, McGraw-Hill!Irwin, 5th edition, 2015.

Holland, Steven, *Discriminant Function Analysis*, Data Analysis in the Geoscience.

Wiryanto, Heru, *Multinomial Logistic Regression*, 2017

J. Meulman, Jacqueline, *Optimal scaling methods for multivariate categorical data analysis* Data Theory Group, Faculty of Social and Behavioral Sciences Leiden University

G. Jacoby, William, *Optimal Scaling*, 2020

Jan de Leeuw, Patrick Mair, Patrick Groenen, *Multivariate Analysis with Optimal Scaling*, 2017

Raveh, Adi, *A Nonmetric Approach to Linear Discriminant Analysis*, Journal of the American Statistical Association, Vol. 84, No. 405 (Mar., 1989), pp. 176-183, Taylor & Francis, Ltd. on behalf of the American Statistical Association, Accessed: 13-03-2020 03:42

Shu, Xiaoling, *CLASSIFICATION AND DECISION TREES*, Knowledge Discovery in the Social Sciences, University of California Press. (2020)