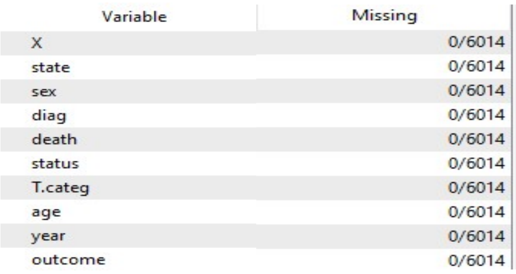
Task 1

The data set that will be used in this report contains 6014 observations and 10 variables. The R command used to extract this result is: dim(`Aids2ann.(1)`) 6014   10 .

The first step before the further analysis of the dataset is to identify if there are any missing data. It appears that there are not any missing data. AmeliaView, an R package has been used to identify if there are any missing data and as shown in the figure on the right the result is zero.



Variable	Missing
X	0/6014
state	0/6014
sex	0/6014
diag	0/6014
death	0/6014
status	0/6014
T.categ	0/6014
age	0/6014
year	0/6014
outcome	0/6014

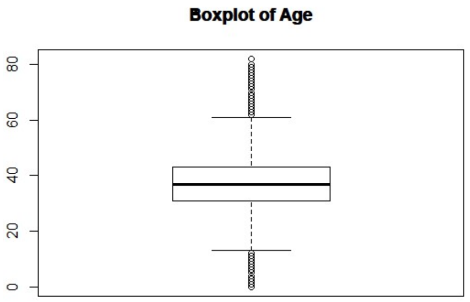
Variable "Age"

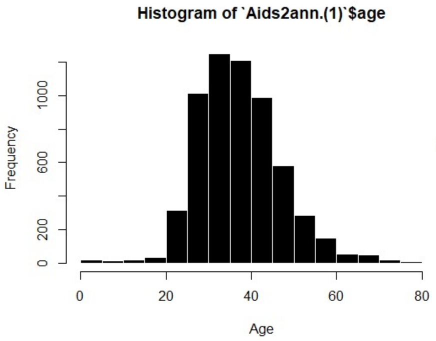
```
range(`Aids2ann.(1)`$age)          [1]  0 82
mean(`Aids2ann.(1)`$age)           [1] 37.74327
median(`Aids2ann.(1)`$age)         [1] 37
quantile(`Aids2ann.(1)`$age)        0%  25%  50%  75% 100%
                                     0   31   37   43   82
sd(`Aids2ann.(1)`$age)             [1] 9.776273
```

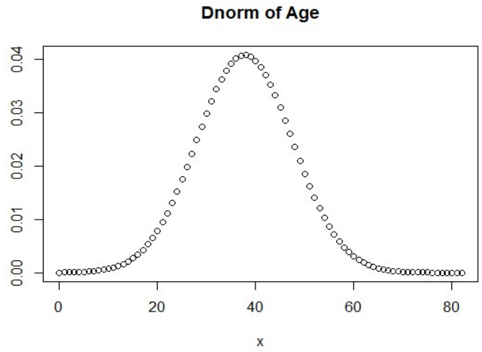
It is visible from the data that the range of age is from newborn of 0 age to the oldest person 82. The values of the mean and the median are really close which indicates that the data are normally distributed. Although judging by the value of the standard deviation and the boxplot below, we can observe that our data are spread out with many outliers. Having many outliers indicates that some data differ significantly between them, something that could cause issues in the data analysis. Someone can observe from the histogram of Age that the distribution is quite normal as well as from the function d norm that has been used in the right figure.

R Commands for the below graphs.

```
hist(`Aids2ann.(1)`$age,col="black",border="white",xlab = "Age",ylab = "frequency")
boxplot(`Aids2ann.(1)`$age)
x <- seq(0,82)
y <-dnorm(x,mean=37.74,sd=9.77)
plot(x,y)
title("Dnorm of Age")
```







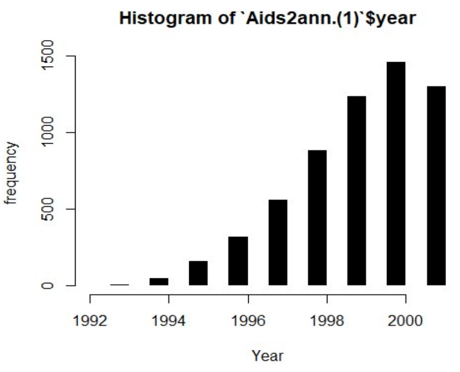
Variable "Year"

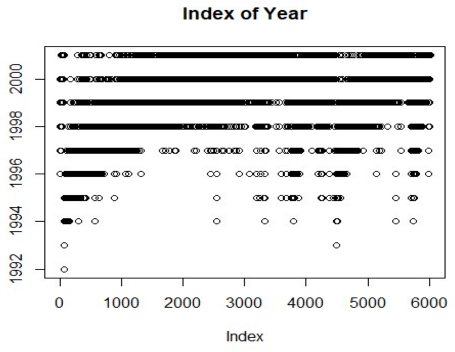
```
range(`Aids2ann.(1)`$year)      [1] 1992 2001
mean(`Aids2ann.(1)`$year)       [1] 1999.023
median(`Aids2ann.(1)`$year)     [1] 1999
quantile(`Aids2ann.(1)`$year)   0%  25%  50%  75% 100%
                              1992 1998 1999 2000 2001
sd(`Aids2ann.(1)`$year)         [1] 1.677312
```

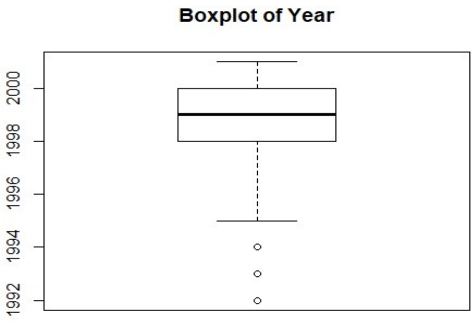
It is observed from the histogram below that the curve has a left skewed behavior with the vast amount of data coming from the years between 1997 and 2001. The boxplot shows that the outliers are not widely spread out as well as the   mean and median values are close. An interesting clue is that the first quantile begins in 1998, 6 years after the first year of the research. Finally, the index of year shows graphically the increase of incidents per year, a constant increase per year is visible except for the last year where there is a decline. The index of year shows how gradually the reported incidents have grown throughout the examined years.

R Commands for the below graphs.

```
hist(`Aids2ann.(1)`$year,col="black",border="white",xlab = "Year",ylab = "frequency")
plot(`Aids2ann.(1)`$year)
title("Index of Year")
boxplot(`Aids2ann.(1)`$year)
title("Boxplot of Year")
```







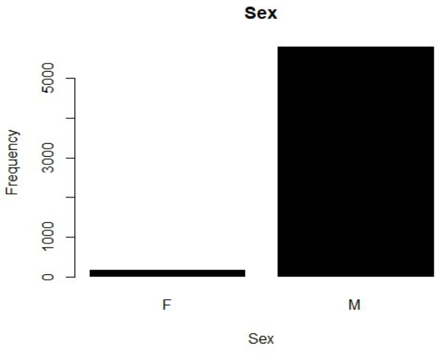
Variable "Sex, Status and State"

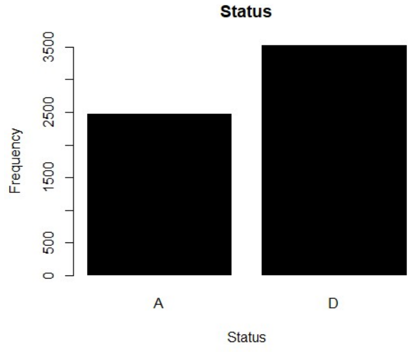
```
table(`Aids2ann.(1)`$sex)
   F    M 
 202 5812 
table(`Aids2ann.(1)`$sex,`Aids2ann.(1)`$status)
       A    D
  F   97  105
  M 2384 3428
summary(`Aids2ann.(1)`$state)
  NSW Other   QLD   VIC 
 3775   544   446  1249 

table(`Aids2ann.(1)`$status)
  A    D 
2481 3533 
```

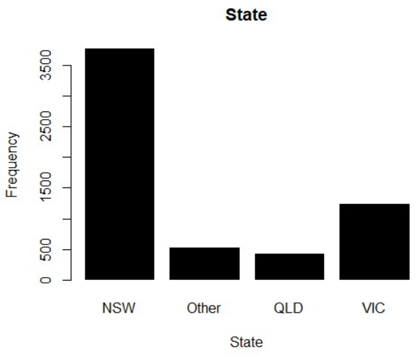
Below three categorical variables have been analyzed, sex, state and status. Most patients in this data set are men covering the stunning 96.6% of the data set, women are significantly lower to 3.4%. The difference seems huge and a further investigation might need to be conducted to check the validity of the collection of the data. The plot for the status shows that a bigger percentage of people have died at the end of the observation as well as the state barplot indicates that the biggest proportion of incidents were reported in the North South Wales state with state Victoria coming second.

R Commands for the below graphs.

```
plot(`Aids2ann.(1)`$sex,col="black",border="white",xlab = "Sex",ylab = "Frequency")
title("State")
plot(`Aids2ann.(1)`$status,col="black",border="white",xlab = "Status",ylab = "Frequency")
title("Status")
plot(`Aids2ann.(1)`$sex,col="black",border="white",xlab = "Sex",ylab = "Frequency")
title("Sex")
```







Possible Issues with the Data Collection

Since Australia is a large developed country and taking into consideration that the virus of HIV is well-known, we assumed that the data were collected accurately. Of course, there will be cases where some death notifications will arrive with delay manipulating slightly the results. For examples, we can observe that 29 patients diagnosed with HIV after their death. The below code helped us to reach to this conclusion.

```
sum(`Aids2ann.(1)`$death>`Aids2ann.(1)`$diag)
[1] 5985
```

```
6014-5985
[1] 29
```

Furthermore, there is a huge disproportion of reported incidents between the states of New South Wales and Victoria which have approximately the same population, 8m and 6.6m respectively (Abs.gov.au, 2020,current data, past data might differ slightly but not significantly), something that reveals the contamination might began from the state of New South Wales and then spread out to the rest of the states or New South Wales state was the one who became first aware of the disease. It is unknown when the states became fully aware of the problem. Another possible issue could be that female patient did not want to reveal that they are homosexuals and contaminated with HIV since back in 90' there was a big prejudice about homosexuality and the disease of HIV itself.

```
table(`Aids2ann.(1)`$T.categ)

 blood    haem     het      hs    hsid      id mother  other 
   187      89     102    5217     168     108     15    128 
```

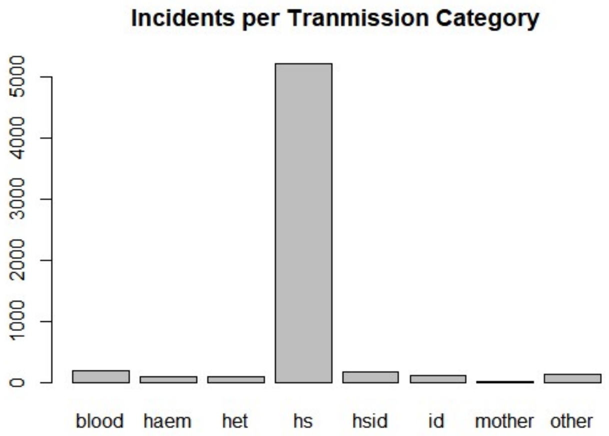
Homosexuals cover the 87% of the population indicating the burst of incidents among this group of people although the huge difference between male and female patients increases the assumption that many women did not come out or that the disease could not spread out between homosexual women because of the nature of the physical contact.

The below table comes to enhance the above assumptions.

```
table(`Aids2ann.(1)`$T.categ=="het")

FALSE  TRUE 
 5912   102 
```

```
table(`Aids2ann.(1)`$T.categ=="het",`Aids2ann.(1
)`$sex)
       
          F    M
  FALSE 147 5765
  TRUE   55   47
```

```
table(`Aids2ann.(1)`$T.categ=="hs",`Aids2ann.(1)
`$sex)
       
           F    M
  FALSE  201  596
  TRUE     1 5216
```



Surprisingly there are more women when it comes to heterosexual patients. On the other hand, when it comes to homosexual patients, only one woman has been reported in the data set. Additionally, it is difficult not to take into consideration the human factor especially when back in the days there was not an automated system where you could report every detail. Since we do not know the way the data were collected, we cannot be totally sure that there are not any misfeeds. It is also unclear at which point the disease was passed to the patient to compare it with the death date. Diagnosis can be reported at a later stage where the possibilities of survival are low, given that the disease has spread through the patient's immune system. Another potential issue could be the nomadic people who move around often and they might have been counted twice. Also, people that did not want to be reported as AIDS patients to avoid the stigma may be missed from the data set.

Task 2

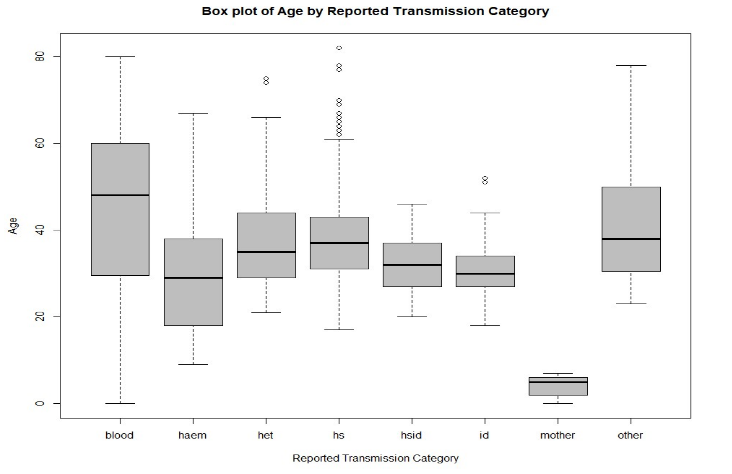
Pairwise association between Age and Reported Transmission Category

The graph 5 shows the relationship between the age and the transmission category. Excluding the category “mother” where the patient obviously got contaminated in birth or in the very early year of their lives and “blood” where the median falls between the age group 40-60, for the rest categories the median falls into the age group 30-40 indicating this age group as a high risk one. The boxplots below show the different skewness of each Transmission category by Age. For examples, “hs”,”hsid”,”id” seem to have normal distributions with the median in the middle of the box. The boxplot of mother and blood seem to have a negative skewed distribution since the median is closer to the upper level of the box. On the other hand, the het and id boxplots seem to have a slight positive skewed distribution since the median is closer to lower level of the box.

Graph 1 : Box plot of Age by Reported Transmission Category

R Command:

```
boxplot(`Aids2ann.(1)`$year~`Aids2ann.(1)`$T.categ,col = "grey", main="Box plot of Year by Reported Transmission Category",xlab="Reported Transmission Category", ylab = "Year")
boxplot(`Aids2ann.(1)`$age~`Aids2ann.(1)`$T.categ,col = "grey",main="Box plot of Age by Reported Transmission Category",xlab="Reported Transmission Category", ylab ="Age")
```



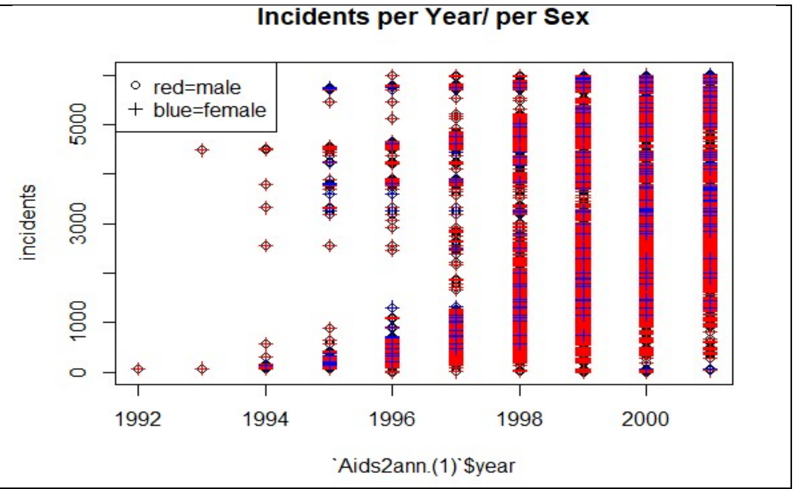
Pairwise association between Incidents per Year/per Sex

For the below graph 7 the variable incidents could be translated as the variable of diagnosis from the point that every diagnosis with a specific date represents a unique incident(in some cases there are more than one incident per specific date, something that has been counted too. The factor X in the data represents the incidents too). It is observed that the incidents/diagnosis have been radically boomed from 1997 and after as well as someone could observe that in the year 1995 and 1996 the population of males and females were approximately at the same level, the following years male is the dominant reported category.

Graph 2 : Incidents per Year/per Sex

R Command:

```
plot(`Aids2ann.(1)`$year,incidents)
index <- `Aids2ann.(1)`$sex=="M"
points(`Aids2ann.(1)`$year[index],incidents[index],col="red",pch=3)
index <- `Aids2ann.(1)`$sex=="F"
points(`Aids2ann.(1)`$year[index],incidents[index],col="blue",pch=3)
legend("topleft", legend=c("red=male","blue=female"), pch=c(1,3)) title("Incidents per Year/ per Sex")
```



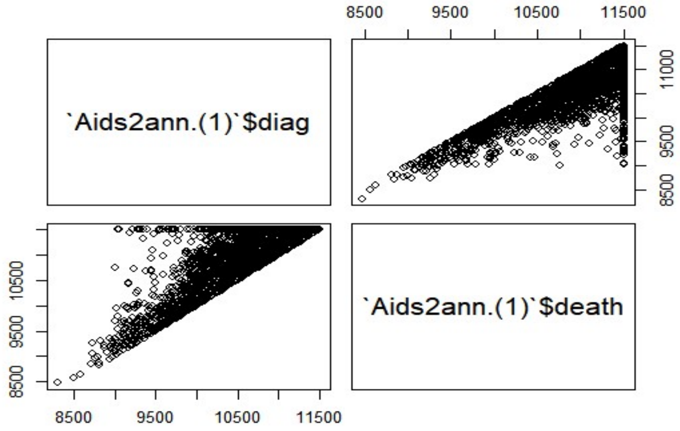
Pairwise association between Death and Diagnosis

The graph 8 shows the pairwise association between death and diagnosis. A correlation test has been conducted showing a strong correlation between the variable death and diagnosis with the test giving the value of 0.71.

Graph 8 : Pairwise association between Death and Diagnosis

R command : `pairs(`Aids2ann.(1)`$diag~`Aids2ann.(1)`$death)`

```
cor(`Aids2ann.(1)`$death,`Aids2ann.(1)`$diag)=0.7066221

cor.test(`Aids2ann.(1)`$death,`Aids2ann.(1)`$diag)
```



```
Pearson's product-moment correlation
data:  `Aids2ann.(1)`$death and `Aids2ann.(1)`$diagt = 77.431, df = 6012, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:

0.6937375 0.7190546

sample estimates:

cor 0.7066221
```

Pairwise association between Status and State – Chi-Square Test

```
chisq.test(`Aids2ann.(1)`$state,`Aids2ann.(1)`$status)

        Pearson's Chi-squared test

data:  `Aids2ann.(1)`$state and `Aids2ann.(1)`$status
X-squared = 18.918, df = 3, p-value = 0.0002843
```

The null hypothesis is: Status is not associated with State

The alternative hypothesis is: Status is associated with State

The chi-square test has been used above to compare two categorical variables. Before we apply the chi-square test, we need to check that all frequencies in all cells are above 5. Since the sample is big we can apply the chi-square where we find that there is a strong association between the status and the state, having a first clue that the survival rates differ throughout the states although the chi-square test cannot tell us the size or the direction of the relationship. We have concluded that the relationship is strong since the p-value is close to zero and if we consider as alpha p-value = 0.05 then we reject the null hypothesis.

Underlying Assumptions

Before we apply the chi-square test, some assumptions need to be stated and to check if these assumptions are satisfied so we can proceed with the implementation of the test. These assumptions are as below:

- We need to make sure that we have two variables which are measured as categorical variables. In our case, we tested the two categorical variables state and status in order to satisfy this assumption.
- Secondly, all frequencies in all cells are above 5. In our case the data set is big, so our assumption automatically is satisfied.
- An assumption that a random sample have been collected for the two categorical variables in the data set.
- The two categorical variables are independent from each other.

TASK3

Before the use of regression, a check of the date set is required. The factor outcome is a binary dependent variable, so the conclusion is that the logistic regression needs to be applied. Plugging different combinations of independent variables into the model, initially we conclude to the below model, called model 1. The model includes the transmitting category “heterosexual”, the year 2000 as well as the factor state and the factor diagnosis. The factors sex and status have not been included since there aren’t any significant association. The factor death despite that the logistic regression shows significant association has not been included since the death and the outcome are two variables with the same result(death) but under different conditions so the model shows association but it is known that association does not imply causation.

DATA ORIGINAL – Model 1

```
dataoriginal <- glm(outcome~factor(T.categ=="het")+factor(year=="2000")+diag+
factor(state),family = binomial,`Aids2ann.(1)`)
summary(dataoriginal)
```

```
Call:
glm(formula = outcome ~ factor(T.categ == "het") + factor(year ==
    "2000") + diag + factor(state), family = binomial, data = `Aids2ann.(1)`)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5437  -0.8468  -0.7043   1.2941   2.2282

Coefficients:
                               Estimate Std. Error z value Pr(>|z|)
(Intercept)                   8.049e+00  5.424e-01  14.839  < 2e-16 ***
factor(T.categ == "het")TRUE -7.539e-01  2.728e-01  -2.764  0.00572 **
factor(year == "2000")TRUE    6.493e-01  7.116e-02   9.125  < 2e-16 ***
diag                         -8.710e-04  5.283e-05 -16.486  < 2e-16 ***
factor(state)Other           -1.216e-01  1.069e-01  -1.138  0.25532
factor(state)QLD              2.542e-01  1.099e-01   2.313  0.02075 *
factor(state)VIC              9.454e-03  7.418e-02   0.127  0.89860
---
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: 6956.0  on 6007  degrees of freedom
AIC: 6970

Number of Fisher Scoring iterations: 4
```

It is observed from the output that the states in general are not significantly associated with the outcome except for Queensland where there is a slight significant association. Judging by this result it is understood that there is a higher risk for heterosexuals to die in year 2000 and some indications that the state of Queensland is riskier for the above combination than the other states.

Before searching in depth, we need to remove the insignificant factors from the model and run it again. Their p values of 0.255>0.05 and 0.898>0.05 respectively show no association. To obtain the best possible model the state of Queensland where there is a slight association has been removed. The R code and output are as below.

Data – Model 2

```
data <- glm(outcome~factor(T.categ=="het")+factor(year=="2000")+age+diag,
            family = binomial,`Aids2ann.(1)`)
summary(data)
```

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7674  -0.8439  -0.6953   1.2838   2.2172

Coefficients:
                               Estimate Std. Error z value Pr(>|z|)
(Intercept)                   7.438e+00  5.522e-01  13.470  < 2e-16 ***
factor(T.categ == "het")TRUE -7.493e-01  2.719e-01  -2.756  0.00585 **
factor(year == "2000")TRUE    6.494e-01  7.129e-02   9.109  < 2e-16 ***
age                           1.601e-02  2.960e-03   5.408 6.36e-08 ***
diag                         -8.698e-04  5.283e-05 -16.465  < 2e-16 ***
---
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: 6934.0  on 6009  degrees of freedom
AIC: 6944
```

The model 2 indicates that all the factors are significantly associated with the dependent variable “outcome” as the p-value result from all the independent variables is close to 0. In order to be sure that there is a difference between the two models we run the Likelihood Ratio Test (LRT) to assess the goodness of fit. The R code and output are as below.

```
lrtest(data,dataoriginal)
```

```
Likelihood ratio test

Model 1: outcome ~ factor(T.categ == "het") + factor(year == "2000") +
    age + diag
Model 2: outcome ~ factor(T.categ == "het") + factor(year == "2000") +
    diag + factor(state)
  #Df LogLik Df  Chisq Pr(>Chisq)
1   5  -3467
2   7  -3478  2 22.057  1.624e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The LRT test provides a p value close to 0 someone can confidently say that there is a significant difference between the two models. Testing this with hypothesis testing, the Ho: There is no difference between the two models vs H1: There is association between the two models. The p-value is 0.00001624 so the null hypothesis is rejected as there is a significant difference between the two models.

Furthermore, there is an alternative way to check the goodness of fit. Plugging the below code, R provides two results of variance. In general, variance is a measurement of the goodness of fit of the model. (Gelman, A. and Hill, J., 2006). Below we have coded in R the equation.

The R code and output are as below.

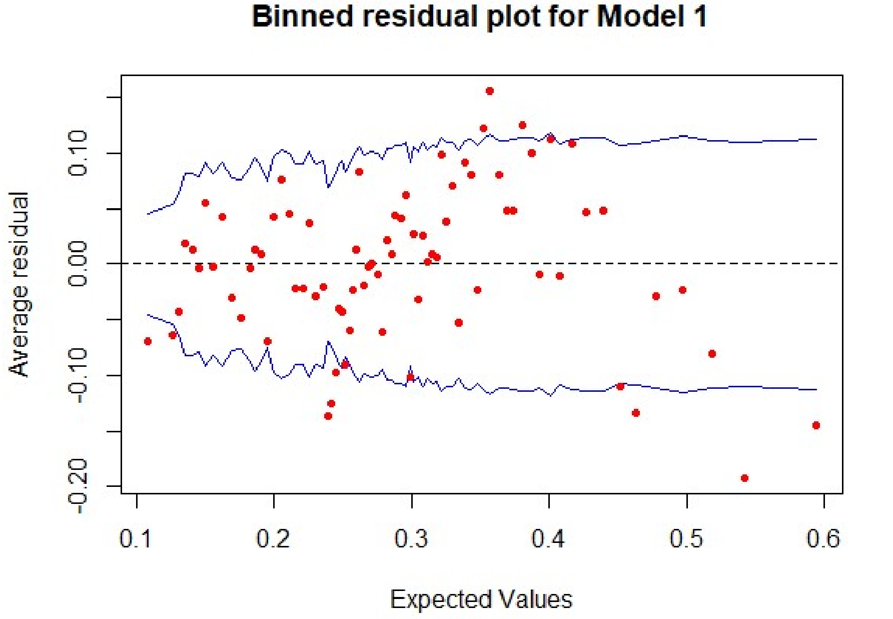
```
  x2= 2*(logLik(data)-logLik(dataoriginal)
  as.numeric(x2)
  [1] 22.05663
  pval=1-pchisq(x2,2)
  as.numeric(pval)
  [1] 1.623543e-05
```

Doing the math with the command as.numeric(x2) we will get the value of 22.057 In the distribution graph of the deviance someone could see that the value 21.057 is on the x axis and the area under the curve for this value in y axis is close to zero explaining that the model 1 and model 2 are significantly different. The p value is given by the command as.numeric(pval).

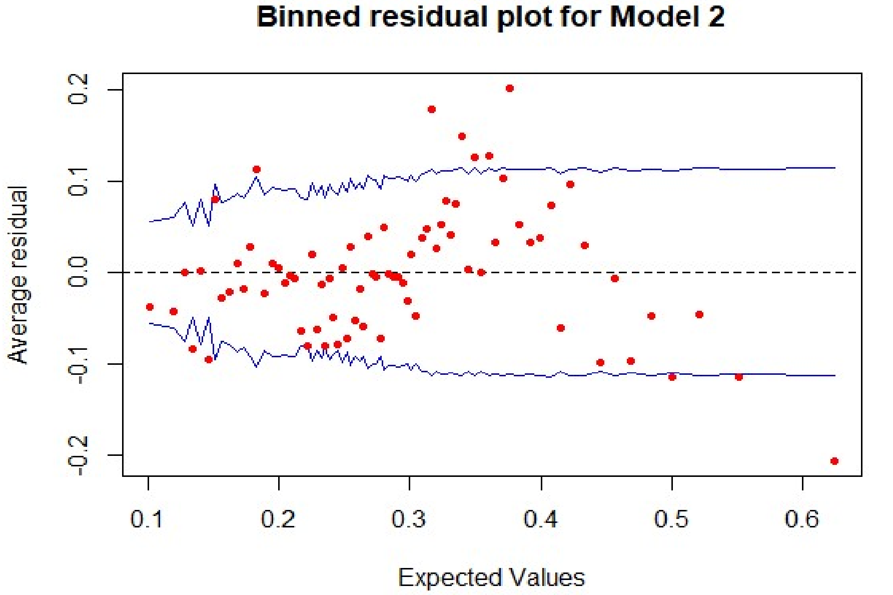
Model 1 vs Model 2 = p-value 0.00001623543

Binned residual plot for Models 1 and 2

```
binnedplot(fitted(dataoriginal),
residuals(dataoriginal, type = "r
esponse"),
nclass = NULL,
xlab = "Expected Values",
ylab = "Average residual",
main = "Binned residual plot for
Model 1",
cex.pts = 0.7,
col.pts = 2,
col.int = "blue")
```



```
binnedplot(fitted(data),
residuals(data, type = "response"
),
nclass = NULL,
xlab = "Expected Values",
ylab = "Average residual",
main = "Binned residual plot for
Model 2",
cex.pts = 0.7,
col.pts = 2,
col.int = "blue")
```



When the logistic regression is used instead of linear regression which is used in linear models, the deviance instead of the standard deviation is used to calculate how well the data fit the model. Deviance is a measure of error. (Gelman and Hill, 2006). Binned residual plots can be used to assess both the overall fit of regression modes for binary outcomes as well as the inclusion of continues variables. For a model to be correct, approximately 95% of the points are expected to be within the confidence limits. (Kasza, 2015). All the models above represent the residual binned plot of models 1 and 2 and we can see gradually as the models progress from model 1 to model 2 the residual outside of the confidence limits are less, coming to the model 2 where the vast majority of the points are included within the confidence limits. Calculating the deviance, the difference is obvious. The model 2 has lower deviance from the model 1.

```
deviance(dataoriginal)
[1] 6956.037
deviance(data)
[1] 6933.98
```

Confidence intervals on parameters

```
confint(data)
```

```
                                    2.5 %        97.5 %
(Intercept)                      6.3597451168  8.5246946560
factor(T.categ == "het")TRUE   -1.3145840458 -0.2422603324
factor(year == "2000")TRUE      0.5096111539  0.7891164765
```

```
age                                  0.0102105120  0.0218142636
diag                                -0.0009738154 -0.0007667032
```

The R output above shows two parameters with negative and two parameters with positive effect on the outcome. The factor “heterosexual” and the diagnosis have a negative effect on the outcome while the year 2000 and the age have positive effect. From the output we can once again confirm the significant association between the independent variables with the outcome since none of these variables include the value 0. If a parameter includes the value 0 between its confidence interval’s limits, means that this parameter is not significant and must be removed from the model.

Interpretation of the results

The model 1 has been reduced taking out the factor state which gave us the model 2 where all the independent variables are significantly associated with the dependent variable outcome. In the model two the p-values of the independent variables are below 0.05 which shows strong association. To compare if the two models have a significant difference between them as well as the goodness of fit of the model, we applied the likelihood ratio test which confirmed the difference between the two models by giving the p-value of 0.00001623543. We checked the validity of the difference with a more manual way too, by applying the code x2= 2*(logLik(data)-logLik(dataoriginal) which gave us the same p-value confirming the first attempt. Additionally, the result of deviance determines which test is better to use, as the reduction of variables that led to model 2 has decreased the variance but increased the significant levels among the independent variables.

Furthermore, the confidence intervals determine the significance of the variables used in the model since none of them contains the value 0 within their limits.The binned residual plots have been used to show how well the data fit in model 1 and model 2. The reduced model 2 included more binned data within its confidence interval’s limits providing the evidence of the better fit.

Analyzing the further the results of model 2, it is noticed that there is a high risk of a heterosexual to die in 2000 given the factors of age and diagnosis as very significant. The model 1 indicates a slight association of the above but adds one more significant factor of state Queensland as the riskier state for the above combination. The report can confidently represent these results as the n=samples size of the data is quite big.

Limitations

It is known that there are several ways to calculate the transformation of the data although the logistic regression is the most common.

We must remember that logistic regression gives odds for each predictor. The odds differ from the risk, and maybe the odds are high, but the risk is low. (Ranganathan, Aggarwal and Pramesh, 2015).

The independent variables must not be highly correlated among them because this might create an issue with estimation. For example, the independent variables of age and diagnosis are included in the model but the correlation between them is not high. Below is the R output.

```
cor(`Aids2ann.(1)`$age,`Aids2ann.(1)`$diag) [1] -0.0006702869
```

In logistic regression, the model must fit one dependent variable and all the others must be independent variables. If the model included two dependent variables, then the level of significance that the result will show is misleading.

Abbreviations

NSW :	**New South Wales**
VIC :	**Victoria**
QLD:	**Queensland**
T.categ:	**Transmission Category**
Haem :	**Haemophilia**
Het :	**Heterosexual**
Hs :	**Homosexual**
Id:	**Drug user**
Hsid:	**Heterosexual drug user**

References

Gelman, A. and Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models*. Cambridge: Cambridge University Press, pp.97-99.

Kasza, J. (2015). Stata Tip 125: Binned Residual Plots for Assessing the Fit of Regression Models for Binary Outcomes. *The Stata Journal: Promoting communications on statistics and Stata*, 15(2), pp.599-604.

Abs.gov.au. (2020). *3101.0 - Australian Demographic Statistics, Jun 2019*. [online] Available at: https://www.abs.gov.au/AUSSTATS/abs@.nsf/mf/3101.0.

Ranganathan, P., Aggarwal, R. and Pramesh, C. (2015). Common pitfalls in statistical analysis: Odds versus risk. *Perspectives in Clinical Research*, 6(4), p.222.

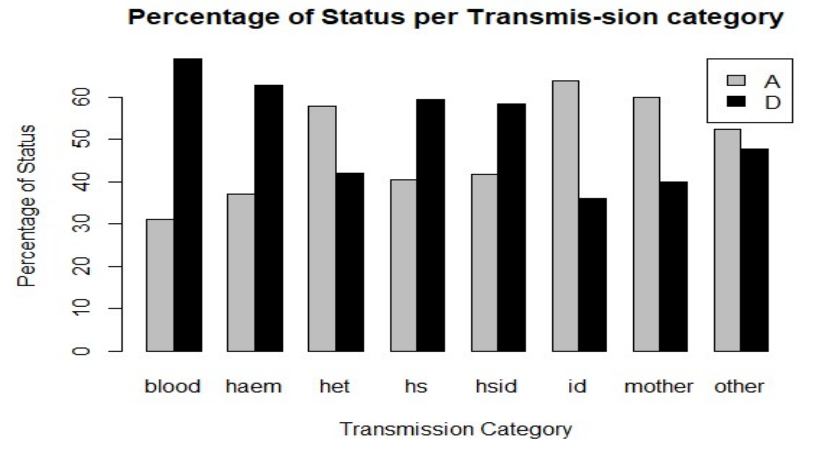
Miller, R. and Siegmund, D. (1982). Maximally Selected Chi Square Statistics. *Biometrics*, 38(4), p.1011.

Appendices

Graph 3 : Percentage of Status per Transmission category

R Command:

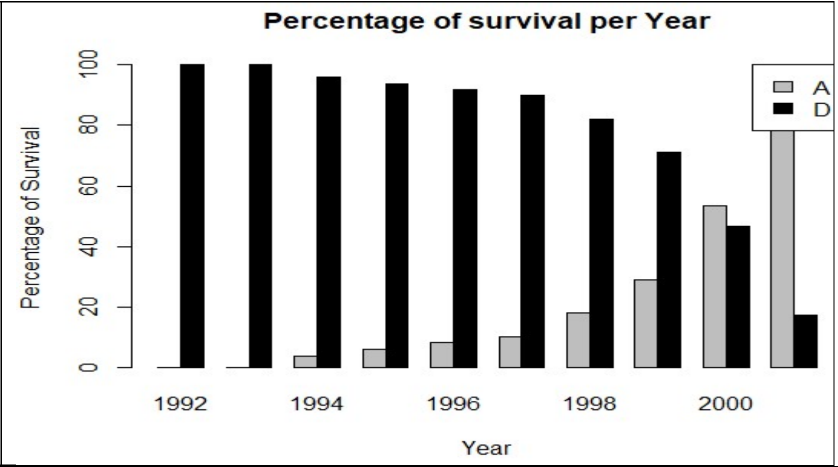
```
table <- table(`Aids2ann.(1)`$status,`Aids2ann.(1)`$T.categ)
barplot(prop.table(table,2)*100,xlab='Transmission
Category',ylab='Percentage of Status',
beside=T,col=c("grey","black"),
legend=rownames(table), args.legend = list(x = "topright"))
title("Percentage of Status per Transmis-sion category
")
```



Graph 4 : Percentage of survival per Year

R command :

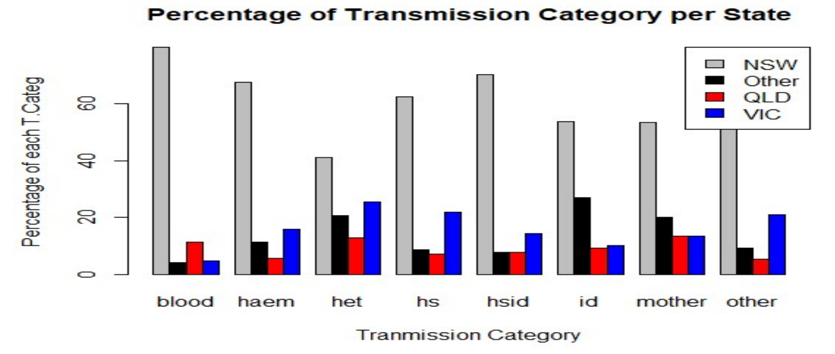
```
table <- table(`Aids2ann.(1)`$status,`Aids2ann.(1)`$year)
barplot(prop.table(table,2)*100,xlab='Year',ylab='Percentage of Survival',beside=T,col=c("gray","black"),
+ legend=rownames(table), args.legend = list(x = "topright"))
title("Percentage of survival per Year")
```



Graph 5: Percentage of T. Category per State

R command :

```
state <- table(`Aids2ann.(1)`$state,`Aids2ann.(1)`$T.categ)
barplot(prop.table(state,2)*100,xlab='Tranmission Category',ylab='Percentage of each T.Categ',beside=T,col=c("gray","black","red","blue"),legend=rownames(state), args.legend = list(x = "topright"))
title("Percentage of Transmission Category per State")
```


Percentage of Transmission Category per State
Percentage of each T.Categ
NSW
Other
QLD
VIC
blood
haem
het
hs
hsid
id
mother
other
Tranmission Category