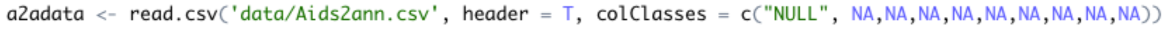
Exploring the data

Possible issues with dataset

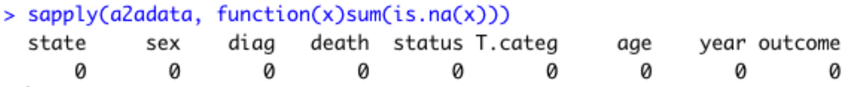
Dataset with which inferences were carried out contained annualized data on patients with AIDS in Australia from 1992 till 2001. The dataset was supposed to contain 6014 observations on 9 variables. To confirm this was correct **dim(a2adata)** code was used. However, it was possible to observe that dimension of the dataset was not as expected, as the codes computed output was: **[1] 6014 10**. An additional unnamed column containing integers, ranging from 1 to 28411, was found. As this seemed like a mistake and irrelevant data, it was decided to disregard the first column of the dataset by only reading the data from the other 9 columns using the **colClasses** argument from the **read.csv** function.



```
a2adata <- read.csv('data/Aids2ann.csv', header = T, colClasses = c("NULL", NA,NA,NA,NA,NA,NA,NA,NA,NA))
```

The dimension was then checked again to check if was as expected; the following was the output of the function: **> dim(a2adata)** output: **[1] 6014 9.**

The dataset was then checked for any potential missing values with the following function:



```
> sapply(a2adata, function(x)sum(is.na(x)))
  state     sex    diag   death  status T.categ     age    year outcome 
      0       0       0       0       0       0       0       0       0 
```

It was possible to conclude from the result and that the dataset did not contain any missing values. **head(a2adata)** returned the first parts of the data frame; from this it was seen that the variable **outcome** had numerical values "0" and "1" and it was considered as an integer. However for the purpose of our investigation it was required to convert the type of the variable by telling R to treat it as a factor: **outcome <- a2adata$outcome, outcomeaf <- as.factor(outcome).** The variables were assigned in advance the following way

```
state <- a2adata$state; sex <- a2adata$sex; diag <- a2adata$diag;
death <- a2adata$death; status <- a2adata$status; tcat <- a2adata$T.categ; age <- a2adata$age; year <- a2adata$year; outcome <- a2adata$outcome; outcomeaf <- as.factor(outcome);
```

Summary statistics - categorical variables

Variable - state

Table 1

States	NSW	Other	QLD	VIC
Frequency	3775	544	446	1249
Relative Frequency	0.628	0.090	0.074	0.208
Approximate%	62.8%	9%	7.4%	20.8%

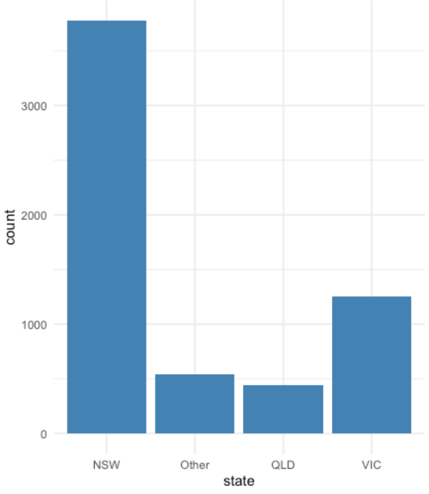


Figure 1

The **state** categorical variable contained 4 different categories: New South Wales (NSW), Queensland (QLD), Victoria (VIC) and Other. The variable state has been appropriately summarised using a table indicating the frequency, relative frequency and the respective percentage. A frequency bar chart was made in order to visualize the data, rather than a pie chart as there were more than 2 categories for the variable and that would have made it more difficult to read. It was possible to observe from the distribution in Figure 1, that the majority (more than 60%) of the observations of AIDS diagnosis were from the NSW. This could be due to NSW being more densely populated than the rest of the states. The least amount of observations was from the QLD state, with only 446 of 6014 observations. observations. Code used to gain the statistics and plots.

```
#check how many categories
> levels(state)
[1] "NSW"   "Other" "QLD"   "VIC"
state.freq <-table(state)
state.freq
rel.freq = state.freq/length(state)
rel.freq
ggplot(a2adata,aes(x = state)) + geom_bar(fill = "steelblue") + theme_minimal()
```

similar code was used for the rest of the categorical variables.

Variable – sex

Sex	F	M
Frequency	202	5812
Relative Frequency	0.033	0.966
Approximate%	3.3%	96%

Table 2

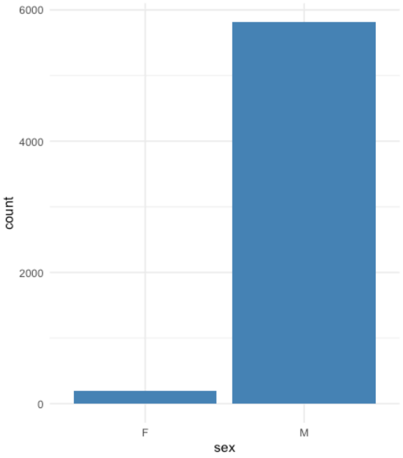


Figure 2

The nominal categorical variable **sex**, was investigated next. There were two categories for sex; "F" for female and "M" for male. Summary of the variable was tabulated in **Table 2** and visually represented in **Figure 2**. From the distribution of the sex variable it was possible to observe that vast majority (96%) of patients diagnosed with AIDS were male and only 202 of the 6014 (3.3%) diagnosed patients were female.

Variable – Year

From visualizing the data in the bar chart in **Figure 3**, it was noticeable that from 1992 till year 2000 the number of observations increased year by year. The number dropped by approximately by 2.7% in the year 2001. The highest percentage of patients diagnosed with AIDS was in the year 2000 with 24.4% of observations.

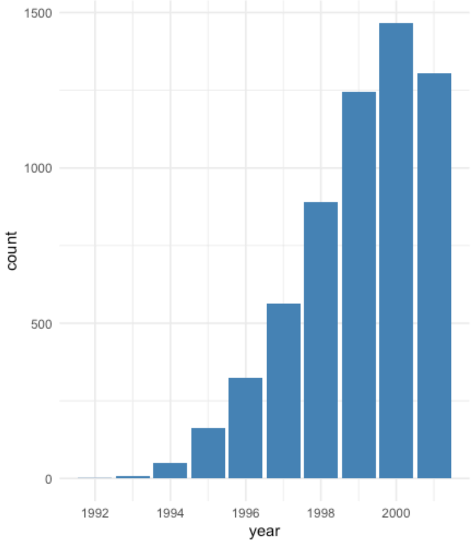


Figure 3

Year	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001
Counts	1	7	51	162	325	562	889	1245	1466	1306
R.F.	0.0002	0.001	0.008	0.027	0.054	0.093	0.148	0.207	0.244	0.217
Approx. %	0.02	0.1	0.8	2.7	5.4	9.3	14.8	20.7	24.4	21.7

Table 3

Variable – T.categ

T.categ variable referred to the AIDS reported transmission category. Again, to check how many categories were present with the categorical variable **levels(tcat)** function was used. The transmission categories found were: **"blood",** receipt of blood, blood components or tissue; **"haem",** haemophilia or coagulation disorder; **"het",** heterosexual contact; **"hs",** male homosexual or bisexual contact; **"hsid"**, as **hs** and also intravenous drug user; **"id",** female or heterosexual male intravenous drug user; **"mother"**, mother with AIDS; **"other",** other or unknown. Summary for the variable was tabulated in **Table 4** and visually represented in **Figure 4.**

category	blood	haem	het	hs	hsid	id	mother	other
Frequency	187	89	102	5217	168	108	15	128
Rel. Freq.	0.031	0.015	0.017	0.867	0.028	0.018	0.002	0.021
%	3.1%	1.5%	1.7%	86.7%	2.8%	1.8%	0.2%	2.1%

Table 4

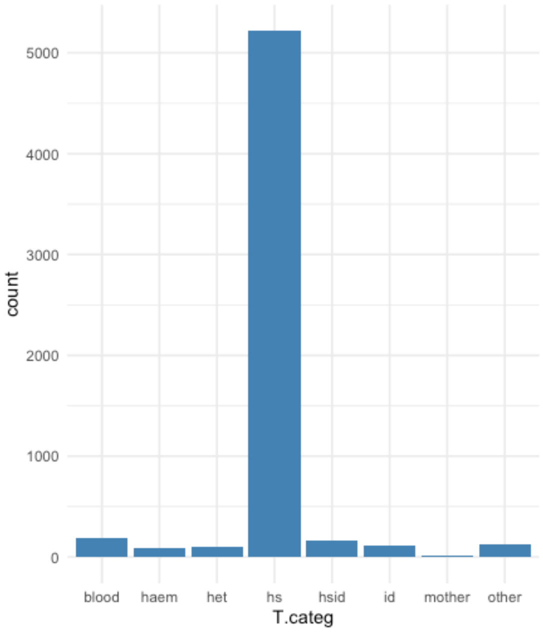


Figure 4

From the bar chart in **Figure 3**, it has been observed that great majority of patients being transmitted AIDS were reported to be in category **hs.** This meant most diagnosed patients had contact with homosexual or bisexual male. From **Table 3**, we can see that almost 87% of the observations were reported under category **hs.** The second most common reported category were patients being diagnosed with AIDS as result of receiving blood, blood components or tissues. The least amount of patients diagnosed with AIDS were under **mother** transmission category; only 15 out of the 6014 observations were reported. This indicated very small portion of patients had inherited AIDS from their mother from birth.

Summary statistics - quantitative variables

Statistic	age	diag
Mean	37.743	10493.05
SD	9.776	594.976
Variance	95.575	353996.2
Range	82	3201
Minimum	0	8302
Maximum	82	11503
Lower quartile	31	10116.00
Median	37	10537
Upper quartile	43	10947.75
IQR	12	831.75
No. of Outliers	158	20

Table 5

Variable – age

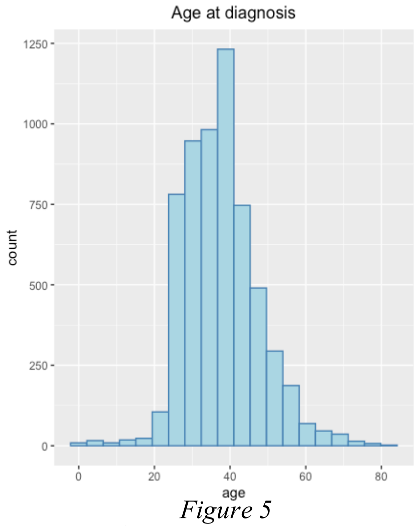


Figure 5

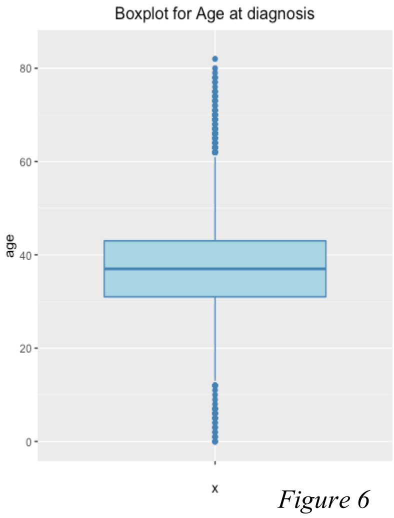


Figure 6

From the histogram of age at the time of diagnosis in **Figure 5** it appeared the data was normally distributed, with no missing data, as also checked earlier in the preliminary tasks. This reflection can be confirmed by the values of the mean and the median from **Table 5**, being very close; this indicated the data being symmetric and approximately close to normal distribution. The boxplot in **Figure 6** helped to visually identify, a number of potential outliers present. To investigate and find the total number of outliers present in the data, a function **find_outliers(x)** was written. In this function, any value was classified as an outlier if it was below $Q1-1.5\times IQR$ or above $Q3-1.5\times IQR$. The total number of outliers identified were 158. The median is less affected by these outliers compared to the mean, therefore median was used as the centre of the data. The code used to gain the statistics and plots is as follow:

```
mean(age), median(age)
sd(age), range(age)
var(age), quantile(age)
IQR(age)
```

FIND OUTLIERS FUNCTION

```
find_outliers <-function(x) {
  lower_threshold = quantile(x,.25) - 1.5*IQR(x)
  upper_threshold = quantile(x,.75) + 1.5*IQR(x)
  result <- which(x<lower_threshold | x>upper_threshold)
  y = length(result)
 return(y) }
```

HISTOGRAM code

```
ggplot(a2adata, aes(x = age)) + geom_histogram(color = "steelblue",fill = "lightblue",
bins=20)+ ggtitle("Age at diagnosis") +theme(plot.title = element_text(hjust = 0.5))
```

BOXPLOT code

```
ggplot(a2adata, aes(x = "", y = age)) +
  geom_boxplot(color = "steelblue",fill = "lightblue") +  ggtitle("Boxplot for Age at
diagnosis")+ coord_cartesian(ylim = c(0, 84)) + theme(plot.title = element_text(hjust = 0.5))
```

Variable – **diag**

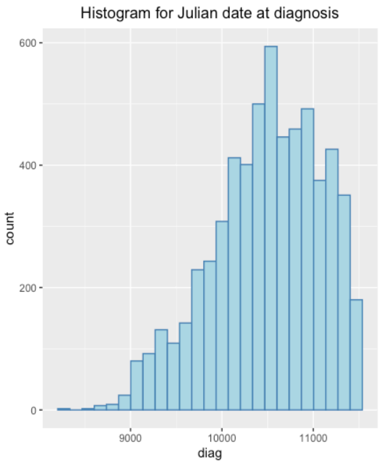


Figure 7

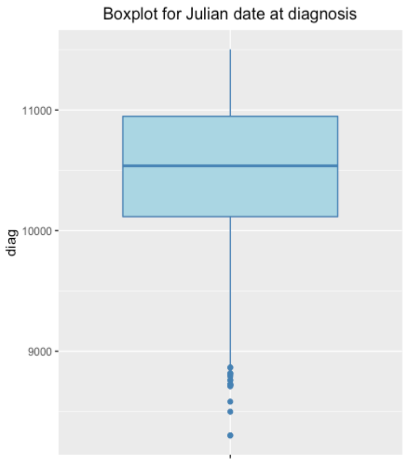


Figure 8

From **Figure 7** the histogram for the Julian date at diagnosis appeared non-symmetrical and it exhibited a left skewed behaviour. Furthermore, it could be noticed that the tail of the distribution on the left hand side is longer than on the right hand side. This meant at earlier Julian date there were less patients diagnosed with AIDS, many more were diagnosed at later Julian date. The reflection of the left skewed distribution can also be confirmed by the value of the mean being smaller than the median from **Table 5.** The boxplot in **Figure 8** visualized the presence of outliers. The exact numbers of outliers were computed using the **find_outliers(diag)**, resulting in 20 outliers. The median being always less affected by skewed data and outliers than the mean, was considered a measure of central tendency; The median Julian date at diagnosis was 10537. The code used to gain statistics and plots were similar to those for variable **age**, not written here due to shortage of space.

Association between pairs

Chi-squared test for independence

The chi-squared statistic:

$$\chi^2 = \sum_{k=1}^{K} \frac{(O_k - E_k)^2}{E_k} \qquad \text{Equation 1}$$

Some of underlying assumptions included: data were obtained through random selection; data in the cells should be in frequencies (or counts) rather than percentage; two variables present and both measured as categories; valid only if all frequencies in all cells of the contingency tables are above 5.

Test for variable status and transmission category (T.categ)

In order conduct a chi-squared test for independence the following hypothesis statements were set; $\boldsymbol{H_0}$ **:** No association between **status** and **T.categ** (they are independent) $\boldsymbol{H_1}$ **:** There is an association (they are not independent). **Data type:** Both variables were categorical variables and they met the underlying assumptions for the chi-squared test mentioned above. From **Table below** it can be seen that the data was a 2 x 8 table (the last rows and columns indicated simply row and column marginal) of cross-tabulated frequencies, and these consisted of the observed values $\boldsymbol{O_k}$.

status	**blood**	**haem**	**het**	**hs**	**hsid**	**id**	**mother**	**other**	**SUM**
A	58	33	59	2116	70	69	9	67	2481
D	129	56	43	3101	98	39	6	61	3533
SUM	187	89	102	5217	168	108	15	128	6014

The null hypothesis was that there was no association between the two categorical variables. The expected frequencies for that case was required to be computed; the values for the expected frequencies was tabulated in **Table below**

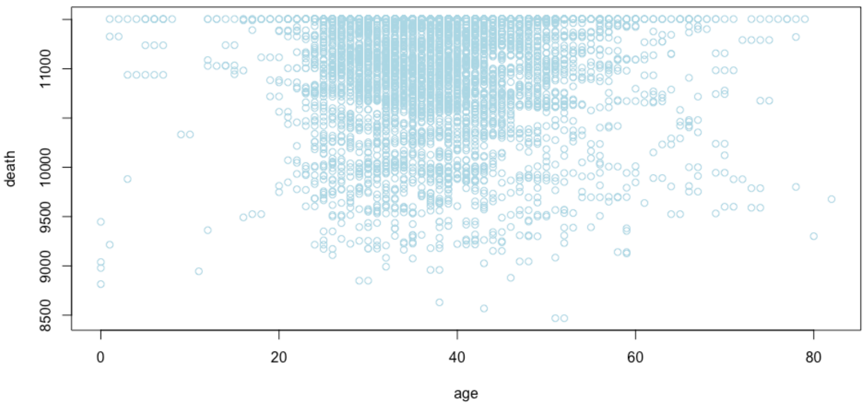
	blood	**haem**	**het**	**hs**	**hsid**	**id**	**mother**	**other**
Expected A	77.145	36.716	42.079	2152.208	69.306	44.554	6.188	52.805
Expected D	109.855	52.284	59.921	3064.792	98.693	63.446	8.812	75.195

In order to get the test statistic and compared to a critical value, **Equation 1** was used. The critical value for the chi-square statistics was determined by the level of significance and the degrees of freedom. Test was carried out at 2 level of significance: 5% and 1%. The degrees of freedom were calculated using the formula ***df* = (n_rows -1)*(n_cols -1) = (2-1) *(8-1) = 7.** From the computation of the chi-squared test statistic, it resulted to be **χ 2 = 52.862**.
The critical values of the χ 2 distribution at 5% significance and 1% significance were calculated to be **14.06714** and **18.47531** respectively. Hence the result was significant at both levels and it was possible to conclude that there was strong evidence against the null hypothesis. Therefore, it was possible to say that an association was present between the patients' status (Alive or dead) at the end of observation and the reported transmission category.

Sex and State

In order conduct a chi-squared test for independence the following hypothesis statements were set H_0 **:** No association between **sex** and **state** (they are independent) H_1 **:** There is an association (they are not independent). Similar procedures were carried out again. Test was carried out at 2 level of significance: 5% and 1%. The degrees of freedom this time were 3 as the cross-tabulation was a 2 by 4 table. From the computation of the chi-squared test statistic, it resulted to be **χ 2 = 19.583**. The critical values of the χ 2 distribution at 5% significance and 1% significance were calculated to be **7.814728**and **11.34487** respectively. Therefore, the result of test statistic was significant at both levels. It was possible to conclude that there's evidence against the null hypothesis. An association was present between the sex of the patient and the state of origin.

Age and Death



Multiple Logistic Regression

Multiple logistic regression model takes in multiple regressors, in the format:

$$p(x, \dots, x_n) \; \frac{e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n}}{1 + e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n}}$$

and the logit of $p(x)$ is

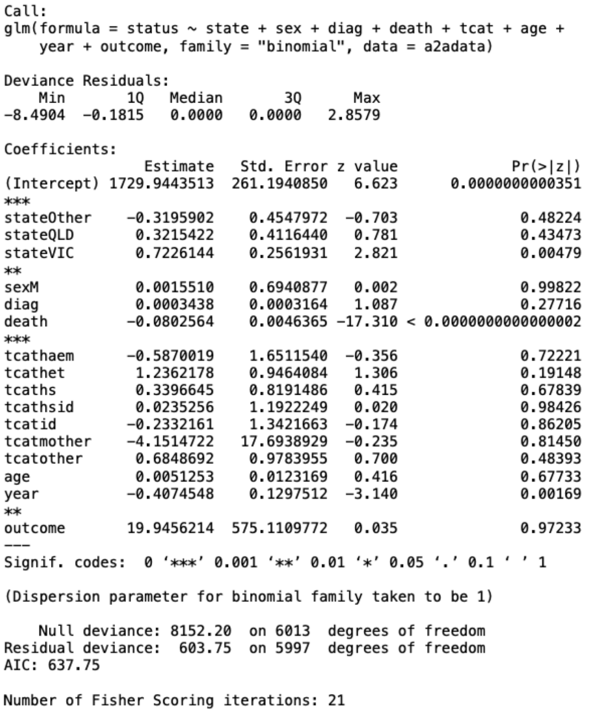
$$logit\big(p(x)\big) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n$$

where β_0 is the intercept of the logistic regression and $\beta_1, \dots, \beta_n$ are the coefficients given in the model and *n* the total number of covariates. Multiple logistic regression is an extension of the linear regression. Whereas linear regression tries to explain correlation between variables, logistic regression tries to estimate the odds of an event to take place or not. Multiple logistic regression can be applied to model associations with multiple covariates, making it a multi-dimensional model.

Logistic model 0: The full model

The probability of survival (status = Alive or Dead) can be modelled in terms of all the explanatory variables. **Logistic model 0** includes all the regressor variables as they are thought to explain the status of a patient and they are included as covariates. To put categorical variables into the regression, dummy variables are used; for example, **SEX** taking values 0 for Female and 1 for Male.

```
logistic.model0 <-glm(status ~ state+sex+diag+death+tcat+age+year+outcome,
data = a2adata, family = "binomial"); summary(logistic.model0)
```



```
Call:
glm(formula = status ~ state + sex + diag + death + tcat + age +
    year + outcome, family = "binomial", data = a2adata)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-8.4904  -0.1815   0.0000   0.0000   2.8579

Coefficients:
                Estimate   Std. Error z value             Pr(>|z|)
(Intercept) 1729.9443513  261.1940850   6.623      0.0000000000351
***
stateOther    -0.3195902    0.4547972  -0.703              0.48224
stateQLD       0.3215422    0.4116440   0.781              0.43473
stateVIC       0.7226144    0.2561931   2.821              0.00479
**
sexM           0.0015510    0.6940877   0.002              0.99822
diag           0.0003438    0.0003164   1.087              0.27716
death         -0.0802564    0.0046365 -17.310 < 0.0000000000000002
***
tcathaem      -0.5870019    1.6511540  -0.356              0.72221
tcathet        1.2362178    0.9464084   1.306              0.19148
tcaths         0.3396645    0.8191486   0.415              0.67839
tcathsid       0.0235256    1.1922249   0.020              0.98426
tcatid        -0.2332161    1.3421663  -0.174              0.86205
tcatmother    -4.1514722   17.6938929  -0.235              0.81450
tcatother      0.6848692    0.9783955   0.700              0.48393
age            0.0051253    0.0123169   0.416              0.67733
year          -0.4074548    0.1297512  -3.140              0.00169
**
outcome       19.9456214  575.1109772   0.035              0.97233
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 8152.20  on 6013  degrees of freedom
Residual deviance:  603.75  on 5997  degrees of freedom
AIC: 637.75

Number of Fisher Scoring iterations: 21
```

The multiple logistic regression model illustrates the following in the coefficients rows:

- **The first numerical row for coefficients**: **intercept** $\beta_0 = 1729.9$, representing the estimated value when all the other variables are equal to zero; the standard error (**Std. Error**) for the intercept value; the **z-value**, in short, the regression coefficient divided by its standard error; **Pr(>|z|),** representing the p-values from the Wald test for significance of the coefficients.
- **Consecutive rows:** The β **coefficients** of the variables; the **standard error** for specified coefficient value; **Pr(>|z|),** representing the p-values from the Wald test for significance of the coefficients.From the output of **Logistic model 0**, the minus two times the log-likelihood is called “Residual deviance”.

Due to the presence of non-linearity and data transformations needed in GLM it is not intuitive to compare fitted values with observations for the goodness of fit. Deviance statistic is used as a measure of the goodness of fit. Our interest lies in comparing nested models, meaning, comparing any proposed reduced model to a fully saturated model (extreme case, where there are as many parameters as observations). At $\alpha = 0.05$, it can be seen that from the p-values from the Wald test, there are many variables which are shown to be redundant in the full model, for any effect on dependent variable.

Logistic model 1: A reduced multiple logistic regression model

The reduced model **Logistic model 1** includes only significant covariates: year, death and state. Due to space limitations it was not possible to show the output. However, comparisons and summaries of results are discussed in a later part.

```
logistic.model1 <-glm(status ~ year+state+death, data = a2adata, family = "binomial");     summary(logistic.model1)
```

The reduced model led to a reduction of significance of the variable year. Therefore, the model was reduced further.

Logistic model 2: A further reduced multiple logistic regression model

The reduced model **Logistic model 2** includes only state and year covariates.

```
logistic.model2 <-glm(status ~ state+death, data = a2adata, family = "binomial");           summary(logistic.model2)
```

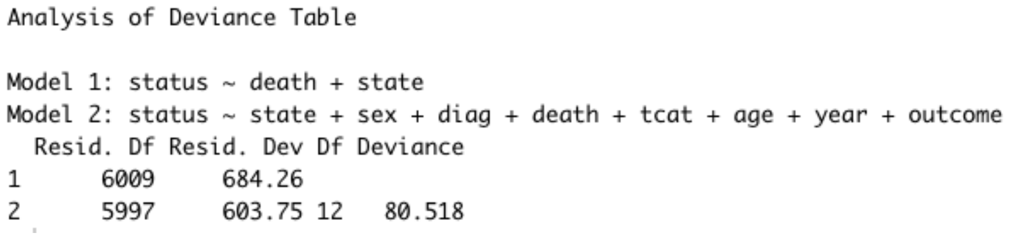
Due to all variables in this model being significant the model cannot be reduced further.

Likelihood Ratio Test (LRT)

The LRT test is a way to assess the goodness of fit and to choose between nested models.
From the full and reduced model above, it's possible to produce analysis of deviance table in RStudio to get the deviances of both models. Considering the full model

$$logit\left(p\left(x_1, x_2, \dots, x_p\right)\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_q x_q + \beta_{q+1} x_{q+1} + \cdots + \beta_p x_p$$

it would be liked to test the null hypothesis $H_0 : \beta_{q+1} = \cdot\cdot\cdot = \beta_p = 0$, against the alternative hypothesis, that at least one coefficient differs from zero.
In order to test $H_{0,}$ all one needs is the fit the full model and the reduced model and compare the respective deviances. As mentioned earlier, the results of the deviances can be read off the analysis of the deviance table, using `anova(logistic.model2, logistic.model0)`.



```
Analysis of Deviance Table

Model 1: status ~ death + state
Model 2: status ~ state + sex + diag + death + tcat + age + year + outcome
  Resid. Df Resid. Dev Df Deviance
1      6009     684.26
2      5997     603.75 12   80.518
```

Therefore, the LRT test statistic is $\chi\ 2\ = Deviance(reduced) - Deviance(full)$ which yields to $\chi\ 2 = 80.51$. Under $H_{0,}$ this test statistic has the degrees of freedom equal to the difference in the number of parameters between full and reduced model: *p – q = 8 - 2 = 6*.

For 1% significance, the critical value from a χ^2_2 distribution is 16.81. Hence we can reject the null hypothesis. Meaning that the full model fits better than the reduced model.

Analysis of covariate

We want to establish if a particular covariate has positive or negative effect on the outcome. We want to test the association between patient survival and gender.

$$LogOdds(Y = 1|male) - LogOdds(Y = 1|female) = \beta_2$$

In order to see if gender is important, hypothesis test is required to be se for β_2,

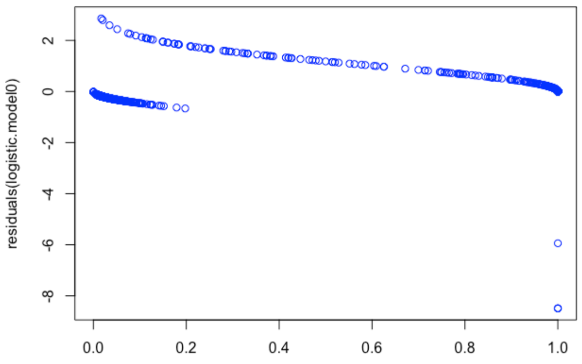
$$H_0 : \beta_2 = 0 \text{ v. } H_1 : \beta_2 \neq 0$$

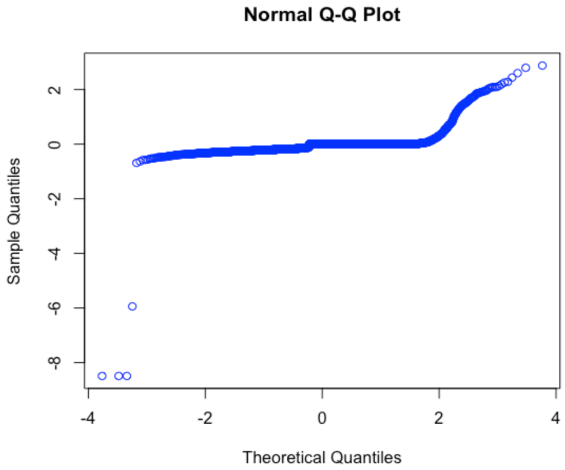
The value for the test statistic can be obtained from the output of the logistic.model0 above. The Wald test statistic is then $z = \frac{\widehat{\beta_2}-0}{SE(\widehat{\beta_2})} = \frac{0.0015510-0}{0.6940877} = 0.002234588$.

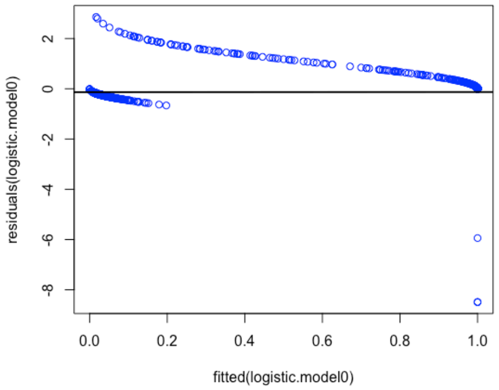
The critical value for 99% significance for the standard Normal *N(0,1)* distribution are ± 2.58. The test statistic is within the range therefore it is slightly insignificant.

Therefore, the probability of survival for patients diagnosed with Aids was not highly associated with gender. Using the test statistic above, it's possible to co construct confidence interval such as $\widehat{\beta_2} \pm z_{crit(\alpha)} * SE(\widehat{\beta_2})$. For a 99% confidence interval on the parameter $\widehat{\beta_2}$ is $0.0015510 \pm 2.58 * 0.6940877 = (1.792297, -1.789195)$. Hence it can be observed that the covariate has no effect on the outcome, thus, probability of survival for patients diagnosed with AIDS has no affect with the gender of the patient.

Residual Plots







Code used:

```
plot(fitted(logistic.model0),residuals(logistic.model0),
col="blue")
lines(lowess(predict(logistic.model0),residuals(logistic
.model0)),col="black",lwd=2)
res<-rstandard(logistic.model0)
qqnorm(res, col="blue")
```

The presence of two lines of plots is normal in the first plot, as we are predicting the probability of for a variable taking values 0 and 1. Points are somewhat smooth curve, indicating non-linearity. The residual measure the variation in the response variable not explained by the regression model. As it can be seen from the plots, there are present a number of outliers. The Q-Q plot isn't so much of a smooth S shape like.

Weakness of this analysis

Logistic regression is a widely used model for it's efficiency which doesn't require much computational resources. Some of the disadvantages is that non-linear problems cannot be solved with logistic regression.

REFERENCES:

McCullagh, P. and Nelder, J. A. (1989). Generalized Linear Models. Chapman & Hall, London

Draper, NR, & Smith, H 1998, Applied Regression Analysis, John Wiley & Sons, Incorporated, New York.

Lecture Notes: https://blackboard.brunel.ac.uk/bbcswebdav/pid-1122592-dt-content-rid-5841118_1/courses/C.CS5606.A.2019-0.TRM1/test%282%29.pdf

https://blackboard.brunel.ac.uk/bbcswebdav/pid-1127356-dt-content-rid-5859701_1/courses/C.CS5606.A.2019-0.TRM1/Lecture%20Slides%20Week%208-9%281%29.pdf

https://blackboard.brunel.ac.uk/bbcswebdav/pid-1124761-dt-content-rid-5850203_1/courses/C.CS5606.A.2019-0.TRM1/glm.pdf