Quantitative Data Analysis

Federico Alessandro Tullio Riva

1944264

Introduction

The aim of this assessment is to run an investigation of the Aids2ann dataset, which contains annualised data from Australian Aids cases until 2001. The analysis was performed using R as it has powerful statistical tools. Each process is justified in detail by showing the related code, the results obtained and an explanation of such results.
The assessment will be divided into three parts. In the first section, the data is explored and the key characteristics are set out. In the second one we begin to point out associations between variables. In the last section we use a logistic regression to determine which variables affect the outcome.

Section 1

The dataset has a dimension of 6014 observations and 9 variables. It was saved into R by typing *Aids2ann <- read.csv("Aids2ann.csv", header=TRUE)*. The observations start precisely in 1992 and end in 2001 (*unique(Aids2ann$year)*).
None of the variables has missing data, this was found with the command *aggr(Aids2ann)* from the library "VIM" which plotted a table with the percentages of missing data in each variable (Figure 1).
We then proceeded to investigate if some quirks were found in the data. The investigation started from the categorical data. For each of the variables in the first place the *unique()* function was used to investigate how many variables we were dealing with, then the *table()* function was used to find out the frequencies of each variable. The *table(Aids2ann$sex)* produced the first interesting data. As we can see from Figure

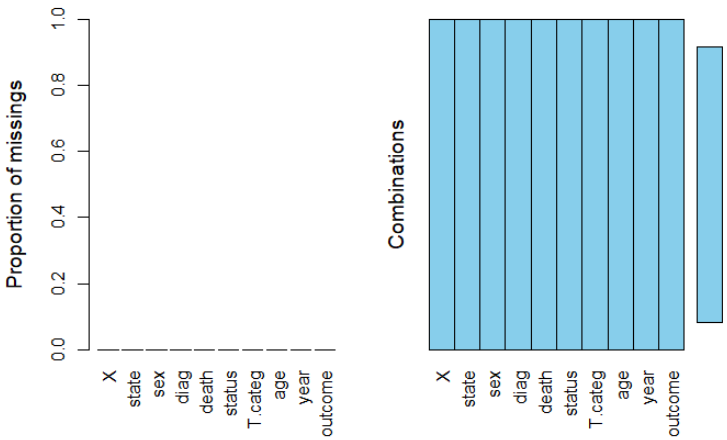

Figure 1: Missing data

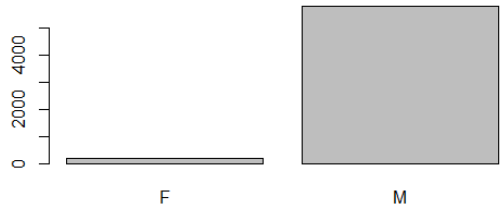

Figure 2: barplot(table(Aids2ann$sex))

2, the number of men is significantly higher than that of the women (the exact numbers are F: 202 and M: 5812). This means that specifically in the regression we would not expect the sex to be very relevant since most of the observations on which we have data portray men.

Furthermore by plotting *barplot(table(Aids2ann$T.categ))* and *barplot(table(Aids2ann$state))* other similar oddities can be found. As shown by Figure 3, in the states variable, there is clear a dominance of the NSW state. In New South Wales there are indeed more aids cases than in the other three states combined (from the output of *table(Aids2ann$state)*: NSW: 3775, Other :544, QLD: 446, VIC:1249). In Figure 4, we can observe than the hs value is by far the most relevant transmission category (blood: 187, haem: 89, het: 102, hs: 5217, hsid: 168, id: 108, mother: 15, other: 128). Both the T.categ and the state values, since they have these two particularities, will produce some particular results when associated to other variables.

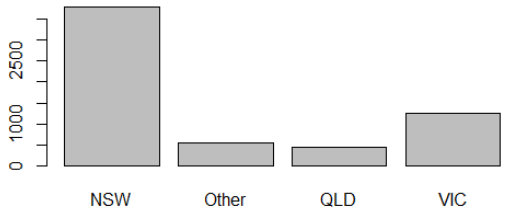

Figure 3: barplot(table(Aids2ann$state))

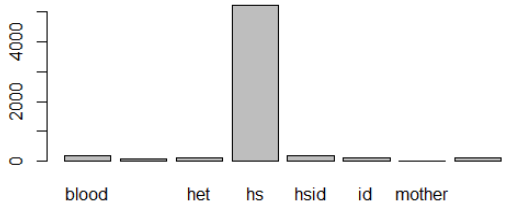

Figure 4: barplot(table(Aids2ann$T.categ))

When investigating numerical data, we have run the same commands to each variable to determine their distribution. The set of command is shown in the text box alongside, where, by investigating one variable, we must change the *var* value with the name of the variable.

```
quantile(Aids2ann$var,p=c(0,0.025,0.25,
0.5,0.75,0.975,1))

h.var <- hist(Aids2ann$var)

mu.var <- mean(Aids2ann$var)

sig.var <- sd(Aids2ann$var)

plot(Aids2ann$var,dnorm(Aids2ann$var,
mean=mu.var,sd=sig.var))

hist(Aids2ann$var,freq=F)

points(Aids2ann$var,dnorm(Aids2ann$var,
mean=mu.var,sd=sig.var))

qqnorm(Aids2ann$var)

qqline(Aids2ann$var)
```

The first variable analysed was the diag, which represented the day aids was diagnosed at the patient (number of days from 1st January 1970). The results are very regular in the *quantile()* function (outcome: 0%: 8302.000, 2.5%: 9174.575, 25%: 10116.000, 50%: 10537.000, 75%: 10947.750, 97.5%: 11416.000, 100%: 11503.000). This should have predicted a normal distribution

when plotted. The histogram and normal line plotted above it, shown in Figure 5, suggest a normal distribution is being created. The line, though, is interrupted before the curve could finish. The problem here is that the end of the observation is set at a specified number of days, but observations keep coming in until the last date available, so we cannot experience the last part of the descending line.

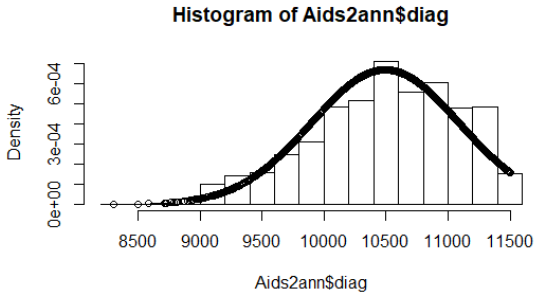

Figure 5: histogram and plot of normal distribution for the diag variable

We obtained a completely different result when running the model on the death variable. By running the first two commands (*quantile()* and *hist()*), we can clearly see that this variable does not have the potential to be distributed normally. It has indeed a spike at the very end, as seen in Figure 6, and by the outcome of the quantile function. The results are the same from the 75% quantile (0%: 8469.000, 2.5%: 9576.975, 25%: 10809.000, 50%: 11337.000, 75%: 11504.000, 97.5%: 11504.000, 100%: 11504.000). This phenomenon is due to the extremely high number of concluded observations at the end of the period. The problem is therefore the same as the diag variable discussed above, but here it is exponentially more interesting. It will produce oddities when related to other variables, which implies that it could not be suitable to make good predictions.

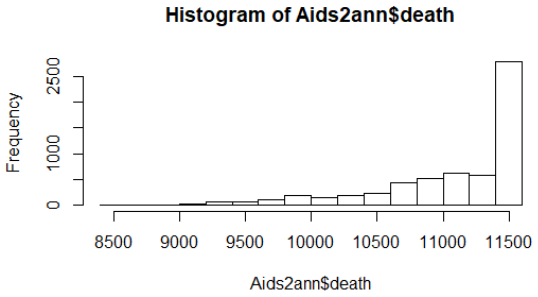

Figure 6: histogram of the death variable

On the contrary, a variable which produced a nearly perfect normal distribution is the age variable. The outcome of the quantiles (0%: 0, 2.5%: 23, 25%: 31, 50%: 37, 75%: 43, 97.5%: 59, 100%: 82) and the histogram with the plotted normal line above clearly show the pattern (Figure 7). This result is really consistent with the previous ones, since the possibilities of contracting Aids are bigger when people are early middle-aged (50% quantile is 37). When individuals are young or old they are less exposed to possible transmission's hazards. The Q-Q

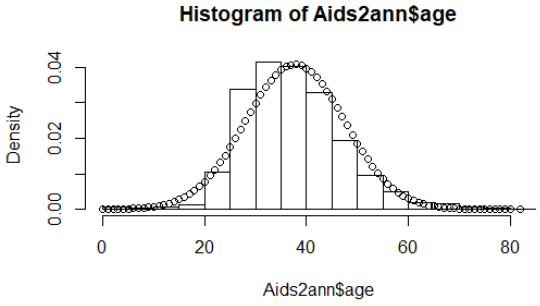

Figure 7: histogram and plot of normal distribution for the age variable

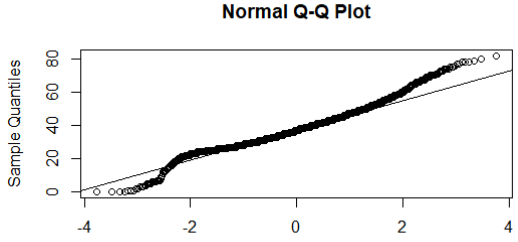

Figure 8: Q-Q plot for the age variable

plot for the age variable (Figure 8) supports our thesis by plotting a good diagonal line. The only point that goes off the trend is from 15 to 20 years, in which there is a sudden spike; this is due to a lack of cases in that age-range. This leads to a higher than expected 2.5% quantile, set at 23, in the dataset.

Section 2

In section 2, we are going to explore some of the pairwise associations between the variables. The first two variables that we confronted are age and T.categ. The aim was to see if in the boxplot we could find any pattern. The results (Figure 9), are quite interesting. All the variables, but one, follow the regular pattern seen before: the average age for each transmission category stands always between 30 and 50. The only exception is for the mother category, in which the ages' mean is set at around 5. This is supported by two t.tests for the mean. Since the

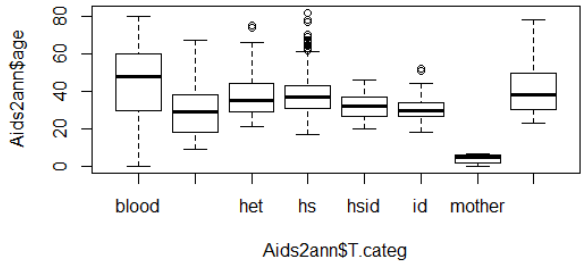

Figure 9: boxplot(Aids2ann$age~Aids2ann$T.categ)

```
    -  Welch Two Sample t-test

data:  Aids2ann$age by Aids2ann$T.categ == "mother"
t = 52.352, df = 15.109, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 32.32302 35.06492
sample estimates:
mean in group FALSE  mean in group TRUE
          37.827305            4.133333

    -  Welch Two Sample t-test

data:  Aids2ann$age by Aids2ann$T.categ == "het"
t = -0.30719, df = 103.53, p-value = 0.7593
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -2.616362  1.914520
sample estimates:
mean in group FALSE  mean in group TRUE
           37.73731            38.08824


    -  Welch Two Sample t-test

data:  Aids2ann$age by Aids2ann$T.categ == "hs"
t = -2.1917, df = 869.29, p-value = 0.02866
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -2.3561021 -0.1298966
sample estimates:
mean in group FALSE  mean in group TRUE
           36.66499            37.90799
```

T.categ has many variables, we ran two tests: *t.test(Aids2ann$age~Aids2ann$T.categ=="mother")* and *t.test(Aids2ann$age~Aids2ann$T.categ=="het").* The first one checks if the difference between the age mean for the mother was significantly different from that of the others, while the second excluded the het category from the group. We expected to find two different outcomes: in the first test a difference in the mean, while in the second no difference. We checked for the results (in the textbox above) and proved our thesis (the first p-value is way smaller than 0,1% which leads us to reject the null hypothesis and state that the two means are different, and the second is way bigger than 10% which means that we cannot reject the null hypothesis: the two means are equal). The extreme drop in the age mean related to the mother category was somehow expected: it is normal that if a mother has Aids, and she does not treat it, her child will be born with Aids as well. Therefore the diagnosis will be done in the first years of the baby's life.

Another t.test was run: *t.test(Aids2ann$age~Aids2ann$T.categ=="hs"),* its results are printed out in the box with the others. We can notice a peculiar characteristic: even though the two means are fairly similar, the p-value is small (2.9% which means we should reject the null hypothesis and, as above, state that the two means are different). This is caused by the *hs* category and by the fact that it comprehends the vast majority of the observations, as we pointed out before.

The second and third analysis were made on the basis of the variable status. In the second investigation the variable was compared to T.categ whether in the third to state. Two chi-square tests for independence were run and the results are printed in the box alongside. Both of them point out that the variable status is highly related to the variable T.categ and state respectively, due to the small p-values (always less than 0.1%). These results mean that there is a correlation between the death of a patient and both the state in which he lived in and the way he got Aids.

```
> chisq.test(Aids2ann$status, Aids2ann$T.categ)
###       Pearson's Chi-squared test
###data:  Aids2ann$status and Aids2ann$T.categ
###X-squared = 52.862, df = 7, p-value = 3.947e-09

> chisq.test(Aids2ann$status, Aids2ann$state)
###Pearson's Chi-squared test
###data:  Aids2ann$status and Aids2ann$state
###X-squared = 18.918, df = 3, p-value = 0.0002843
```

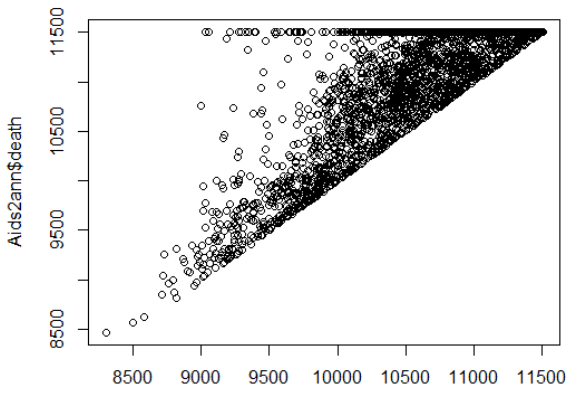

Figure 10: plot(Aids2ann$diag, Aids2ann$death)

The last analysis between two variables was done by looking at the scatterplot of the diag and death

variables (Figure 10). The plot includes two patterns: the diagonal one displays the time passed from diagnosis to death for each patient, while the horizontal one, at the top, coincides with the end of observation. The diagonal, therefore, is what we are interested in. We can see that, at the beginning of the observation, the life expectancy when diagnosed with Aids was not long (the first three points show people who were diagnosed and then died around 8500 days after 01-01-1970). We can see some improvements in the average life expectancy with patients diagnosed around 10000 days after 01-01-1970 due to the "cloud" of points getting bigger. This increase matches also a higher number of patients taken in, so the growth in the average life expectancy could be caused either by an improvement in therapies or just as a result of having more observations.

Section 3

Now we will try to create the best regression model to make predictions on the outcome. The regression will investigate how likely it is for a patient to die in a particular year. The regression used is the logistic one. At the beginning we started with a backward selection method. We had all the variables in the model and we took out those that were not significant (*mod <- glm(outcome~factor(T.categ)+age+year+factor(state)+diag+death+factor(sex), family = binomial, data = Aids2ann)*). At the end of this process, we were left with just two variables: death and year. R, though, did get us a *Warning message: glm.fit: fitted probabilities numerically 0 or 1 occurred.* From this data we could find out that one of the two variables made the outcome go directly to 0 or 1. On a second thought, the death variable could not fit in the regression model because it already specifies when the patient death specifically happens. Therefore, it is not a suitable variable to fit in the regression model.

Having now just the year as a variable, we proceeded with a forward selection method, in which we have thought about which variables could be the most accurate to fit in the model. The first variable that was added was the age, since it previously had a good distribution. The model *mod <- glm(outcome~age+year, family = binomial, data = Aids2ann)* turned out to have both of the variables which were significant (each p-value<0,1%) so we proceeded to add a third variable. The chosen variable was T.categ, and its model and results are in the textbox below. Many of the T.categ dummy variables are not significant in the model, but two of them are (het and id). We proceeded by running an *lrtest(mod, mod1)* and found out that there was significance in the difference between the two models, which suggested that the new one, with the T.categ inserted, was better than the one before.

For every other variable that we tried to add, the lrtest gave us a p-value>10% which meant that

there was not a strong difference between the models. Therefore, since when two models are not significantly different we prefer to use the one with less variables, we kept *outcome~factor(T.categ)+age+year* as our final logistic regression model. When running the *confint()* function on the model, we find out the confidence intervals for each covariate. Every

```
> mod1 <- glm(outcome~factor(T.categ)+age+year, family = binomial, data = Aid
s2ann)
> summary(mod1)

Call:
glm(formula = outcome ~ factor(T.categ) + age + year, family = binomial,
    data = Aids2ann)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.3510  -0.8465  -0.7658   1.4134   2.0266

Coefficients:
                       Estimate Std. Error z value Pr(>|z|)
(Intercept)          263.467113  33.707946   7.816 5.45e-15 ***
factor(T.categ)haem    0.040302   0.279923   0.144  0.88552
factor(T.categ)het    -0.910690   0.308999  -2.947  0.00321 **
factor(T.categ)hs     -0.222486   0.158404  -1.405  0.16015
factor(T.categ)hsid   -0.291241   0.236499  -1.231  0.21815
factor(T.categ)id     -0.696145   0.300818  -2.314  0.02066 *
factor(T.categ)mother -0.125644   0.678414  -0.185  0.85307
factor(T.categ)other  -0.140953   0.247262  -0.570  0.56864
age                    0.015162   0.002999   5.056 4.28e-07 ***
year                  -0.132418   0.016878  -7.846 4.30e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 7272.8  on 6013  degrees of freedom
Residual deviance: 7160.3  on 6004  degrees of freedom
AIC: 7180.3

Number of Fisher Scoring iterations: 4

> lrtest(mod1, mod)
Likelihood ratio test

Model 1: outcome ~ factor(T.categ) + age + year
Model 2: outcome ~ death
  #Df  LogLik Df  Chisq Pr(>Chisq)
1  10 -3580.2
2   2 -3045.7 -8 1068.9  < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


> confint(mod1)
                             2.5 %       97.5 %
(Intercept)          197.402417281 329.56038221
factor(T.categ)haem   -0.515740828   0.58413329
factor(T.categ)het    -1.540930371  -0.32380850
factor(T.categ)hs     -0.530466135   0.09136894
factor(T.categ)hsid   -0.758831712   0.16955787
factor(T.categ)id     -1.306052096  -0.12200801
factor(T.categ)mother -1.651905321   1.09704985
factor(T.categ)other  -0.630350131   0.34047716
age                    0.009289334   0.02104704
year                  -0.165511803  -0.09934025
```

covariate (in the T.categ variable) which does not produce a significant p-value has a confidence interval astride zero. Among the others het, id and year have a negative impact on the regression, while the intercept and age contribute to the model with a positive effect. By running the *residual()* function on the model and plotting the results over the years, we could see that the regression has some errors that span from -1 to up to 2 years, which should not be a bad result (Figure 11). The more significant issue is that none of the residual falls close to zero, which means that each prediction made will have an error.

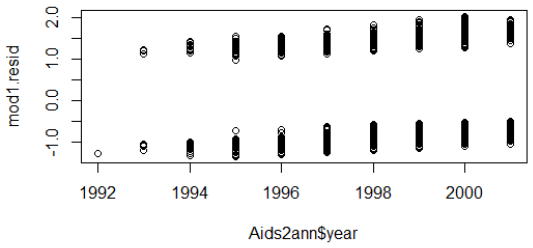

Figure 11: plot(Aids2ann$year, mod1.resid)

The regression model as it is represents the best way to make predictions about the dataset that we have. Its strengths lie in the low number of variables used to create the model itself. Furthermore, the two numerical variables have a really high p-value, which adds effectiveness to the equation. The weakness found in the model is that the T.categ produce many covariates, most of which are not significant. The problem, as we stated before, may come from the fact that many of the observations fall in the hs category. The result of this is a fairly good prediction model which, unfortunately, will never give the exact result.

Conclusion

The aim of the assessment was to investigate thoroughly the Aids2ann dataset, find quirks in the observations, and try to create a regression model which suited the data to make predictions. We could observe some singularities in the dataset, related to the high number of observations we had of just one category over some variables. This fact led to difficulty in reaching a very good prediction model. The other issue which made some results difficult to interpret was due to the end of observations set to a sudden date. Some results, given this problematic, were harder to read and it has not been easy to find patterns within them.

The resulting logistic regression turned out to be a simple model, which, although, could never predict the exact year of death for a patient.