POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Project report

Dementia Desease: Nonparametric Models for prediction and Survival Analysis

Author: F. Di Filippo, E. Manfrin, E. Musiari, E. Palli

Course: Nonparametric Statistic

Academic year: 2021 - 2022

Contents

Chapter 1

INTRODUCTION

The aim of our project is to use nonparametric methods to study the Dementia desease, thanks to the Open Access Series of Imaging Studies (OASIS) made available by the Washington University Alzheimer's Disease Research Center, Dr. Randy Buckner at the Howard Hughes Medical Institute (HHMI) at Harvard University, the Neuroinformatics Research Group (NRG) at Washington University School of Medicine, and the Biomedical Informatics Research Network (BIRN).
Dementia describes a group of symptoms associated with a decline in memory, reasoning or other thinking skills. Many different types of dementia exist, and many conditions cause it.
Dementia is not a normal part of aging. It is caused by damage to brain cells that affects their ability to communicate, which can affect thinking, behavior and feelings.
To study this desease we have done a first explorative analysis of the dataset, we have used depht measures to find the outliers, we have performed a survival analysis on the converted patients. As last step we have performed regression and prediction with different methods (nonparametric logistic regression, robust logistic regression , conformal prediction) to identify patients which are demented from the one that are not demented.

Chapter 2

DATASET PRESENTATION AND FIRST ANALYSIS

2.1. Dataset

We have two datasets available:

- Cross-sectional MRI Data in Young, Middle Aged, Nondemented and Demented Older Adults: This set consists of a cross-sectional collection of data of 416 subjects aged from 18 to 96. The subjects are all right-handed and the collection includes both men and women. 100 of the subjects over the age of 60 have been clinically diagnosed with very mild to moderate Dementia's disease (AD). Moreover, a reliability data set is included containing 20 nondemented subjects visited again within 90 days of their initial session.
- Longitudinal MRI Data in Nondemented and Demented Older Adults: This set consists of a longitudinal collection of 150 subjects aged from 60 to 96. Each subject was scanned on two or more visits, separated by at least one year for a total of 373 imaging sessions. The subjects are all right-handed and the collection includes both men and women. 72 of the subjects were characterized as nondemented throughout the study. 64 of the included subjects were characterized as demented at the time of their initial visits and remained so for subsequent scans, including 51 individuals with mild to moderate Dementia's disease. Another 14 subjects were characterized as nondemented at the time of their initial visit and were subsequently characterized as demented at a later visit.

2.2. Variables

In our datasets we have the variables:

- ID, which is an identification number for each patient, we have use it to select the first visit of each patient to create a new dataset with independent components.
- M.F, which is the gender.
- Hand, which is the dominant hand, we have not use this parameter since all the patients are right-handed.
- Age, which is the age of the patient at each visit.
- EDUC, which is the educational level, with values ranging from 1 to 5.
- SES, which is the socio-economic status, with values ranging from 1 to 5.
- MMSE, which is the Mini Mental State Examination, a test with values ranging from 1 to 30 to establish how is the mental state of the patient (patient with high value have less probability of being demented).
- CDR, which is the Clinical Dementia Rating,a value ranging between 0 and 3 which explains the grade of dementia of the patient.
- eTIV, which is the Estimated Total Intracranial Volume.
- nWBV, which is Normalize Whole Brain Volume.
- ASF, Atlas scaling factor, which is excluded from the analysis since it's based on eTIV.
- Delay, day from the first visit.
- Label: in Longitudinal MRI Data the classification of the patients in Demented, NonDemented and Converted is already present in the Variable 'Group'. While in Cross-sectional MRI Data we used the variable CDR: NonDemented are those whit $CDR = 0$, the other are classified as Demented.

2.3. Training and Test set

To compute the division in Training and Test set we merge together the Cross-sectional MRI Dataset and the first visit of Longitudinal MRI Data, to obtain independent samples. Then we used 271 data as Training set and the other as Test set.

2.4. ANOVA and Permutation tests

To motivate our study we started performing a multivariate permutation test to assess if there is statistical evidence of the difference between the patients diagnosed as *Demented* and *Non demented*. The test:

$$H_0 : X_{Demented} \overset{\text{d}}{=} X_{Nondemented} \quad vs \quad H_1 : X_{Demented} \overset{\text{d}}{\neq} X_{Nondemented}$$

means that we test the equality of the distribution of the two groups of patients. We decided to use the scaled dataset, since the order of magnitude of our variable is very different, and to apply as test statistic the infinite norm of the difference between the Tukey median of the two groups. The p-value resulting from our test is

$$p_{groups} = 0.002$$

meaning that we reject the null hypothesis of the two groups being statistically equivalent.

Moreover, we are particularly interested in the importance of variables M.F and Age (the sex and age of the patients) in the prediction of the state of the patient (namely, if demented or not). We can start analyzing these two variables with respect to the grouping: sex can be thought as categorical variable used to explain the dependent variable Label. We will use the ANOVA (Analysis of Variance) test, considering sex as a "treatment", to study if there are differences between the groups means.
We can formalize this:

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij} \qquad \forall i = 1, \dots, I, \ \forall j = 1, \dots, n_i$$

where I is the number of factor levels, Y_{ij} is the j^{th} observation for the level i of the factor, n_i the number of replicates for the level i if the factor, n the total number of patients, μ the overall mean, α_i is the i^{th} treatment effect, $\epsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$ independents.
We study the existence of an effect of the factor on the output variable.

$$H_0 : \ \alpha_1 = \ldots = \alpha_I = 0 \quad vs \quad H_1 : \ \exists i \ st \ \alpha_i \neq 0$$

First we want to test the variable M.F.
Since we want only to use independent observations, from the longitudinal dataset we select only the first visit of each patient, and we add them to the cross-sectional dataset in order to start the analysis with all the independent patients at their first visit.
First of all, we check the hypothesis of normality of the residuals using the Shapiro-Wilk test, where the null hypothesis is that the residuals are gaussian. We choose a significance level of 0.05 and we get as p-value of the test:

$$p_{MF} < 2.2 \times 10^{-16}$$

Clearly, we have to reject the null hypothesis, meaning that the residual are not gaussian. For this reason, we should perform the analysis of variance in a fully permutational setting. As before, the null hypothesis is that the factor is not significant. The new test gives us the p-values:

$$p_{MF} = 0.011$$

Hence, for $\alpha = 0.05$ we reject the null hypothesis, namely sex is a significant factor in the prediction of the state of the patients. These results will be useful in our study.

After having tested the importance of the gender, we now want to perform a permutation test to check the importance of the variable Age with respect to the grouping *Demented, Nondemented*: the null hypothesis states that the variable is not statistically significant with respect to the grouping; the alternative hypothesis is the complementary of the null one:

$$H_0 : \ Age_{Demented} \overset{\text{d}}{=} Age_{Nondemented} \quad vs \quad H_1 : \ Age_{Demented} \overset{\text{d}}{\neq} Age_{Nondemented}$$

The p-value resulting from our study is:

$$p_{Age} = 0$$

this means that also the Age is important in the grouping of our patients.

Chapter 3

OUTLIER DETECTION

3.1. Depth measures

Outlier detection is the analysis aimed at finding anomalies in our dataset. There exists many methods to perform this analysis and we will start using the depth measures.

3.1.1. Longitudinal dataset

Since we are in a p-dimensional case with $p >> 2$, the best option is to represent the depths through a bagplot matrix. We performed it on the *longitudinal dataset*, obtaining:

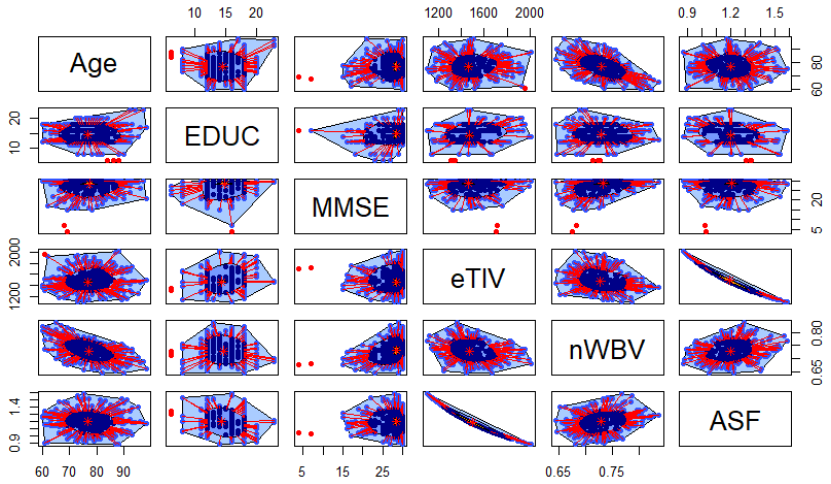


Figure 3.1: Bagplot matrix of longitudinal dataset

Using this bagplot we want to try to understand which comparisons are reasonable to find outliers using the depth measures. However, the outlier detection becomes more difficult when the dimension increases, since some outliers could be wrongly flagged as genuine points, or some good points could be wrongly flagged as outliers (respectively the masking and swamping phenomena).
We will focus our analysis on the points that from 3.1 we can suppose as anomalous, visualizing them two at a time using the bagplot function (hence the Tukey measure).

Age VS MMSE

We can find two points are outside the shell of the comparison between Age and MMSE, ie considered as *outliers*. Through the command `pxy.outlier` we can identify them, finding:

x	y
68	7
69	4

which are at lines 101, 102.

eTIV VS MMSE

We can find that one point is marked as outlier, more precisely:

x	y
1701	4

which is at line 102.

nWBV VS MMSE

One point is considered as outlier; we can identify it, finding:

x	y
0.676	4

which is at line 102.

ASF VS MMSE

One point is marked as outlier; we can identify it, finding:

x	y
1.032	4

which is at line 102.

Age VS EDUC

There are three points considered as outliers; we can identify them, finding:

x	y
84	6
86	6
88	6

which are at lines 79, 80, 81.
This means that these three lines have an educational level really lower than the others.

eTIV VS EDUC

There are three points marked as outliers; we can identify them, finding:

x	y
1310	6
1320	6
1348	6

which are at lines 79, 80, 81.

nWBV VS EDUC

There are three points marked as outliers; we can identify them, finding:

x	y
0.727	6
0.724	6
0.713	6

which are at lines 79, 80, 81.

ASF VS EDUC

There are three points considered as outliers; we can identify them, finding:

x	y
1.339	6
1.329	6
1.302	6

which are at lines 79, 80, 81.

Age VS eTIV

One point is marked as outlier; we can identify it, finding:

x	y
61	1957

which is at line 140.

Conclusions

The patient OAS2_0048 at line 101 results as anomalous in the fourth visit for MMSE vs Age.
The patient OAS2_0048 at line 102 results as anomalous in the fifth visit for both MMSE vs Age, and MMSE vs the last 3 columns (eTIV, nWBV, ASF). This is not surprising since these three columns are resulted to be dependent in our analysis. This patient is quite young in these two visit (68 and 69 years old).
The patient OAS2_0040 at lines 79, 80, 81 anomalous in all his 3 visits for EDUC vs Age and EDUC vs the last 3 columns (eTIV, nWBV, ASF). This is caused by the very low level of education compared to the other patients analysed.
The patient OAS2_0066 at line 140 results as anomalous in the first visit only for Age vs eTIV. However, this patient results not to be anomalous with respect to nWBV and ASF. This could be a result of an erraneous transcription. The patient in his first visit is quite young (61 years old).

3.1.2. Training dataset

We proceed with the analysis considering the training set that we have explained before. In this way, we can find out if it does exist any patient that is an outlier with respect to the others in our Dementia research.

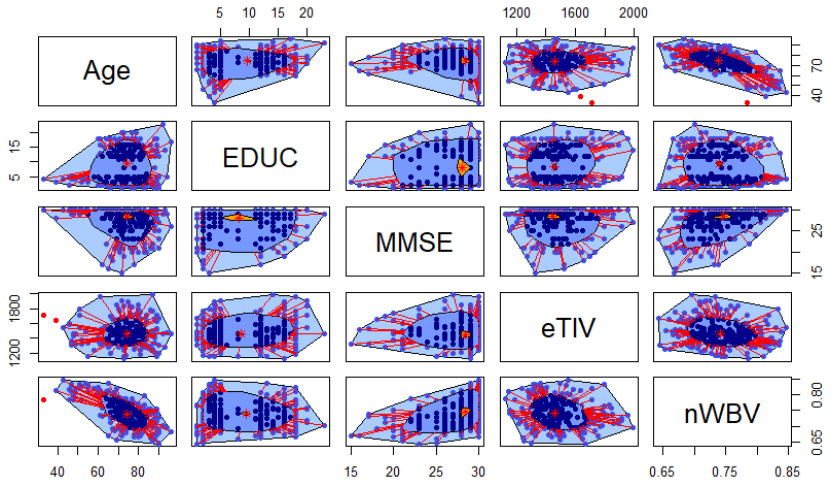
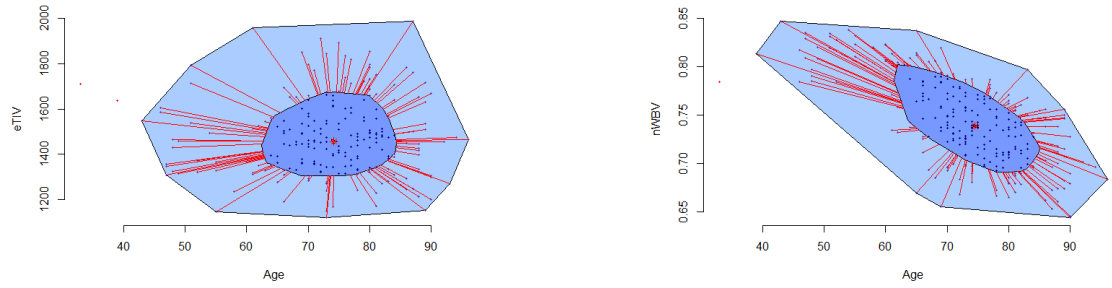


Figure 3.2: Bagplot matrix of training dataset

Now our aim is to identify if there are anomalous patients, focusing our analysis on the points that from 3.2 we can suppose as anomalous. In this case, it seems to have problems only with Age VS eTIV and Age VS nWBV.

Age VS eTIV and Age VS nWBV



(a) Bagplot Age VS eTIV

(b) Bagplot Age VS nWBV

Figure 3.3

From 3.3a we can outline that two points are outside the shell, i.e. considered as *outliers*. Through the command `pxy.outlier` we can identify them, finding:

x	y
39	1636
33	1709

which are at lines 145, 175.
From 3.3b we can see that only one point is marked as outlier; we can identify it, finding:

x	y
33	0.784

which is at line 175.

Conclusions

In the training dataset, we can outline that: at line 175 there is a patient anomalous both for Age vs eTIV and Age vs nWBV; at line 145 there is a patient anomalous only for Age vs eTIV. This is due to the fact that they have a particular age: they are respectively 33 and 39 years old.

In general, when we deal with medical datasets, we expect to find outliers due to the extreme medical conditions of certain patients. For this reason, we decided not to cancel any of the patients marked as outliers.

Chapter 4

LOGISTIC REGRESSION

As next step of our analysis we tried to implement different types of logistic regression models in a nonparametric framework, in order to make an effort in developing a model that can help predicting if a new patient is demented or not.

4.0.1. Covariates

Since the covariate MMSE is a reliable indicator in the detection of demented patients, and neurologists exploite mainly this indicator to diagnose the Alzheimer desease, we decided to consider two different approaches: the first one is to fit a model with MMSE as unique covariate, trying to exploit all the information from it. The second one is to fit a model that doesn't take into account MMSE and understand if it could be an additional instrument for neurologists. In this case we fitted a Generalized additive model using as covariates: sex, Age, Education, nWBV and eTIV.

4.1. MMSE model

In this part we introduce 3 different nonparametric logistic models using MMSE as covariate and we discuss their goodness of fit and their performances in prediction on the test set.

4.1.1. Global polynomial regression

The first model we built up is a polynomial regression model. We fitted it using as regressors an orthogonal basis of polynomials up to degree 7 in order to reduce it. We started by comparing the deviance between each model from degree 1 up to degree 7 in order to find the optimal degree for the polynomial regression. From degree 4 to degree 5 the gain in information is low and from 5 up to 7 we have no more gain in performance (4.1).Indeed we perfomed an Anova χ^2-test between degree 4 and degree 5 polynomial model to compare the two models (4.2):

$$\log \frac{p}{1-p} = \beta_0 + \beta_1 \cdot MMSE + \beta_2 \cdot MMSE^2 + \beta_3 \cdot MMSE^3 + \beta_4 \cdot MMSE^4 + \beta_5 \cdot MMSE^5$$

H0: $\beta_5 = 0$ vs H1: $\beta_5 \neq 0$

```
Analysis of Deviance Table

Model 1: label ~ poly(MMSE, degree = degree)
Model 2: label ~ poly(MMSE, degree = degree)
Model 3: label ~ poly(MMSE, degree = degree)
Model 4: label ~ poly(MMSE, degree = degree)
Model 5: label ~ poly(MMSE, degree = degree)
Model 6: label ~ poly(MMSE, degree = degree)
Model 7: label ~ poly(MMSE, degree = degree)
  Resid. Df Resid. Dev Df Deviance
1       269     212.03
2       268     211.50  1   0.5252
3       267     210.97  1   0.5337
4       266     205.77  1   5.1976
5       265     203.77  1   2.0018
6       264     203.77  1   0.0000
7       263     203.77  1   0.0000
```

Figure 4.1

```
> anova(mdl_poly[[3]], mdl_poly[[4]], test = 'Chisq')
Analysis of Deviance Table

Model 1: label ~ poly(MMSE, degree = degree)
Model 2: label ~ poly(MMSE, degree = degree)
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       267     210.97
2       266     205.77  1   5.1976  0.02262 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> anova(mdl_poly[[4]], mdl_poly[[5]], test = 'Chisq')
Analysis of Deviance Table

Model 1: label ~ poly(MMSE, degree = degree)
Model 2: label ~ poly(MMSE, degree = degree)
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       266     205.77
2       265     203.77  1   2.0018   0.1571
> anova(mdl_poly[[4]], mdl_poly[[6]], test = 'Chisq')
Analysis of Deviance Table

Model 1: label ~ poly(MMSE, degree = degree)
Model 2: label ~ poly(MMSE, degree = degree)
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       266     205.77
2       264     203.77  2   2.0018   0.3676
```

Figure 4.2

We can conclude that in terms of goodness of fit the degree 4 seem to be better, according also to the Akaike criterion.

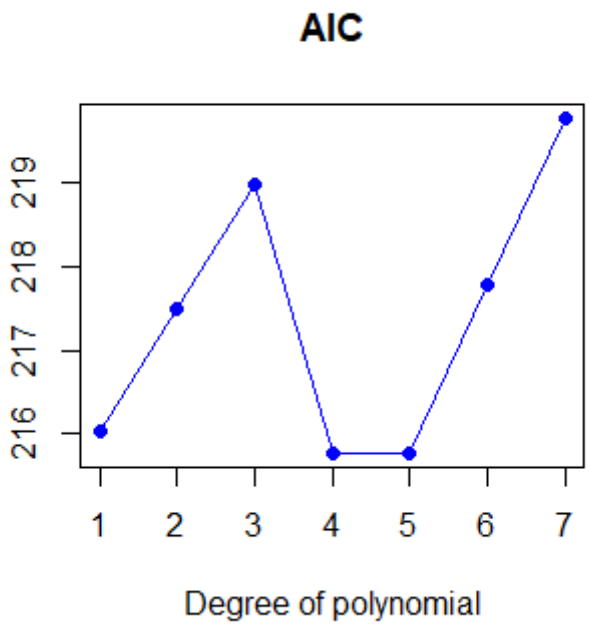


Figure 4.3: the minimum is reached by the 4-th degree polynomial model

The model also provides a quite good McFadden $R^2 = 0.446$. Moreover we compared the performances in prediction on the test set, and the degree 4 polynomial model reaches the maximum value in term of area under roc curve.

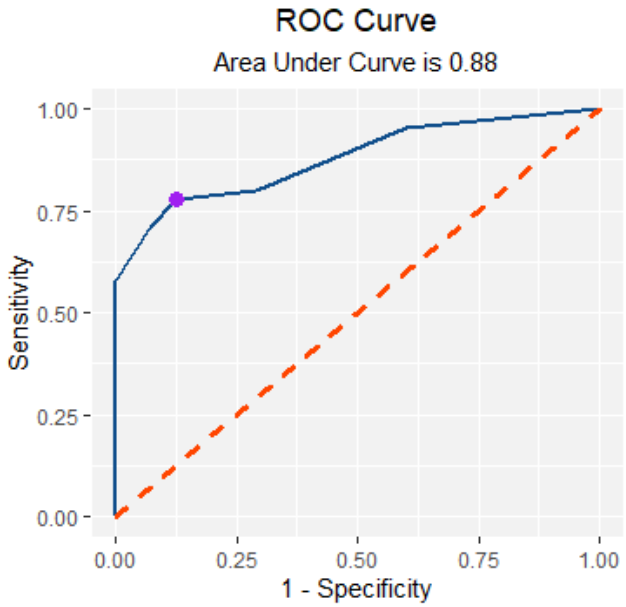


Figure 4.4: ROC curve for polynomial regression model

Prediction scores:

- AUC = 0.88
- Sensitivity = 0.78
- Specificity = 0.87
- Accuracy = 0.83
- Precision = 0.83
- F1 score = 0.80

As we can observe on the right tail the logistic curve (4.5) has a flaw in his shape, starting increasing from a certain value of MMSE: this doesn't reflect the idea that for higher values of MMSE we get lower probability to be demented.

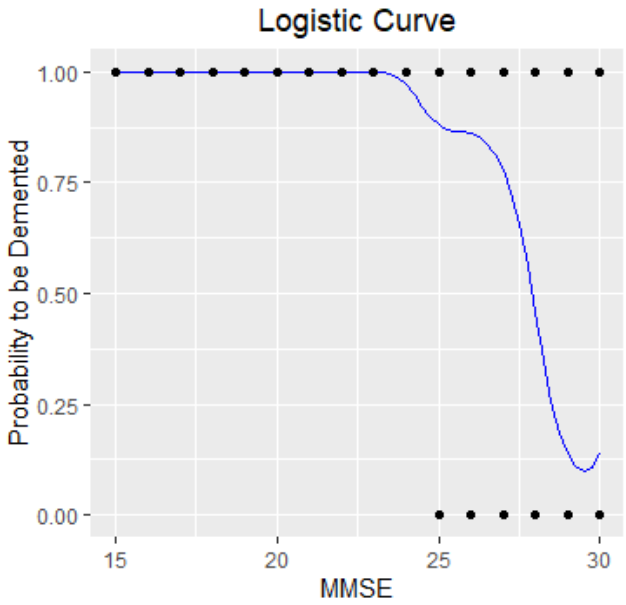


Figure 4.5: ROC curve for polynomial regression model

4.1.2. Local Likelihood

As next step we considered a Local Likelihood logistic model with gaussian kernel. The aim is to maximize the local likelihood function to obtain the coefficients betas:

$$\ell_{x,h}(\boldsymbol{\beta}) := \sum_{i=1}^{n} \ell(Y_i, \eta(MMSE_i - x)) K_h(x - MMSE_i). \tag{5.14}$$

$$\ell(Y_i, \eta(MMSE_i - x)) = Y_i \log(\text{logistic}(\eta(MMSE_i - x))) + (1 - Y_i) \log(1 - \text{logistic}(\eta(X_i - x))). \tag{5.15}$$

$$\eta(x) := \beta_0 + \beta_1 x. \tag{5.16}$$

Where $K_h(x)$ is the Gaussian Kernel.
h is the smoothing parameter, that states the fraction of data around x that contribute to smooth the curve. We selected the default value h=0 (no smoothing). As we can see from the plot below we managed to correct behaviour of the fit of the polynomial regression curve for high values of MMSE, obtaing a monotonically decreasing curve.

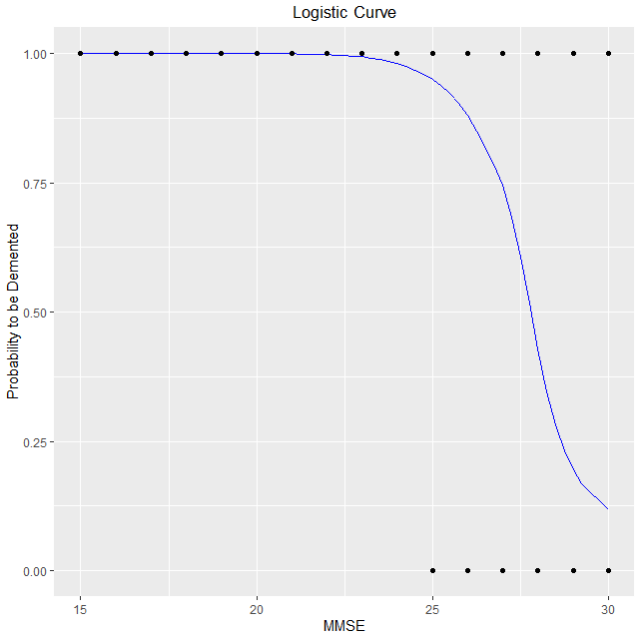


Figure 4.6

Also this method provides good results in terms of prediction:

- AUC = 0.88
- Sensitivity = 0.78
- Specificity = 0.87
- Accuracy = 0.83
- Precision = 0.83
- F1 score = 0.80

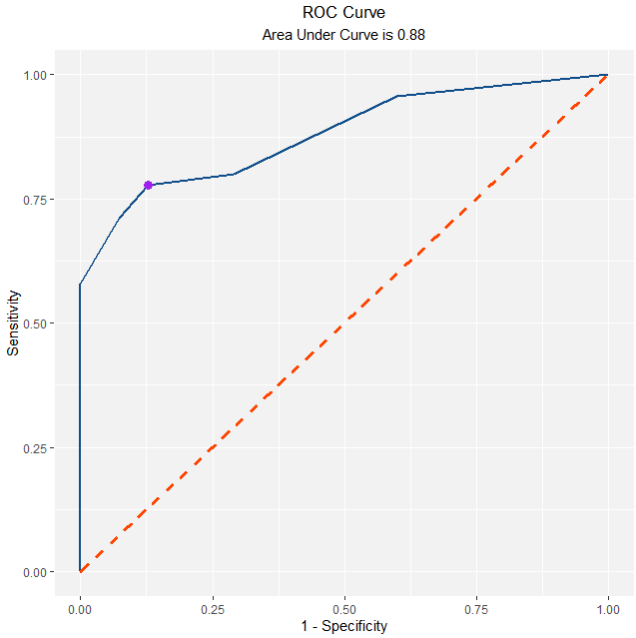


Figure 4.7

4.1.3. Cubic Spline Model

Then we fitted a linear spline logistic model with 1 knot placed at the median. We tried to increase the degree of the spline in order to get a smoother regression curve, but it results in not significant regressors. Since the distribution of MMSE is characterized by majority of high values, within that region we would like to increase the number of knots. However its support is descrete, thus we obtained an overfitted curve. At the end we opted for the initial model:

$$\log \frac{p}{1-p} = \beta_0 + \sum_{j=1}^{K+1} \beta_j g_j \left(MMSE\right)$$

$K = 1$ knot, placed at the median of MMSE.

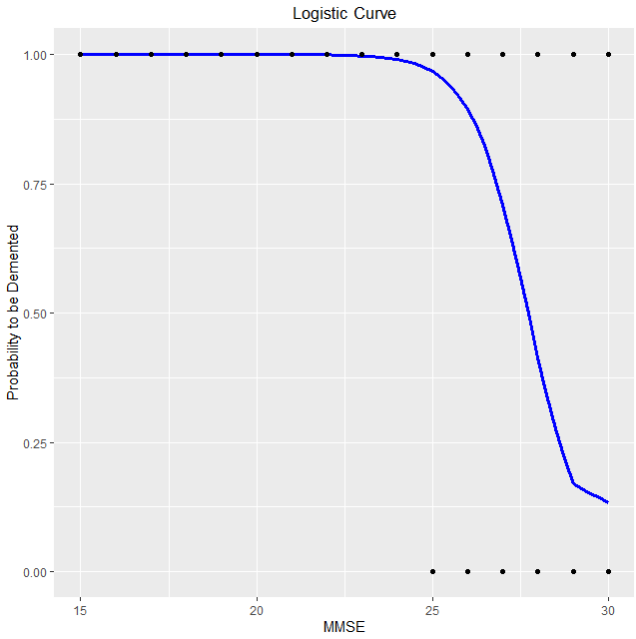


Figure 4.8: logistic curve for spline regression model

$R^2 = 0.4377$, AIC $= 216.03$
In terms of prediction we have no gain and no loss with respect to the other models (4.9):

- AUC = 0.88
- Sensitivity = 0.78
- Specificity = 0.87
- Accuracy = 0.83
- Precision = 0.83
- F1 score = 0.80

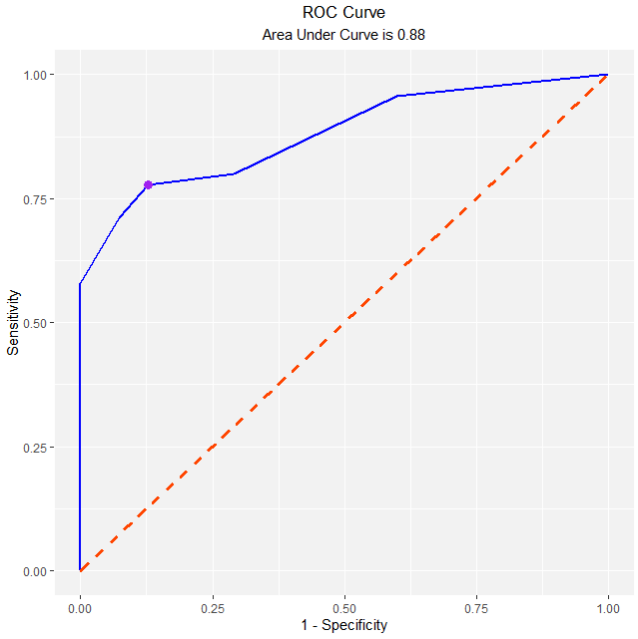


Figure 4.9: ROC curve for spline regression model

4.1.4. Comparison with Parametric Logistic Model

As final analysis we wanted to compare these models with a parametric logistic regression one, in order to study if a nonparametric approach can effectively provide a benefit in terms of goodness of fit and prediction. As we can see from the table 4.1 below, the values of R^2 indicate that nonparametric models tend to fit slightly better and also Akaike information is in favour of global polynomial model. On the other hand the two different approaches seem to be similar under the predictive perspective.

	R^2	AIC	AUC
PARAMETRIC GLM	0.429	216.0	**0.879**
GLOBAL POLYNOMIALS	**0.446**	215.8	**0.879**
LOCAL LIKELIHOOD	-	-	**0.879**
CUBIC SPLINE	0.4377	**213.03**	0.854

Table 4.1: for the local likelihood we used the funcion locfit, which does not provide values of deviance, null deviance and AIC, so we can't extract the fitting performances.

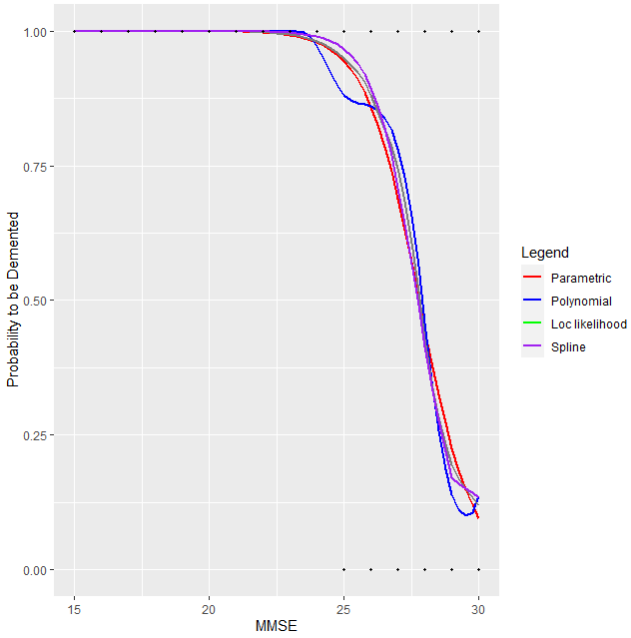


Figure 4.10: The plot show a similar behaviour in fitting for the three models

4.2. Generalized Additive Model

In this section we wanted to develop a model to be used by neurologists in addition to the MMSE to reach a more accurate result in terms of prediction of the pathology. We started implementing a Generalized Additive model considering as numerical covariates Age, Education, nWBV, eTIV, the dummy variable for sex and its interaction with each of the numerical variables.

$$\ln \frac{p}{1-p} = \beta_0 + \beta_1 Sex + f_2\left(EDUC : Sex\right) + f_3\left(nWBV : Sex\right) + f_4\left(Age : Sex\right) + f_5\left(eTIV : Sex\right)$$

As we can deduce from the plot 4.11 and from the high p-values, the model can be reduced removing the interaction of sex with age and nWBV.

```
Family: binomial
Link function: logit

Formula:
label ~ M + s(EDUC, bs = "cr") + s(I(EDUC * M), bs = "cr") +
    s(nWBV, bs = "cr") + s(I(nWBV * M), bs = "cr") +
    s(Age, bs = "cr") + s(I(Age * M), bs = "cr") +
    s(eTIV, bs = "cr") + s(I(eTIV * M), bs = "cr")

Parametric coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   -1.741      1.674  -1.040    0.298
M              1.366      2.274   0.601    0.548

Approximate significance of smooth terms:
                edf Ref.df Chi.sq  p-value
s(EDUC)       4.020e+00      9 23.801 2.10e-05 ***
s(I(EDUC * M)) 7.163e+00      9 13.157  0.03848 *
s(nWBV)       1.000e+00      9 20.595 2.94e-06 ***
s(I(nWBV * M)) 1.315e+00      9  1.224  0.34620
s(Age)        6.570e+00      9 19.192  0.00371 **
s(I(Age * M))  9.041e-05      9  0.000  1.00000
s(eTIV)       5.291e+00      9  9.051  0.07269 .
s(I(eTIV * M)) 4.344e+00      9  7.991  0.04780 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

R-sq.(adj) =  0.389   Deviance explained = 40.1%
UBRE = 0.055142  Scale est. = 1         n = 271
```

Figure 4.11: Summary of the complete model

In this way also the intercept of the dummy become significant (p-value = 0.006), and performing an Anova χ^2 test: H0:$\beta_{sex*age} = \beta_{sex*educ} = 0$ vs H1 = $H0^c$ P-value = 0.09
We accept H0 at level 5%, the two models are statistically equal. The reduced model is:

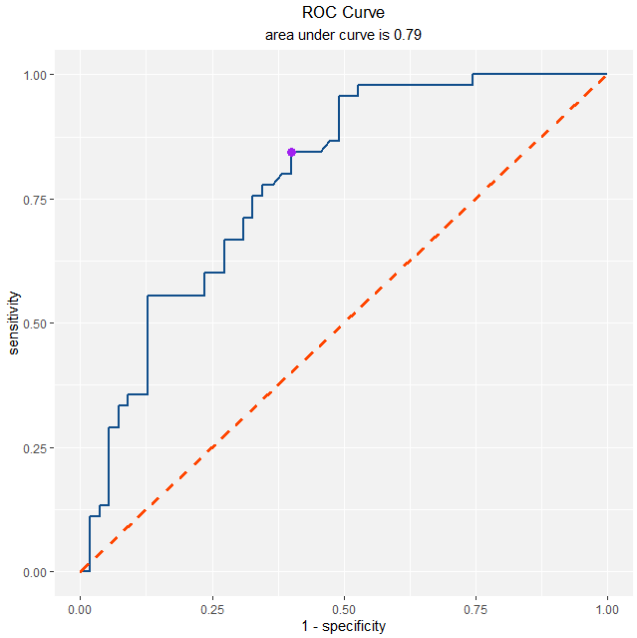


Figure 4.12: ROC curve for the GAM model

$$\ln \frac{p}{1-p} = \beta_0 + \beta_1 Sex + f_2\left(EDUC\right) + f_3\left(nWBV\right) + f_4\left(nWBV * Sex\right) + f_5\left(Age\right) + f_6\left(eTIV\right) + f_7\left(eTIV * Sex\right)$$

Moreover in terms of fitting the two model perform similarly:
Model 1: $R^2 = 0.4$ $R^2_{adj} = 0.389$ AIC $= 285.9$
Model 2: $R^2 = 0.36$ $R^2_{adj} = 0.36$ AIC $= 284.5$
The Akaike information criterion suggest to reduce the model in order to reduce its complexity.
From a prediction point of view the AUC for the two models is almost equal (0.789 and 0.785 respectively for the first and the resuced one), so we decided to reduce it. We labelled as demented whoever has an estimated probability of being an Alzheimer's patient greater that 33% (deduced from the ROC curve), taking into account that is more important to maximize sensitivity rather than specificity.
It shows a satisfying performance on the test set:

- Sensitivity = 0.84
- Specificity = 0.6
- Accuracy = 0.71
- Precision = 0.63
- F1 score = 0.72

Chapter 5

ROBUST LOGISTIC REGRESSION

5.1. Introduction to the analysis

After having performed the nonparametric logistic regression we decide to implement a robust logistic regression in particular to see if the preceding models are effected by outliers and/or by the different distributions of demented and non-demented.
We have implemented different models, with and without B-splines. We have implemented a model with all the covariates and then as in the chapter before we have considered first a model with only the variable that we have found to be the most informative (MMSE), and a model with all the other parameters except this one. To do so we have used the command *glmrob* which fit generalized linear models by robust methods, since our regression is logistic we use the binomial family. For all this model we have considered the training dataset to train the model and the test set to test the model.

5.2. Some theory about *glmrob*

This part refers to **[2]** and **[3]**.
We consider the following set of estimating equations for the robust estimation of β:

$$\sum_{i=1}^{n} \psi(y_i, \mu_i) = 0 \tag{5.1}$$

where $\psi(y, \mu) = \nu(y, \mu) \cdot w(x) \cdot \mu' - a(\beta)$, $a(\beta) = \frac{1}{n} \sum_{i=1}^{n} \mathtt{E}[\nu(y_i, \mu_i)] \cdot (x_i) \cdot \mu_i'$, with the expectation taken with respect to the conditional distribution of $y|x$, $\nu(\cdot, \cdot)$, $w(x)$ are robustness weight functions, and $\mu_i = \mu_i(\beta) = g^{-1}(x_i^T \beta)$.The constant $a(\beta)$ ensures the Fisher consistency of the estimator.
The estimator defined in (5.1) is an M-estimator characterized by the score function $\psi(y_i, \mu_i) = \nu(y_i, \mu_i) \cdot w(x_i) \cdot \mu_i' - a(\beta)$. It is asymptotically normally distributed with variance $\Omega = M(\psi, F)^{-1} Q(\psi, F) M(\psi, F)^{-1}$, where $M(\psi, F) = -\mathtt{E}[\frac{\partial}{\partial \beta} \psi(y, \mu)]$ and $Q(\psi, F) = E[\psi(y, \mu)\psi(y, \mu)^T]$. Its influence function is proportional to ψ, which is bounded with respect to y if $\nu(y, \mu)$ is bounded, and with respect to x if $w(x)$ is suitably chosen to down-weight leverage points.
Moreover, the estimating equations can be obtained by differentiating the robust quasi-likelihood function
$Q_M(y, \mu) = \sum_{i=1}^{n} Q_M(y_i, \mu_i)$,
with respect to β, where the functions $Q_M(y_i, \mu_i)$ can be written as
$Q_M(y_i, \mu_i) = \int_{\tilde{s}}^{\mu_i} \nu(y_i, t) w(x_i) dt - \frac{1}{n} \sum_{j=1}^{n} \int_{\tilde{t}}^{\mu_j} E[\nu(y_j, t) w(x_j)] dt$, with $\tilde{s}$ such that $\nu(y_i, \tilde{s}) = 0$, and $\tilde{t}$ such that $E[\nu(y_i, \tilde{t})] = 0$.
For binomial models we have the particular case
$\nu(y_i, \mu_i) = \psi_c(r_i) \frac{1}{V^{1/2}(\mu_i)}$
with $r_i = \frac{y_i - \mu_i}{V^{1/2}(\mu_i)}$ being the Pearson residuals and ψ_c the Huber function defined by

$$\psi_c(r) = \begin{cases} r & |r| \leq c \\ c \cdot sign(r) & |r| > c \end{cases} \tag{5.2}$$

5.3. Complete model

As said before, the first model has as covariates all the variables except CDR. This parameter tells us if a patient is demented and so ,with this parameters, we have already a classification in demented and non demented.
The model from which we have started is

$$log(\frac{p}{1-p}) = \beta_1 \cdot M.F + \beta_2 \cdot EDUC + \beta_3 \cdot nWBV + \beta_4 \cdot Age + \beta_5 \cdot MMSE + \beta_6 \cdot eTIV$$

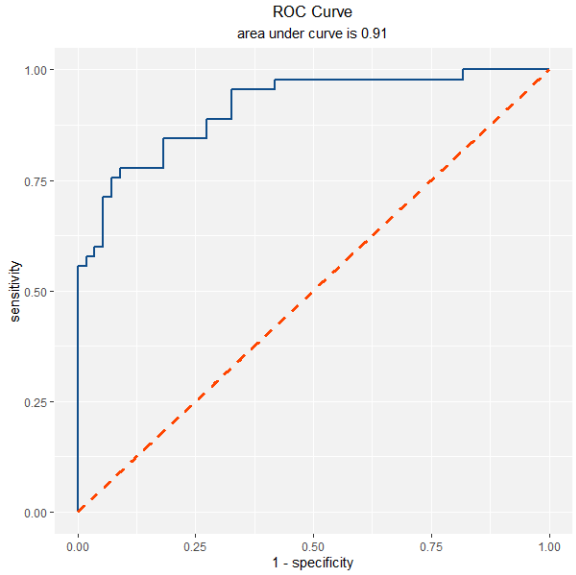


Figure 5.1: ROC curve complete model

These are the ROC curve and the values of the estimator:

$$F1 - score = 0.7916667 \qquad accuracy = 0.8 \qquad precision = 0.745098$$
$$sensitivity = 0.875 \qquad specificity = 0.7333$$

Then we consider the same model with B-splines

$$log(\frac{p}{1-p}) = \beta_1 \cdot M.F + \beta_2 \cdot bs(EDUC, degree = 3) + \beta_3 \cdot bs(nWBV, degree = 2) + \beta_4 \cdot bs(Age, degree = 2)$$
$$+\beta_5 \cdot bs(MMSE, degree = 2) + \beta_6 \cdot bs(eTIV, degree = 3)$$

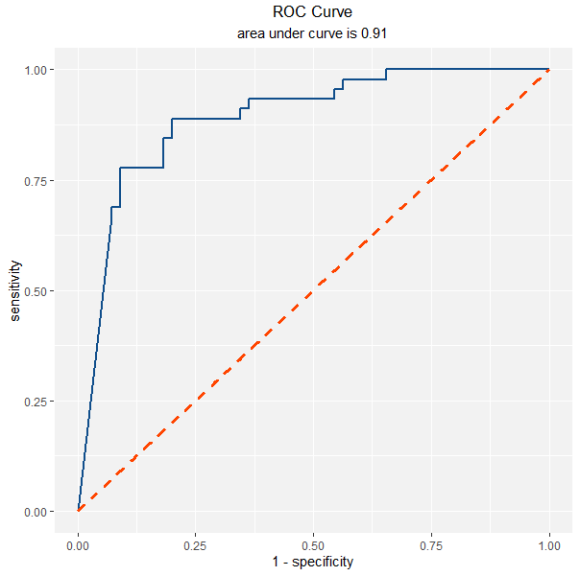


Figure 5.2: ROC curve complete model with splines

With values

$$F1 - score = 0.8080808 \qquad accuracy = 0.81 \qquad precision = 0.7407407$$
$$sensitivity = 0.8888889 \qquad specificity = 0.7454545$$

We have also reduced the model, but with less covariates, the performances worsten a little. The best reduced model are: Without B-splines:

$$log(\frac{p}{1-p}) = \beta_1 \cdot M.F + \beta_2 \cdot MMSE$$

With ROC curve

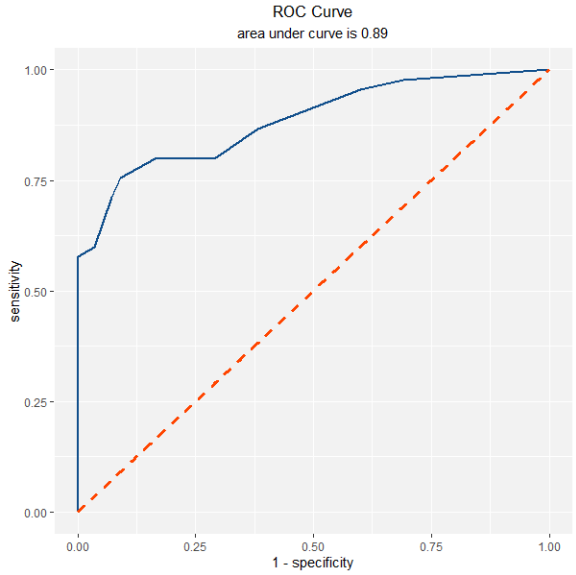


Figure 5.3: ROC curve M.F and MMSE model without B-splines

and values

$$F1 - score = 0.742268 \qquad accuracy = 0.75 \qquad precision = 0.6923077$$

$$sensitivity = 0.8 \qquad specificity = 0.7090909$$

And with B-splines: $log(\frac{p}{1-p}) = \beta_1 \cdot M.F + \beta_2 \cdot bs(MMSE, degree = 2) + \beta_3 \cdot bs(eTIV, degree = 3)$
With ROC curve:

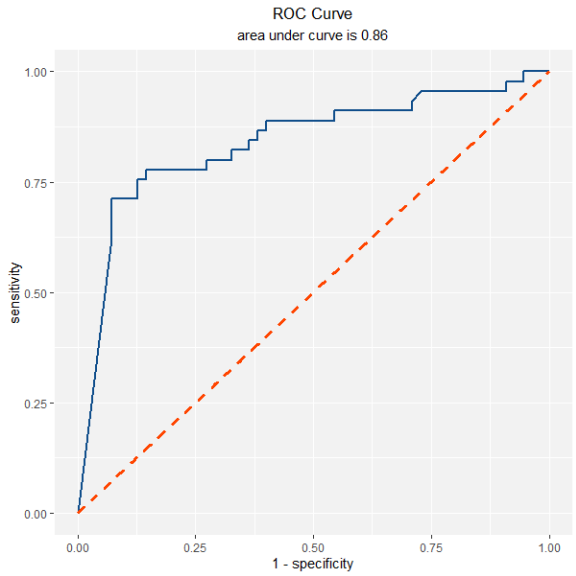


Figure 5.4: ROC curve reduced model with B-splines

and values:

$$F1 - score = 0.7368421 \qquad accuracy = 0.75 \qquad precision = 0.7$$

5.4. MMSE model

Then we performed two models with only the MMSE parameter, with and without B- splines:

$$log(\frac{p}{1-p}) = \beta \cdot MMSE$$

$$log(\frac{p}{1-p}) = \beta \cdot bs(MMSE)$$

This two models confirm that MMSE is a very significative factor, in particular in the first case we have

$$F1 - score = 0.8045977 \qquad accuracy = 0.83 \qquad precision = 0.8333333$$

$$sensitivity = 0.7777778 \qquad specificity = 0.8727273$$

Adding the splines reduced the performances.

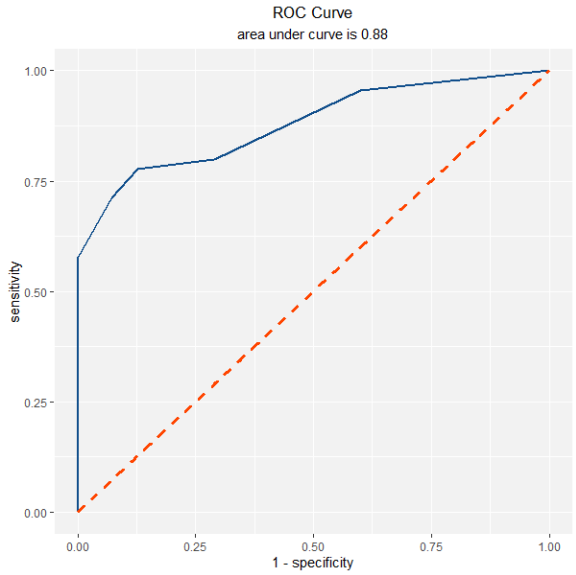


Figure 5.5: ROC curve MMSE model without splines

5.5. Models without MMSE

Since we have seen that the most influential covariate is the MMSE we built robust regression models without this parameter to see what are the most influential parameters that we can be used in prediction if we do not know the Mini Mental State Examination of a patient. As first model we use

$$log(\frac{p}{1-p}) = \beta_1 \cdot M + \beta_2 \cdot bs(Age) + \beta_3 \cdot bs(nWBV) + \beta_4 \cdot bs(eTIV)$$

and we find the following ROC curve

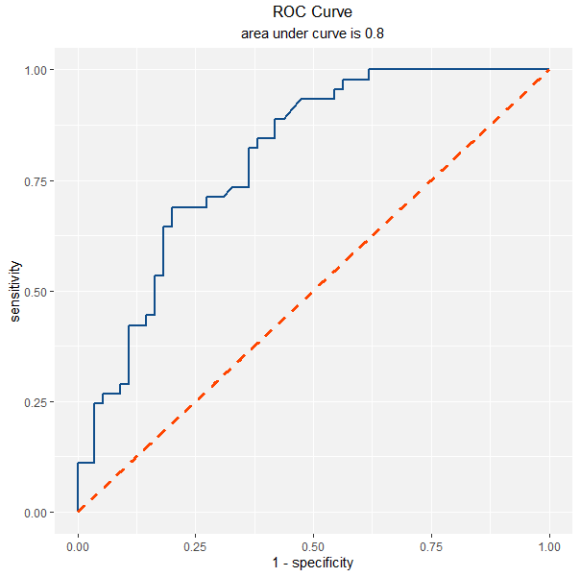


Figure 5.6: ROC curve model without MMSE

with the following values:

$$F1 - score = 0.733945 \qquad accuracy = 0.71 \qquad precision = 0.625$$

$$sensitivity = 0.6888889 \qquad specificity = 0.7636364$$

Then we reduced the model considering covariates with and without the splines . As final model we consider the one with only nWBV:

$$log(\frac{p}{1-p}) = \beta \cdot bs(nWBV)$$

Finding the following ROC curve

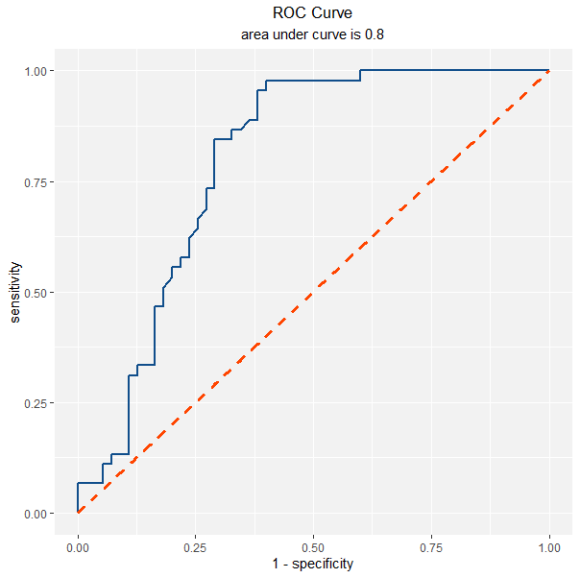


Figure 5.7: ROC curve model with only nWBV

With the following values:

$$F1 - score = 0.754717 \qquad accuracy = 0.74 \qquad precision = 0.6557377$$

$$sensitivity = 0.7777778 \qquad specificity = 0.6727273$$

Another good model is the following one: $log(\frac{p}{1-p}) = \beta_1 \cdot M + \beta_2 \cdot nWBV + \beta_3 \cdot eTIV$
With the following ROC curve

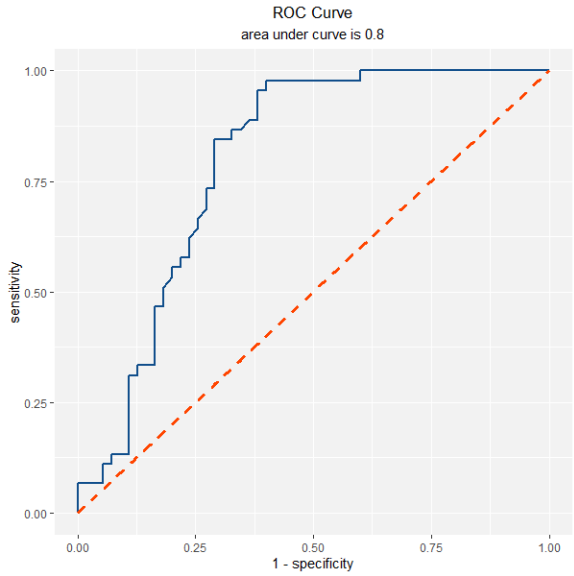


Figure 5.8: ROC curve

and values

$$F1 - score = 0.7927928 \qquad accuracy = 0.77 \qquad precision = 0.666667$$
$$specificity = 0.7090909 \qquad sensitivity = 0.7333333$$

5.6. Conclusions

In conclusion we can say that the best model is the one with all the covariates and with splines. If we want to consider only the MMSE the model is better without the splines, and if we want to make prediction without the use of the MMSE the best model consider the gender, the NWBV and the eTIV without the use of splines.

Chapter 6

CONFORMAL CLASSIFICATION: Inductive Conformal Prediction (ICP)

This part refers to [**1**]

6.1. Introduction to the method

For this kind of analysis we used the dataset divided in training and test set presented before, in order to satisfy the assumption of iid data, to ensure exchangeability.
This dataset X_{train} is composed by $n = 271$ samples and X_{test} is composed by $n = 100$. We considered 6 feaures: Age, EDUC, SES, MMSE, eTIV, nWBV.
The label space is $Y \in (Demented, NonDemented)$ so the output of the conformal prediction will be based on the classes: (Demented, NonDemented, Demented-NonDemented). The object space is denoted by $X \in R^p$, where $p = 6$ is the number of features. We assume that each observation consists of a sample and its label, and its space is given as Z := X × Y.

6.1.1. Conformity measure

The conformity measure can be defined as: $1 - A$. Where A is a Nonconformity measure.

NonconformityMeasure
A nonconformity measure is a measurable function $A : Z \times Z \to R$ such that A(Z_1, Z_2) does not depend on the ordering of observations in the set Z_1.

Since we are tackling a Classification problem, we need to use a particular conformity measure that exploit the classification. The one presented in [**1**] uses the random forest method.

Random Forests Method
This method is based on Classification trees. Then the conformity score is the proportion of votes for each class: the ratio between the number of trees in the forest voting for a given class divided by the total number of trees, which is 100.

$$\alpha_i(y) = \frac{\sharp\ trees voting\ for\ class\ y}{\sharp\ of\ trees}$$

We denote by $\alpha_i(y)$, the conformity score for i^{th} observation for class y. Each component $\alpha_i(y)$ that corresponds to the sample (x_i, y_i) is computed the equation above based on the augmented sample $z_1, ..., z_n, z_{n+1} = (x_{new}, y)$.
Then p-value describes the lack of conformity of the new observation x_{new} to the training set Z.

$$p_y = \frac{|z_i \in Z :\ y_i = y, \alpha_i(y) < \alpha_{new}(y)| + u_i * |z_i \in Z :\ y_i = y, \alpha_i(y) = \alpha_{new}(y)|}{n_y + 1}$$

where $u_i \sim U[0, 1]$, n_y denotes the number of observations having the true label as class-y in the training set. The p-value $p(y) = p_y$, $y \in Y$ lies in $(\frac{1}{n_y+1}, 1)$. The smaller the p(y) is, the less likely the true pair is (x_{new}, y). Multiplying the borderline cases by u_i results in what are known as smoothed conformal predictors.

6.1.2. Transductive Conformity Prediction (TCP)

Given a training dataset Z and a new observation x_{new}, the TCP, corresponding to a nonconformity measure A, checks each of a set of hypothesis (for all possible labels) for the new observation x_{new}, assigns to it a p-value, and finds the prediction region for the test set x_{new} at a significance level $\epsilon \in (0, 1)$.
The fully on-line mode of TCP can be very computationally demanding: the learning algorithm updated for each new data point. Therefore we introduce an off-line version of the method which uses a batch of sample at each iteration.

6.1.3. Inductive Conformal Prediction (ICP)

It's a batch-mode version of the TCP, in which the training set of n= 271 samples is partitioned into two different sets:

1. Proper Training set: $Z_p = z_1, ..., z_q$ of size q=271*0.8
2. Calibration set: $Z_c = z_{q+1}, ..., z_n$ of size n-q

The idea is to check how well the calibration set conforms to the proper training set.
The ICP p-value is then computed as

$$p_y = \frac{|z_i \in Z_c : \ y_i = y, \alpha_i(y) < \alpha_{new}(y)| + u_i * |z_i \in Z_c : \ y_i = y, \alpha_i(y) = \alpha_{new}(y)|}{n_y + 1}$$

where $u_i \sim U[0, 1]$ and n_y denotes the number of observations having the true label as class-y in the calibration set.

6.1.4. Algorithm

Input: (training dataset: Z, test data: xnew, label set: Y, a nonconformity measure: A)
Output: p-values
partition Z into proper training set Z_p and calibration set Z_c
Compute nonconformity scores: $\alpha_i(y) = A(Z_p, z_i)$for each $z_i \in Z_c$
Compute nonconformity score for test observation: $\alpha_{new} = A(Z_p, (x_{new}, y))$ for each $y \in Y$
return p-values

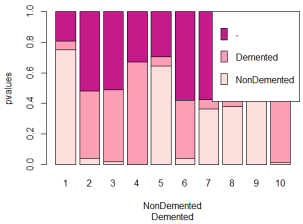


Figure 6.1: p-values of the first 10 sample test

6.2. Evaluation of performace

A predictor makes an error when the predicted region does not contain the true label, that is $y \notin |\Gamma^\epsilon|$. Given a training dataset Z and an external test set Z_t, and $|Z_t|$= m. Suppose that a conformal predictor gives prediction regions as $\Gamma^\epsilon_1, ..., ..., \Gamma^\epsilon_m$, where $\epsilon = 0.05$ then the error rate is defined as follows(using seed=3):

Error rate

$$ER^\epsilon = \frac{1}{m} \sum_{i=1}^{m} I_{y_i \notin \Gamma^\epsilon_i} = 0.1$$

where y_i is the true class label of the i-th test case and I is an indicator function. The efficiency can be computed as the ratio of predictions with more than one class over number of observations in the test set.

Efficiency

$$EFF^\epsilon = \frac{1}{m} \sum_{i=1}^{m} I_{|\Gamma^\epsilon_i|>1} = 0.5$$

The Observed fuzziness is defined as the sum of all p-values for the incorrect class labels.

Observed Fuzziness

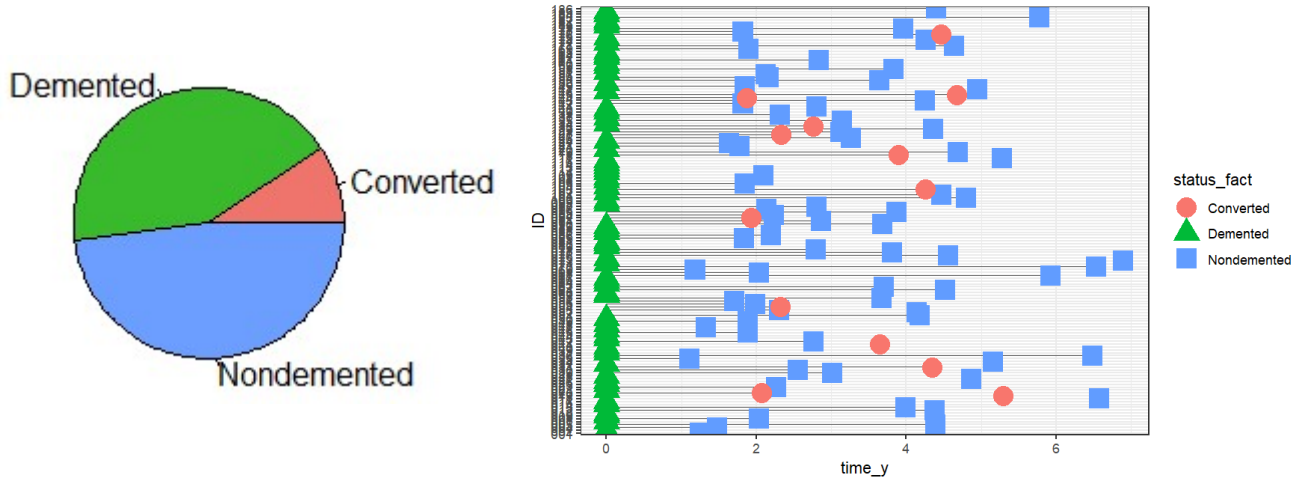
$$ObsFuzz = \frac{1}{m} \sum_{i=1}^{m} \sum_{y_i \neq y} p_i^y = 0.15$$

We note that for the above measure of performances, smaller values are preferable.
We can notice that except for Efficency the values are quite small, and thus we can deduce that for majority of the samples not just one class is identified.

Chapter 7

SURVIVAL ANALISIS

As shown before our dataset 'cross-longitudinal' collects data from 100 patients and for each of them there are from 1 to 5 visits. This is a perfect example of censored data, where the event is the occurrence of the Dementia.
As we can see from the image below 7.1, where time is expressed in years, we had a lot of patients which presented the deasese at the beginning of the study. This fact influenced a lot the analysis and affected the results, which may differ from the usual results about Dementia.



(a) pie plot of the dataset

(b) time to event data plot

Figure 7.1: In blue the censored data, the other ID have experienced the event: in green the ID which present the desease at the first visit, in red the ones which got the desease during the observation

7.1. Kaplan-Meier estimator

Our main goal was to study the Survival function, and how it's influenced by the features of the patient.

$$S(t) = P(T < t)$$

To estimate this function we can use the Kaplan-Meier estimator:

$$\hat{S}(t) = \prod_{i:\ t_i^* < t} p_i$$

where p_i = probability of surviving time t_i^*.
This are the results if we don't consider any additional feature

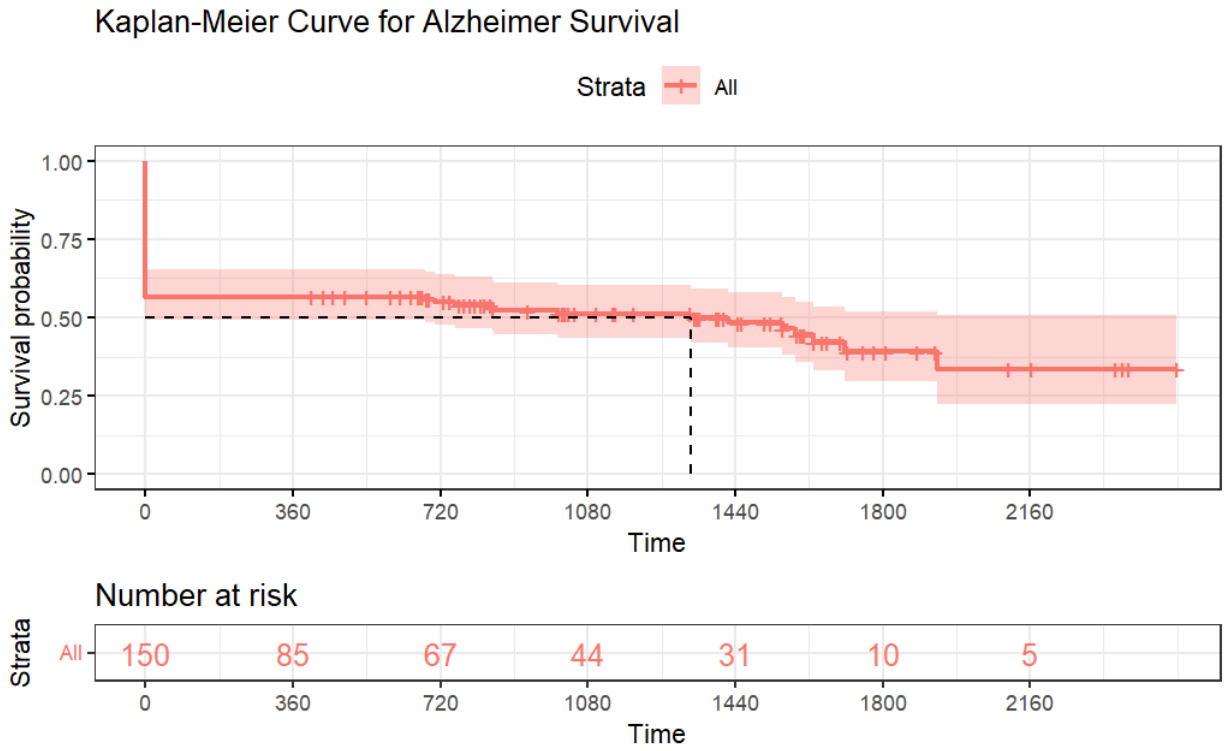


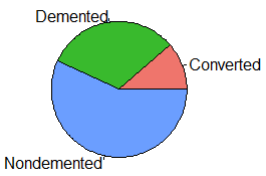
Figure 7.2

As mentioned before a lot of patients presented the desease at the beginning of the study, thus our Survival probability is less the 100% at time 0, it's around 60%. It has possibly influenced also the median time of survival, which is around 3 and a half years. We can also notice that at the end of the study the probability of surviving is larger then zero, that's given by the fact that not all patients got the desease.

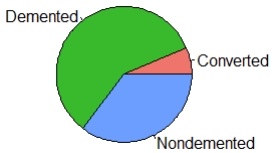
Kaplan-Meier estimator with Gender classification

$$S(t) \sim Gender$$

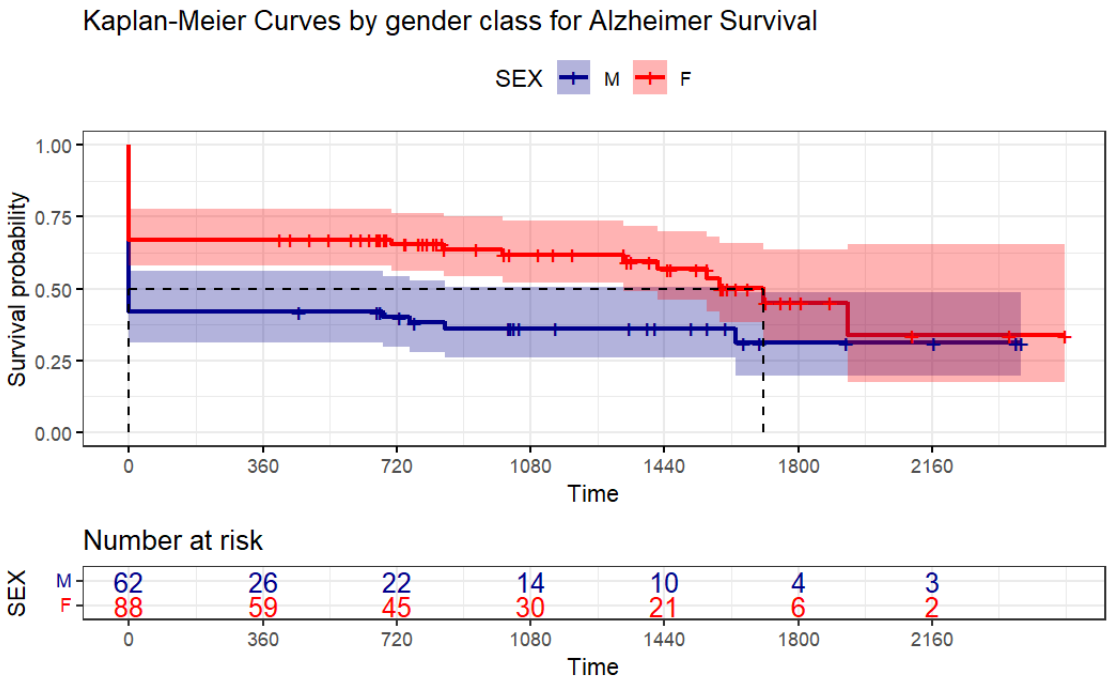
To get more information about the Survival function we added the influece of some features: first of all applying a permutational anova test on our data we found out that the Gender is influential, thus we use it as a covariate. From the plot 7.3c we observe a difference in the curves of the estimator until the last years, where the curves overlap. Also the survival median looks different: For females it's almost 5 years, while Male starts with an estimate of the Survival probability wich is already less then 50%.



(a) Pie plot of the female time to event dataset



(b) Pie plot of the male time to event dataset



(c) Kaplan Meier estimator with gender classification

Figure 7.3

To understand if there's statistical evidence to affirm so, we apply the Long-rank test:

$$H_0 : \ S_{male}(\cdot) = S_{female}(\cdot) \ \ vs \ \ H_1 : \ S_{male}(\cdot) \neq S_{female}(\cdot)$$

The p-value of the test is 0.01. We can conclude that there's evidence to reject H_0 with a confidence level $\alpha = 5\%$ and thus the curves divided by sex differ.
Computing the Hazard ratio we can confirm that being a Male is a risk factor for our analysis, indeed $HR_{M,F} = 1,57$

7.2. Cox proportional hazards regression analysis: Age

Kaplan-Meier curves and logrank tests are more useful when the predictor variable is categorical, which is not the case. Thus we use the Cox model, which is able to to assess simultaneously the effect of several risk factors on survival time. We would like to examine the effect of the Age variable as a risk factor via a univariate regression model:

$$h(t) = h_0(t) \exp(\beta_{Age} Age)$$

The variable Age is distributed as shown in the histogram 7.5
Analyzing the result given by the model:

1. Regression coefficient is negative, -0.037: older patients are less at risk
2. Statistical significance: the variable age is statistically significant at 2%. (i.e. the beta coefficient of the variable is statistically different from 0)
3. Global statistical significance of the model: the likelihood-ratio test, the Wald test, and the score logrank statistic, all gave a p-value of 0.01. The model is globally significant (already expected because the only variable was stated as significant)

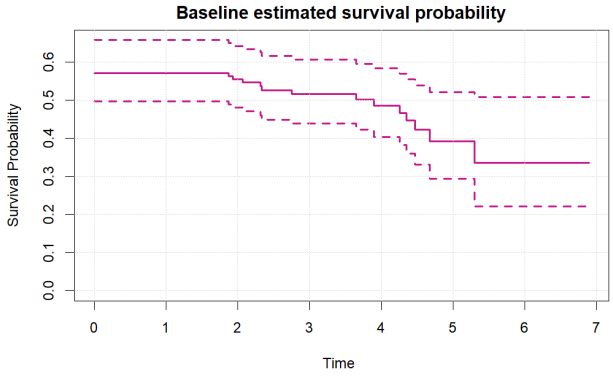


Figure 7.4: cox model wrt Age variable

4. Hazard Ratio and Confidence interval 95%:

Lower bound	Fit	Upper bound
0.9351	0.9629	0.9916

Every passing year the risk decreases since the HR is <1.

7.3. Cox proportional hazards regression analysis: Age and Sex

To go more in depth we added the effect of the gender class:

$$h(t) = h_0(t) \exp(\beta_{Age} Age + \beta_{Gender} Gender)$$

Results are as follow:

- Variable Age: the result looks similar to the previous one, still significant at 2% and every year the risk decreases by $3,7\%$
- Variable Sex:
  1. Regression coefficient is positive: being male is associated with bad prognostic.
  2. Statistical significance: the variable age is statistically significant at 2%. (i.e. the beta coefficient of the variable is statistically different from 0)
  3. Global statistical significance of the model: the likelihood-ratio test, the Wald test, and the score logrank statistic, all gave a p-value of 0.002. The model is globally significant.

Hazard Ratio and Confidence interval 95%

Variable	Lower bound	Fit	Upper bound
Age	0.9358	0.9640	0.9931
Sex	0.5595	0.3583	0.8737

We obtained the same results as seen before in term of interpretation of the HRs.

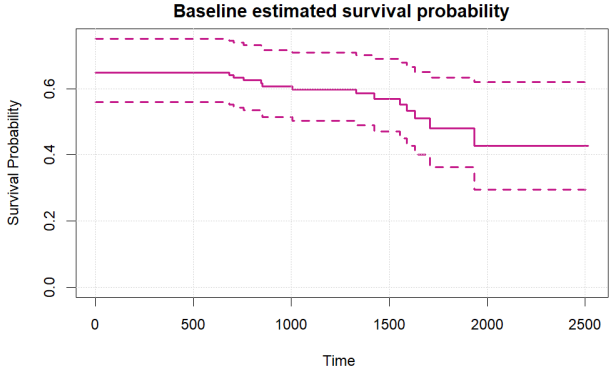


Figure 7.5: cox model wrt Age and Sex variables

Chapter 8

CONCLUSION

After our analysis we can conclude that we have found efficient logistic methods both robust and non robust, which can be used when we don't have the complete history of the patients.
While, if we want to take in consideration the clinical history of patients we conclude that, using our sample patients, male are more at risk of Dementia and also that getting older provide less possibility of contracting the desease.

Bibliography

[1] Niharika Gauraha, Ola Spjuth ,*conformalClassification: A Conformal Prediction R Package for Classification*

[2] Eva Cantoni, *Analysis of Robust Quasi-deviances for Generalized Linear Models*, Department of Econometrics, University of Geneva

[3] Eva Cantoni, Elvezio Ronchetti, *Robust Inference for Generalized Linear Models*, Journal of the American Statistical Association, published by Taylor Francis