CS5606 QUANTITATIVE DATA ANALYSIS: ASSESSMENT/COURSEWORK FOR 2019/20

Emma Luk
1830215@brunel.ac.uk

Task 1. Explore the data. Plot and produce summary statistics to identify the key characteristics of the data (for some of the variables listed above) and produce a report of your findings. 5 - 10 tables or figures are expected accompanied by a description of your main findings. The topics that you might choose to discuss include: possible issues with the data collection, identification of possible outliers or mistakes in the data, role of missing data (if any) and distribution of the variables provided.

1.1 Data description & research question

It is important that we understand factors affecting the survival of people with AIDS. Its importance stems from the evolving definition of AIDS which has implications for defining and estimating the incubation distribution. (B D Ripley, P J Solomon, 1992). The first step, before any calculations or plotting of data, is to decide what type of data one is dealing with.

There are a number of typologies, **table 1.1.1 and 1.1.2**. describes and meaning the variables. The basic distinction is between quantitative variables (for which one asks, "how much?") and categorical variables (for which one asks, "what type?").

Categorical variables are grouped state of origin: New South Wales and Australian Capital Territory, Other, Queensland, and Victoria; male/female; alive/dead; Reported transmission category; and survived/ died. Numerical variables are diag, death, age and year. The diag and death are using Julian date of diagnosis.

Research questions:
The questions will be "What kinds of trends are in the data? What kinds of predictions are possible? What conclusions can we make?"

Numerical variable meanings:

Variable	Meaning
diag	**Julian date of diagnosis (the number of days since 1970-01-01)**
death	**Julian date of death or end of observation (the number of days since 1970-01-01)**
age	**Age (years) at diagnosis**
year	**The year of observation (normal calendar)**

Table 1.1.1 numerical variable meanings

Categorical variable meanings:

Variable	Meaning (Ripley and Solomon, 1992)
state	**Grouped state of origin**
NSW	New South Wales and Australian Capital Territory
Other	Other
QLD	Queensland
VIC	Victoria
sex	**Sex of patient**
F	Female
M	Male
status	**Alive or dead at the end of observation**
A	Alive
D	Dead
T.categ	**Reported transmission category**
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 or at risk of HIV infection
other	other or unknown
outcome	**'1' if the patient died in the year of observation specified in 'year', '0' if survived**
0	if survived
1	if the patient died in the year of observation specified in 'year

Table 1.1.2 categorical variable meanings

1.2 Data preparation and cleaning

This section explains data preparation and cleaning the dataset so it can be used effectively during an investigation. The Aids2ann dataset didn't need too much cleaning, this project needed to create separate column for survival days, which number of days he/she was alive after diagnosis.

The diag and death are using Julian date of diagnosis (**Figure 1.2.1**) Therefore, it had converted nonstandard date to standard date formatting (yyyy-mm-dd) and created a new separate column called "diagnosis-minus-death". Also converted outcome variable into factor variable in the **figure 1.2.2.**

state	sex	diag	death	status	T.categ	age	year	outcome
NSW	M	10905	11081	D	hs	35	1999	0
NSW	M	10905	11081	D	hs	36	2000	1
NSW	M	11029	11096	D	hs	53	2000	1
NSW	M	9551	9983	D	hs	42	1996	0

Figure 1.2.1. the diag and death are using Julian date of diagnosis

state	sex	diag	death	status	T.categ	age	year	outcome	diagnosisminusdeath
NSW	M	1999-11-10	2000-05-04	D	hs	35	1999	Survived	176
NSW	M	1999-11-10	2000-05-04	D	hs	36	2000	Died	176
NSW	M	2000-03-13	2000-05-19	D	hs	53	2000	Died	67
NSW	M	1996-02-25	1997-05-02	D	hs	42	1996	Survived	432

Figure 1.2.2 converted nonstandard date to standard date formatting (yyyy-mm-dd) and created the separate column called "diagnosis minus death".

```
> #uploading data set
> rm(list=ls())
> library(ggplot2)
> library(plyr)
> library(forcats)
> Aids2ann <- read.csv("Aids2ann.csv")
> View(Aids2ann)
> #converting the date from julian format to standard format
> Aids2ann$diag <- as.Date(Aids2ann$diag,origin="1970-01-01")
> Aids2ann$death <- as.Date(Aids2ann$death,origin="1970-01-01")
> #number of days he/she was alive after diagnosis
> Aids2ann$diagnosisminusdeath <- Aids2ann$death- Aids2ann$diag
> #convert them into number of days
> Aids2ann$diagnosisminusdeath<-as.numeric(Aids2ann$diagnosisminusdeath)
> #convert outcome variable into factor variable
> Aids2ann$outcome<-factor(Aids2ann$outcome,labels=c("Survived","Died"))
```

Figure 1.2.3 R output for data preparation and cleaning

1.3 Data Exploration

Summary statistics of variables:

	Count	Min	Lower quartile	Median	Mean	Upper quartile	Max	Range	IQR	Standard Deviation	Missing Values
diag	6014	1992-09-24	1997-09-12	1998-11-07	1998-09-24	1999-12-22	2001-06-30	none	none	none	0
death	6014	1993-03-10	1999-08-06	2001-01-15	2000-04-25	2001-07-01	2001-07-01	none	none	none	0
age	6014	0	31	37	37.74	43	82	82	12	9.78	0
year	6014	1992	1998	1999	1999	2000	2001	none	none	none	0
diagnosis minus death	6014	0	250	496	579	801	2470	2470	551	445.79	0

Table 1.3.1 Summary statistics of variables

(Table 1.3.1) The mean age in the dataset is 37.74 years while the median age is 37 years. The minimum age is 0 (new born) while maximum age is 82 years. This indicates that there is considerable variation in age.

Although reported, **table 1.3.1** the summary statistics for diag and death do not much intuitive meaning since these variables are of date type. The same can be said about the year variable. The variable diagnosis-minus-death tells us about the time an individual has survived after diagnosis. The mean stands at 579 days while median is equal to 496 days.

Summary statistics of categorical variables:

Variable	Frequency	Relative Frequency	%
state			
Count	6014		
NSW	3775	0.6277	62.77 %
Other	544	0.0905	9.05 %
QLD	446	0.0741	7.41 %
VIC	1249	0.2077	20.77 %
sex			
Count	6014		
F	202	0.0336	3.36 %
M	5812	0.9664	96.64 %
status			
Count	6014		
A	2481	0.4125	41.25 %
D	3533	0.5875	58.75 %
T.categ			
Count	6014		
blood	187	0.0311	3.11 %
haem	89	0.0148	1.48 %
het	102	0.0170	1.70 %
hs	5217	0.8674	86.74 %
hsid	168	0.0279	2.79 %
id	108	0.0180	1.80 %
mother	15	0.0025	0.25 %
other	128	0.0213	2.13 %
outcome			
Count	6014		
0	4253	0.7072	70.72 %
1	1761	0.2928	29.28 %

Table 1.3.2 Summary statistics of categorical variables

We use the following commands for the variable. Codes Used in R Studio to generate results above for variables: Instead of using the summary () command, we opt to manually compute summary statistics as shown in **Table 1.3.1, 1.3.2** Summary statistics of variables to obtain a broader set of statistics.

```
> #$diag
> summary(Aids2ann$diag)
> length(Aids2ann$diag)
> quantile(Aids2ann$diag)
> range_Diag <- max(Aids2ann$diag) - min(Aids2ann$diag)
> range_Diag
> IQR_Diag <- quantile(Aids2ann$diag, .75) - quantile(Aids2ann$diag,.25)
> IQR_Diag
> sd(Aids2ann$diag)
```

Figure 1.3.3 R output for summary statistics. Similar codes were used for the other variables

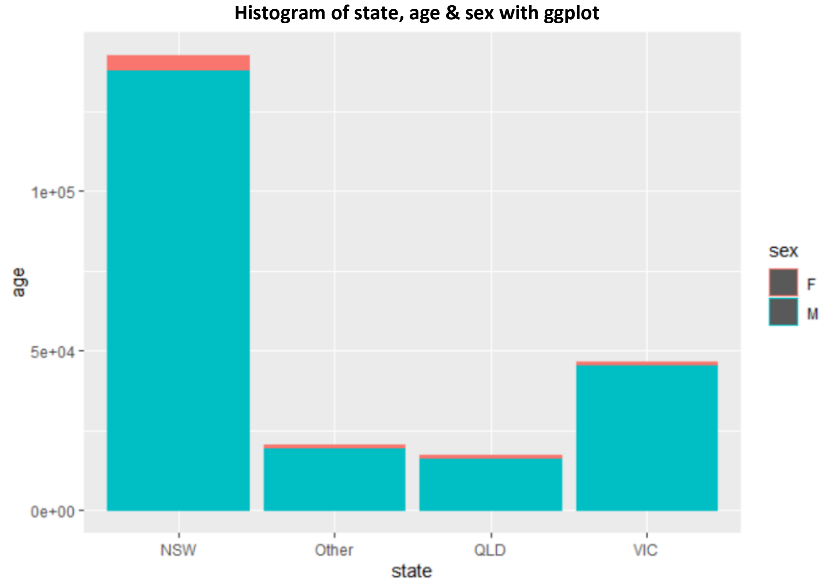


Figure 1.3.4: Histogram illustrating the relationship between state, age and sex with ggplot

```
> #histogram of age across different categories
> plot1<- ggplot(Aids2ann,
+ aes(y = age, x = state ,color= sex)) +
+ geom_histogram(stat="identity")
> plot1
```

Figure 1.3.5 R output for histogram illustrating the relationship between state, age and sex with ggplot

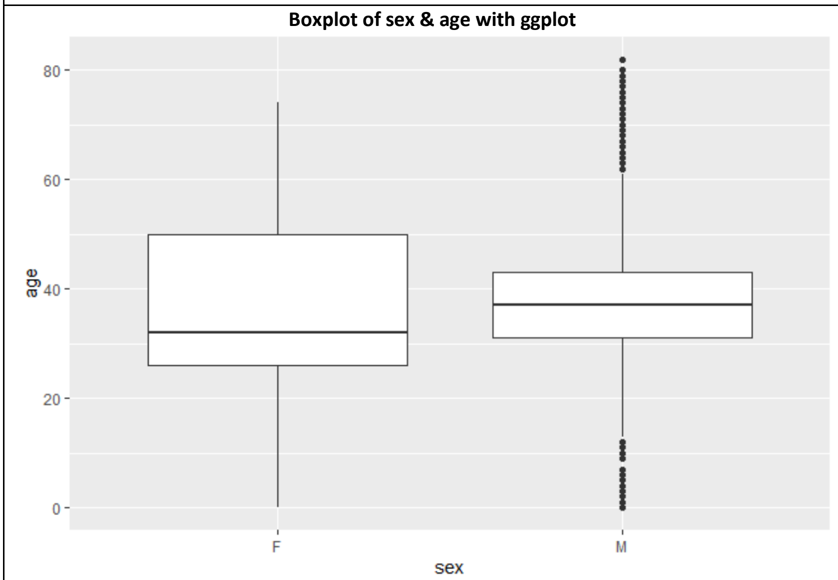


Figure 1.3.6: Boxplot illustrating the outliers and relationship between sex and age with ggplot

```
> #boxplot of age across different categories
> plot2 <- ggplot(Aids2ann, aes(x = sex, y = age)) + geom_boxplot()
> plot2
```

Figure 1.3.7 R output for boxplot illustrating the outliers and relationship between sex and age with ggplot

Distribution of the variables

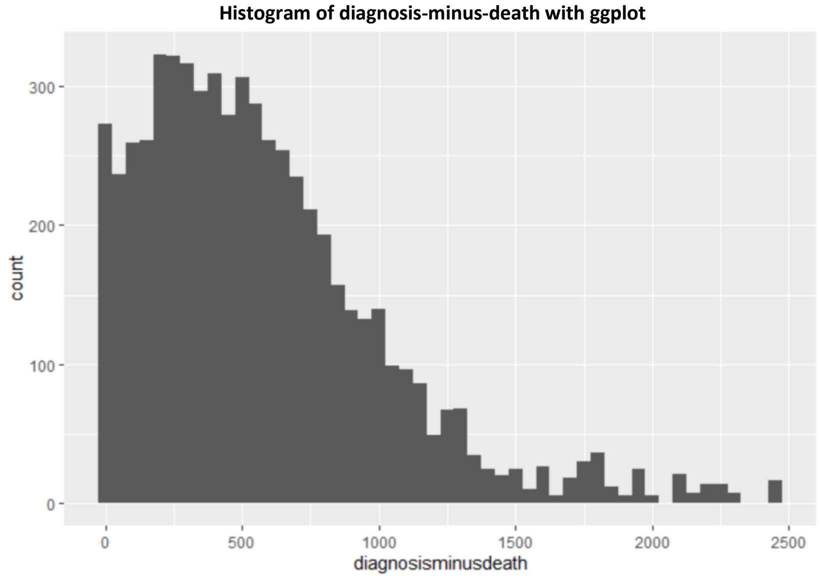


Figure 1.3.8: Histogram representing the distribution of diagnosis-minus-death with ggplot

```
> plot3<- ggplot(Aids2ann, aes(x = diagnosisminusdeath)) + geom_histogram(stat="bin",binwidth = 50)
> plot3
```

Figure 1.3.9 R output for histogram representing the distribution of diagnosis-minus-death with ggplot

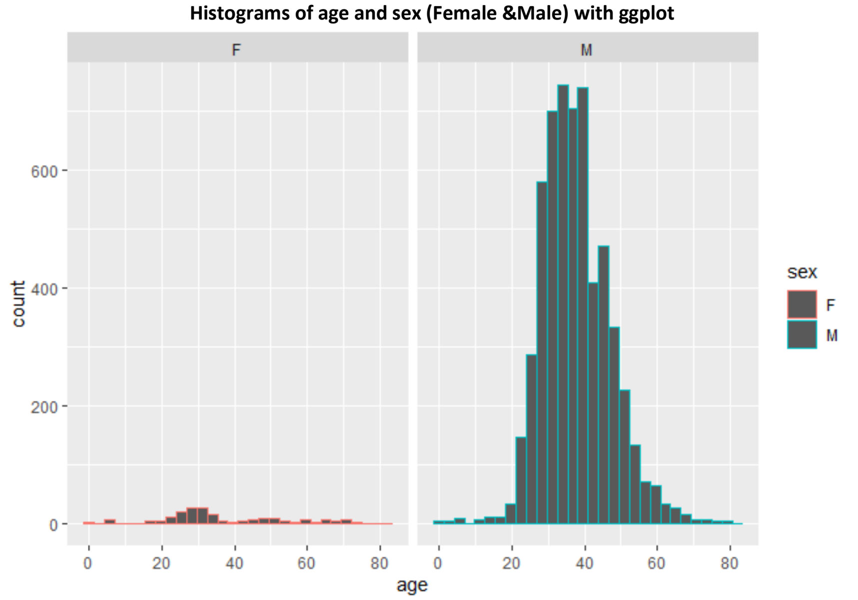


Figure 1.3.10: Histograms representing the distribution of age and sex (Female & Male) with ggplot

```
> #distribution of age
> plot4<- ggplot(Aids2ann, aes(age,color=sex))+ facet_wrap(Aids2ann$sex) +
+   geom_histogram(bins=30)
> plot4
```

Figure 1.3.11 R output for histograms representing the distribution of age and sex (Female & Male) with ggplot

Samples and populations:

This data is very large, so this project is working with a subset of the data and focusing only on nine variables and 6014 observations.

Issues with the data collection

In the **table 1.3.2**, there was a total of 6014 Patients and the first problem that had been detected with data was gender/sex, which is a categorical variable.

There are 202 observations (3.36%) for females and 5812 observations (96.64%) for males. Since the total number of males far exceed the number of females, the conclusions drawn from this dataset may not be widely applicable to women. Thus, the male-bias (in terms of count of observations) thus make our analysis more apt for males than females. One way to address this issue can be to include only those males and females who have similar characteristics in general.

Outliers or mistakes in the data

Boxplot is a graphical method of displaying distribution of a variable. It is drawn with the help of 5 number summary – Minimum, Maximum, Median, First Quartile and Fourth Quartile. The above boxplots (see **figure 1.3.6**: Boxplot illustrating the outliers and relationship between sex and age with ggplot) illustrates that there are numerous upper outliers and lower outliers for male age. The 1.5 IQR criterion tells us that any observation with an age that is below 13 or above 61 can be considered an outlier for males. The 1.5 criterion does not exhibit any outliers in age for females.

Distribution of the variables

- **Table 1.3.1** Summary statistics of variables reporting the distribution of "diagnosis-minus-death" is strongly skewed to the right. In this case the mean (579) is greater than the median (496), hence further satisfying that the data is not normally distributed for this variable. **Figure 1.3.8** affirms our suspicion as the histogram of this variable is skewed to the right.
- **Table 1.3.1,** the mean age in the dataset is 37.74 years while the median age is 37 years. The minimum age is 0 (new born) while maximum age is 82 years. This indicates that there is considerable variation in age.
- Distributions of age across both sexes seem nearly normal (see **figure 1.3.10**: Histograms representing the distribution of age and sex (Female & Male) with ggplot). The Age distribution for males looks symmetric. The same may not be said about age distribution females – although it nearly replicates a bell curve. Summary statistics of variables reporting when the data is symmetric and normally distributed, the mean is roughly close to the median; **Table 1.3.1** but in this case the mean (37.74) is greater than the median (37), hence further satisfying that the data is not normally distributed.

Task 2. What are the pairwise associations between variables in the dataset? Use correlation analysis, scatter plots, box plots, chi-squared tests to test for associations between pairs. You should choose 3-4 associations to investigate. What are the underlying assumptions of the statistical test that you applied? Are the assumptions satisfied? What do these test results mean?

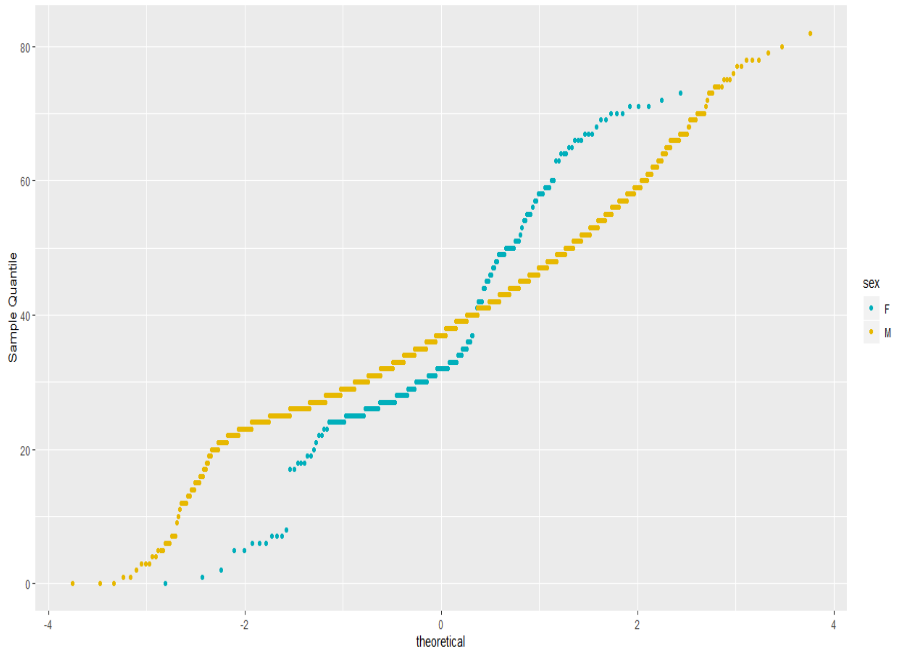
(**Figure 2.1**) We conduct Shapiro Wilk test of normality for Age. However, this test requires that the number of observations should be between 3 and 5000. Aids2ann dataset has more than 5000 observations and hence Shapiro Wilk test of normality cannot be conducted.

Hence, we conduct an alternative test of normality called Anderson-Darling normality test – (**Figure 2.1**)

```
> #Is Age normally distributed?
> shapiro.test(Aids2ann$age)
Error in shapiro.test(Aids2ann$age) :
  sample size must be between 3 and 5000
> library(nortest)
> normalitytest <- ad.test(Aids2ann$age)
> normalitytest

        Anderson-Darling normality test

data:  Aids2ann$age
A = 33.047, p-value < 2.2e-16

> #     Anderson-Darling normality test - null hypothesis of normal
ity is rejected at 5% level.
```

Figure 2.1 R output for Anderson-Darling normality test

H_0: Age *is normally distributed*
H_1: Age *is* not *normally distributed*
The p-value of the test statistic of the Anderson-Darling normality test was less than level of 5% significance, hence the null hypothesis was rejected. We therefore conclude that the data is not normally distributed. This violation of normality could impact the conclusion of the two sample t-test that we perform later on.
However, the qqplot of age for both genders shows that it may be approximately normal as some of the points will lie on the straight line.




```
#qqplot of age
plot5 <-ggplot(Aids2ann, aes(sample=age)) +
  stat_qq(aes(color = sex)) +
  scale_color_manual(values = c("#00AFBB", "#E7B800"))+
  labs(y = "Sample Quantile")
plot5
```

Figure 2.2 R output and the qqplot of the age variable for both sexes

(Figure 2.3) Here we test if the mean age across both genders is equal or not. The t-test require that the original data is normally distributed. In our case, the Age data is not normally distributed as per Anderson-Darling test. However, we still conduct a t-test across two groups. The mean age of male was 37.76 and the mean age of female was 37.13, the mean difference stands at 0.63.

Welch's t test:
H_0: $\mu_1 = \mu_2$
H_1: $\mu_1 \neq \mu_2$
Where μ_1 is the mean age of male and μ_2 being the mean age of female.

(Figure 2.3) The p-value of the test statistic is 0.5976. The results show there is no sufficient evidence to reject the null hypothesis as the p

```
> #Are the average age of male and female equal?
> t.test(Aids2ann$age[Aids2ann$sex=="M"], Aids2ann$age[Aids2ann$sex=="F"])

        Welch Two Sample t-test

data:  Aids2ann$age[Aids2ann$sex == "M"] and Aids2ann$age[Aids2ann$sex == "F"]
t = 0.52872, df = 205.32, p-value = 0.5976
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.735381  3.007205
sample estimates:
mean of x mean of y
 37.76462  37.12871
```

Figure 2.3 R output for t-test

<table>
<tr><td>value is greater than or equal to level of significance of 5%. This implies that the average age of male does not statistically differ from the average age of female at 5% level.</td><td></td></tr>
<tr><td>(Figure 2.4) Welch’s t test:
H_0: $\mu_1 = \mu_2$
H_1: $\mu_1 \neq \mu_2$
Where μ_1 is the mean age of survived and μ_2 being the mean age of died.

We also conduct a t-test of whether the average age of those who survived differs from average age of those who died. Those who died, their average age was 1.49 years more than average age of those who survived.
In this case, the p-value of the test statistic is less than level of significance of 5%. Hence, we may reject the null hypothesis of equality of means across both groups. This implies that the average age of those who survived differs significantly, to the average age of died.</td><td>> #Is the average age same across both outcomes?
> t.test(Aids2ann$age[Aids2ann$outcome=="Survived"], Aids2ann$age[Aids2ann$outcome=="Died"])

Welch Two Sample t-test

data: Aids2ann$age[Aids2ann$outcome == "Survived"] and Aids2ann$age[Aids2ann$outcome == "Died"]
t = -5.2399, df = 3077.3, p-value = 1.715e-07
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -2.0525517 -0.9347363
sample estimates:
mean of x mean of y
 37.30590 38.79955

Figure 2.4 R output for t-test</td></tr>
</table>

```
> #Is the average age same across states?
> fit = lm(age~state,Aids2ann)
> anova(fit)
Analysis of Variance Table

Response: age
            Df Sum Sq Mean Sq F value  Pr(>F)
state        3    619  206.32    2.16 0.09058 .
Residuals 6010 574077   95.52
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 2.5 R output for Analysis of Variance (ANOVA) table

(Figure 2.5) We also test whether the average age of the patients is same across all the states. Thus, the null hypothesis is whether average age is same across all the states. The alternative hypothesis states otherwise. The p-value of the ANOVA F statistic is 0.09058. The null hypothesis that average age is same across all states cannot be rejected at 5% level (as p-value >0.05). However, the same can be rejected at 10% level of significance. It is important to note that ANOVA test is that continuous variable (age) is normality distributed. However, that assumption seems to be violated as per Anderson-Darling normality test. (**Figure 2.1)**

```
> # test of independence between outcome and state
> chisq.test(Aids2ann$outcome,Aids2ann$state)

        Pearson's Chi-squared test

data:  Aids2ann$outcome and Aids2ann$state
X-squared = 6.5233, df = 3, p-value = 0.08875

> #Ho: Status and state are independent
> #We will reject Ho at 5% level.
```

Figure 2.6 R output for Chi-squared test

(Figure 2.6) I also check if the two categorical variables- outcome and state are independent. The null hypothesis is that both these variables are independent. The Chi-square test statistic, which is computed under the assumption that null hypothesis is true, has a p-value of 0.08875. The null hypothesis can be rejected at 10% level of significance. Thus, the two categorical variables may be dependent on each other.

```
> #test of independence between state and sex
> chisq.test(Aids2ann$state,Aids2ann$sex)

        Pearson's Chi-squared test

data:  Aids2ann$state and Aids2ann$sex
X-squared = 19.583, df = 3, p-value = 0.0002071

> #Ho: State and sex are independent
> #We reject Ho at 5% level.
```

Figure 2.7 R output for Chi-squared test

(Figure 2.7) Lastly, we run a Chi-square test of independence on state and sex. The null hypothesis will be that both state and sex are independent while the alternative hypothesis states otherwise. The p-value of the Chi-square test statistic is less than level of significance of 5%. Hence, we may reject the null hypothesis that both state and sex are independent.

```
> #correlation between number of days one surives and their age
> round(cor(Aids2ann$diagnosisminusdeath,Aids2ann$age),2)
[1] -0.03
```

Figure 2.8 R output for correlation of age and diagnosisminusdeath

(Figure 2.8) The dataset has very few continuous variables for which we can compute correlation. We thus look at only the correlation of *age* and *diagnosisminusdeath* to see if there is any correlation between age and the number of years one survives after diagnosis. A negative but nearly zero correlation of these 2 continuous variables indicate that they are not correlated at all.

Task 3. Use logistic regression to establish which variables affect the outcome, i.e. how likely for a particular patient to die in a particular year. Use the Likelihood Ratio Test (LRT) to assess the goodness of fit. Use confidence intervals on parameters to establish if a particular covariate has positive or negative effect on the outcome. Discuss the interpretation of the results and check the residuals plot. Discuss any weakness of this analysis and its effectiveness to answer the question above.

In task 3, we model probability of death as a function of independent variable –

$$P(Y=1) = \frac{e^{\beta_0+\beta_1 X_1+\beta_2 X_2+..+\beta_k X_k+u_i}}{1+e^{\beta_0+\beta_1 X_1+\beta_2 X_2+..+\beta_k X_k+u_i}}$$

The dependent variable in the regression model assumes only 2 values – either dead or survived in the year. Hence, we use logistic regression methodology to proceed further. The independent variables that are the considered in the regression model are – Sex, Age and State. Of the three regressors in the model, Sex and State are categorical variables while Age is a continuous variable. I also interact age with other categorical variables in the regression.

(Figure 3.1) The estimation of the logistic regression is done through Maximum Likelihood Estimation. In R, we can use glm package and select binomial family to run a logistic regression. **Below is the regression result for this exercise:**

```
> mylogit <- glm(outcome~sex+age+age*sex+state*age+state,data = Aids2ann, family = "binomial")
> summary(mylogit)

Call:
glm(formula = outcome ~ sex + age + age * sex + state * age +
    state, family = "binomial", data = Aids2ann)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.2024  -0.8444  -0.7910   1.4772   2.0931

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)    -2.295634   0.432538  -5.307 1.11e-07 ***
sexM            0.727685   0.438926   1.658 0.097342 .
age             0.032250   0.009790   3.294 0.000987 ***
stateOther      0.791010   0.450544   1.756 0.079144 .
stateQLD        0.815970   0.399205   2.044 0.040955 *
stateVIC        0.170568   0.301631   0.565 0.571742
sexM:age       -0.013790   0.010045  -1.373 0.169825
age:stateOther -0.024985   0.011583  -2.157 0.030996 *
age:stateQLD   -0.016918   0.009938  -1.702 0.088689 .
age:stateVIC   -0.005694   0.007720  -0.737 0.460840
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

(Figure 3.1) We can also look at the odds ratios coefficients of these regressors –

```
exp(coef(mylogit))
  (Intercept)           sexM            age     stateOther       stateQLD       stateVIC
    0.1006976      2.0702823      1.0327756      2.2056222      2.2613676      1.1859787
     sexM:age age:stateOther   age:stateQLD   age:stateVIC
    0.9863050      0.9753246      0.9832241      0.9943226
```

Figure 3.1 R output for odds ratio in logistic regression

Below is the 95% confidence interval for these regression coefficients (Figure 3.2):

```
> round(confint(mylogit),2)

                  2.5 % 97.5 %
(Intercept)       -3.18  -1.48
sexM              -0.11   1.62
age                0.01   0.05
stateOther        -0.10   1.67
stateQLD           0.02   1.59
stateVIC          -0.42   0.76
sexM:age          -0.03   0.01
age:stateOther    -0.05   0.00
age:stateQLD      -0.04   0.00
age:stateVIC      -0.02   0.01
```

Figure 3.2 R output for confidence interval of the coefficients

(Figure 3.2) The null hypothesis of each of these coefficients is that its hypothesized value of the true parameter is equal to 0. If 0 is not contained in the 95% confidence interval, then we can reject the null hypothesis and conclude that the coefficient is statistically significant from 0. The coefficients of variables that are significant at 95% level are – *age, stateQLD* and interaction variable – *stateOther*. Ceteris paribus, the results indicate that there is a positive association between *age* and predicted probability of death. Thus, the probability of death increases with age.
However, this relationship between age and predicted probability of death may be different across states and sex. This will be captured by the interaction terms. We look at the interaction graphs to see how relationship between predicted probability of death and age may evolve across different age groups. We will use interactions package in R for this purpose.

This is shown in below diagrams (Figure 3.3):

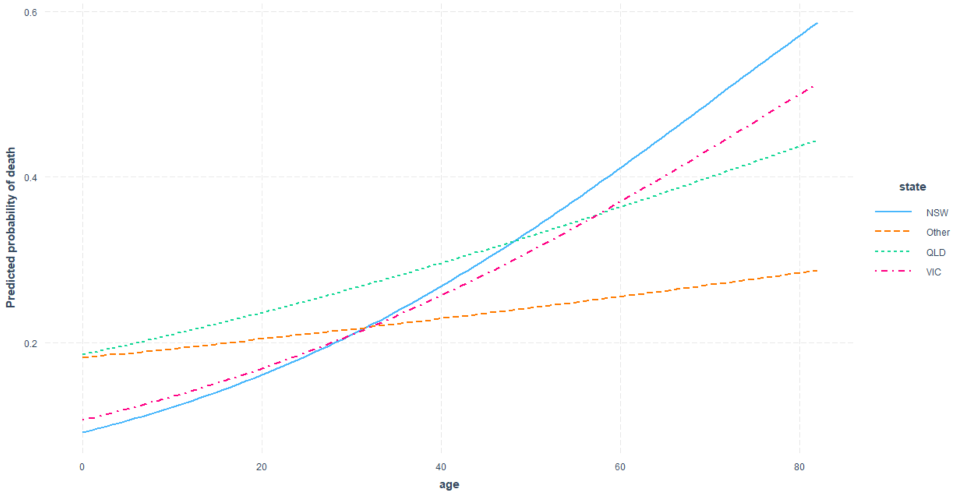


Figure 3.3 Relationship between predicted probability of death and age across different states

(Figure 3.3) As shown here, in all the states, the predicted probability of death increases as age increases. The predicted probability of death is the highest for young people in the state of QLD. However, for middle age and elderly, predicted probability of death is the highest in state of NSW. Given age, the predicted probability of death in VIC closely tracks the predicted probability of death in NSW.

Next, we also look at the how age affects the predicted probability of death across sexes –

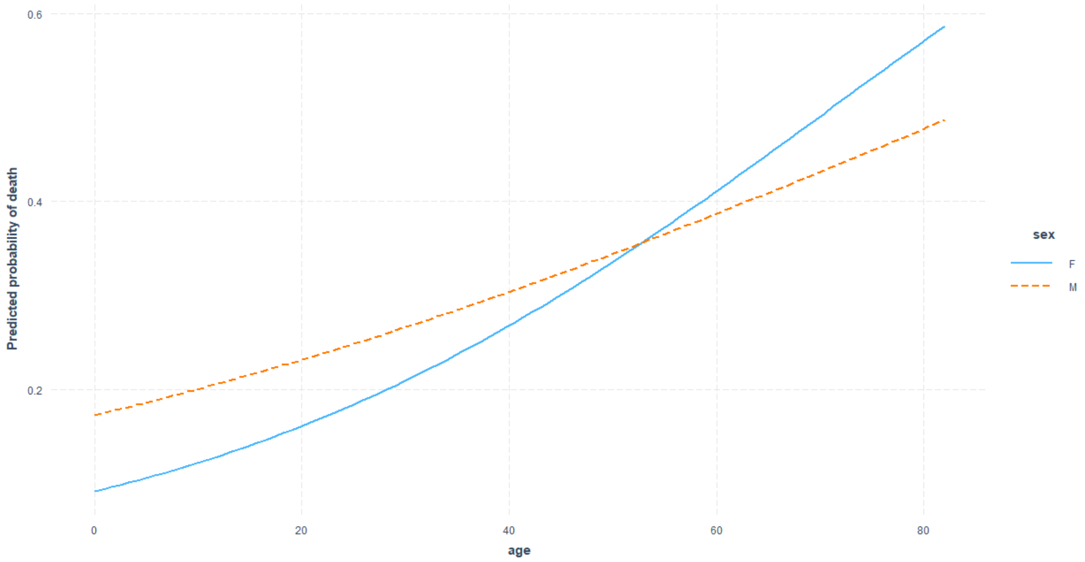


Figure 3.4 Relationship between predicted probability of death and age for both genders

R output for Figure 3.3 & 3.4 :

```
interact_plot(mylogit, pred = age, modx = state,y.label = "Predicted probability of death")
interact_plot(mylogit, pred = age, modx = sex, y.label = "Predicted probability of death")
```

(Figure 3.4) Given age less than 55 years (approximately), the predicted probability of death is higher for male than females. However, the predicted probability of death is higher for females than males given that their age is more than 55 years(approximately).

Residual analysis

(Figure 3.5) I also assess the behaviour of the residuals by looking at the binned residual plot using the binned plot function from the arm package. The red lines represent the ± 2 standard errors (SE) – essentially 95% confidence interval. Almost all of the fitted values lie with in the 95% SE band which implies that may be a good model.

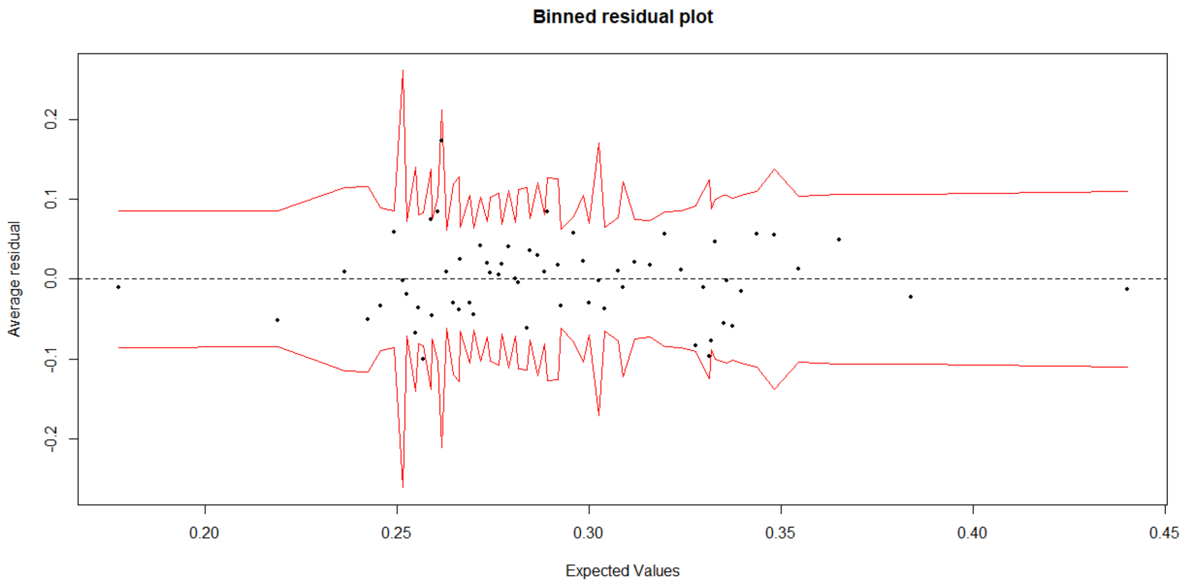


Figure 3.5 Binned residual plot

R output for Figure 3.5 :

```
binnedplot(fitted(mylogit),
           residuals(mylogit, type = "response"),
           nclass = NULL,
           xlab = "Expected Values",
           ylab = "Average residual",
           cex.pts = 0.6,
           col.pts = 1,
           col.int = "red")
```

(Figure 3.5) We also perform a likelihood ratio test that is Logistic Regression's equivalent of the F-test of joint significance of Multiple Linear regression. This test the validity of this model against a constant only model. **Below are the results: Binned residual plot**

```
> anova(mylogitconstant,mylogit)
Analysis of Deviance Table

Model 1: outcome ~ 1
Model 2: outcome ~ sex + age + state + age * sex + state * age
  Resid. Df Resid. Dev Df Deviance
1      6013     7272.8
2      6004     7228.4  9   44.418
> #p-value is less than 1% and 5% => We reject H0.
> 1-pchisq(35.673,df=9)
[1] 4.530792e-05
```

Figure 3.6 Likelihood ratio test

(Figure 3.6) Since the p-value of the test statistic is less than the alpha of 1% and 5%, we may reject the null hypothesis that states all the slope coefficients are jointly equal to 0.

Limitations of the analysis

Apart from the fact that there is an overrepresentation of male which could lead to spurious findings, the data set had limited number of continuous independent variables. This limits our understanding of how certain continuous variables such as income and education could be impacting the mortality since those with higher income and education are more likely to able to afford better treatment for diseases. Higher income and social status are linked to better health. The greater the gap between the richest and poorest people, the greater the differences in health. Similarly, low education levels are linked with poor health, more stress and lower self-confidence.

Appendix 1 – R Output for Task 1

We use the following commands for the variable. Codes Used in R Studio to generate results above for variables: Instead of using the summary() command, we opt to manually compute summary statistics as shown in Table 1.2 Summary statistics of variables to obtain a broader set of statistics.

```
> #$diag
> summary(Aids2ann$diag)
> length(Aids2ann$diag)
> quantile(Aids2ann$diag)
> range_Diag <- max(Aids2ann$diag) - min(Aids2ann$diag)
> range_Diag
> IQR_Diag <- quantile(Aids2ann$diag, .75) - quantile(Aids2ann$diag,.25)
> IQR_Diag
> sd(Aids2ann$diag)
```

Similar codes were used for the other variables

```
> #summary stats
> Aids2ann<-Aids2ann[,2:11]
> summary(Aids2ann)
   state        sex            diag                     death               status        T.categ
 NSW   :3775   F: 202   Min.   :1992-09-24   Min.   :1993-03-10   A:2481   hs       :5217
 Other: 544    M:5812   1st Qu.:1997-09-12   1st Qu.:1999-08-06   D:3533   blood   : 187
 QLD   : 446            Median :1998-11-07   Median :2001-01-15            hsid    : 168
 VIC   :1249            Mean   :1998-09-24   Mean   :2000-04-25            other   : 128
                        3rd Qu.:1999-12-22   3rd Qu.:2001-07-01            id      : 108
                        Max.   :2001-06-30   Max.   :2001-07-01            het     : 102
                                                                           (Other): 104
      age              year          outcome       diagnosisminusdeath
 Min.   : 0.00   Min.   :1992   Survived:4253   Min.   :   0
 1st Qu.:31.00   1st Qu.:1998   Died    :1761   1st Qu.: 250
 Median :37.00   Median :1999                   Median : 496
 Mean   :37.74   Mean   :1999                   Mean   : 579
 3rd Qu.:43.00   3rd Qu.:2000                   3rd Qu.: 801
 Max.   :82.00   Max.   :2001                   Max.   :2470
```

```
> #uploading data set
> rm(list=ls())
> library(ggplot2)
> library(plyr)
> library(forcats)
> Aids2ann <- read.csv("Aids2ann.csv")
> View(Aids2ann)
> #converting the date from julian format to standard format
> Aids2ann$diag <- as.Date(Aids2ann$diag,origin="1970-01-01")
> Aids2ann$death <- as.Date(Aids2ann$death,origin="1970-01-01")
> #number of days he/she was alive after diagnosis
> Aids2ann$diagnosisminusdeath <- Aids2ann$death- Aids2ann$diag
> #convert them into number of days
> Aids2ann$diagnosisminusdeath<-as.numeric(Aids2ann$diagnosisminusdeath)
> #convert outcome variable into factor variable
> Aids2ann$outcome<-factor(Aids2ann$outcome,labels=c("Survived","Died"))
> levels(Aids2ann$outcome)
[1] "Survived" "Died"

> table(Aids2ann$outcome)

Survived     Died
    4253     1761
```

```
> #data structure
> str(Aids2ann)
'data.frame':	6014 obs. of  11 variables:
 $ X                  : int  1 2 21 3 31 4 5 51 6 61 ...
 $ state              : Factor w/ 4 levels "NSW","Other",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ sex                : Factor w/ 2 levels "F","M": 2 2 2 2 2 2 2 2 2 2 ...
 $ diag               : Date, format: "1999-11-10" "1999-11-10" ...
 $ death              : Date, format: "2000-05-04" "2000-05-04" ...
 $ status             : Factor w/ 2 levels "A","D": 2 2 2 2 2 2 2 2 2 2 ...
 $ T.categ            : Factor w/ 8 levels "blood","haem",..: 4 4 4 4 4 2 4 4 4 4 ...
 $ age                : int  35 36 53 42 43 44 39 40 36 37 ...
 $ year               : int  1999 2000 2000 1996 1997 1996 1997 1998 1997 1998 ...
 $ outcome            : Factor w/ 2 levels "Survived","Died": 1 2 2 1 2 2 1 2 1 2 ...
 $ diagnosisminusdeath: num  176 176 67 432 432 77 275 275 373 373 ...
```

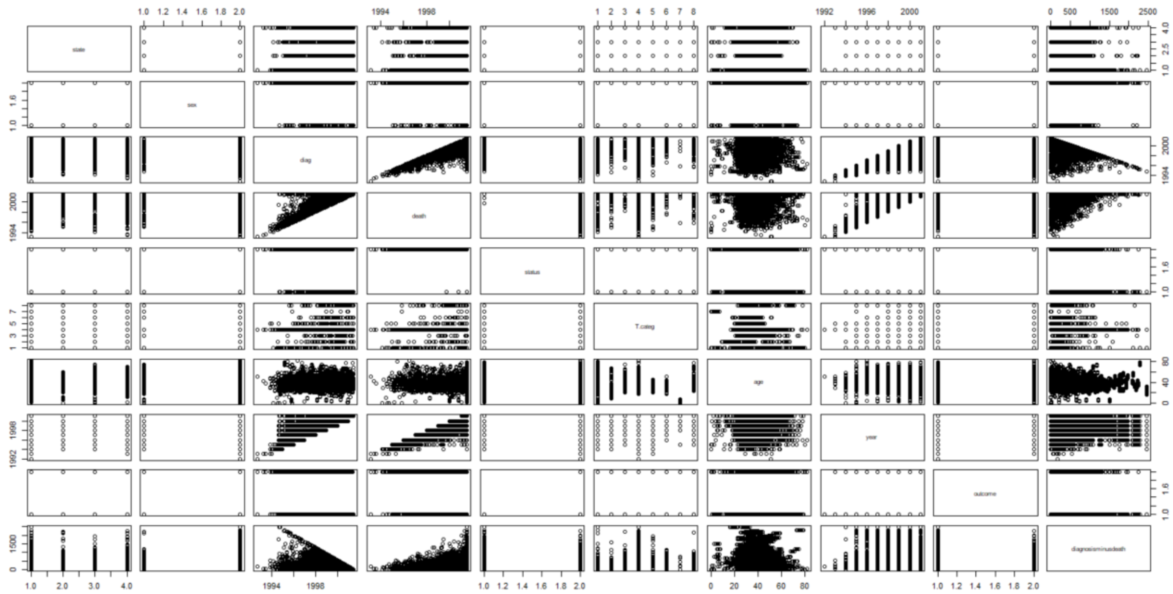
```
> str(Aids2ann)
'data.frame':	6014 obs. of  10 variables:
 $ state              : Factor w/ 4 levels "NSW","Other",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ sex                : Factor w/ 2 levels "F","M": 2 2 2 2 2 2 2 2 2 2 ...
 $ diag               : Date, format: "1999-11-10" "1999-11-10" ...
 $ death              : Date, format: "2000-05-04" "2000-05-04" ...
 $ status             : Factor w/ 2 levels "A","D": 2 2 2 2 2 2 2 2 2 2 ...
 $ T.categ            : Factor w/ 8 levels "blood","haem",..: 4 4 4 4 4 2 4 4 4 4 ...
 $ age                : int  35 36 53 42 43 44 39 40 36 37 ...
 $ year               : int  1999 2000 2000 1996 1997 1996 1997 1998 1997 1998 ...
 $ outcome            : Factor w/ 2 levels "Survived","Died": 1 2 2 1 2 2 1 2 1 2 ...
 $ diagnosisminusdeath: num  176 176 67 432 432 77 275 275 373 373 ...
```

Summary for diagnosisminusdeath

```
> #$diagnosisminusdeath
> summary(Aids2ann$diagnosisminusdeath)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
      0     250     496     579     801    2470 
> length(Aids2ann$diagnosisminusdeath)
[1] 6014
> quantile(Aids2ann$diagnosisminusdeath)
  0%  25%  50%  75% 100% 
   0  250  496  801 2470 
> range_diagnosisminusdeath <- max(Aids2ann$diagnosisminusdeath) - min(Aids2ann$diagnosismin
usdeath)
> range_diagnosisminusdeath
[1] 2470
> IQR_diagnosisminusdeath <- quantile(Aids2ann$diagnosisminusdeath, .75) - quantile(Aids2ann
$diagnosisminusdeath,.25)
> IQR_diagnosisminusdeath
75% 
551 
> sd(Aids2ann$diagnosisminusdeath)
[1] 445.7868
```

```
> #tabulating the cateogorical variable
> table(Aids2ann$state)

  NSW Other   QLD   VIC 
 3775   544   446  1249 
> table(Aids2ann$sex)

   F    M 
 202 5812 
> table(Aids2ann$outcome)

Survived     Died 
    4253     1761 
> table(Aids2ann$T.categ)

 blood    haem     het      hs    hsid      id mother  other 
   187      89     102    5217     168     108     15    128 
> table1 = table(Aids2ann$outcome,Aids2ann$state)
> round(prop.table(table1,2),2)
          
            NSW Other  QLD  VIC
  Survived 0.70  0.74 0.67 0.72
  Died     0.30  0.26 0.33 0.28
> table2 = table(Aids2ann$outcome,Aids2ann$sex)
> round(prop.table(table2,2),2)
          
              F    M
  Survived 0.74 0.71
  Died     0.26 0.29
```

```
> table3 = table(Aids2ann$state,Aids2ann$sex)
> round(prop.table(table3,2),2)

          F    M
  NSW   0.59 0.63
  Other 0.17 0.09
  QLD   0.08 0.07
  VIC   0.15 0.21
```



Histogram of state, age & sex (check this age) with ggplot

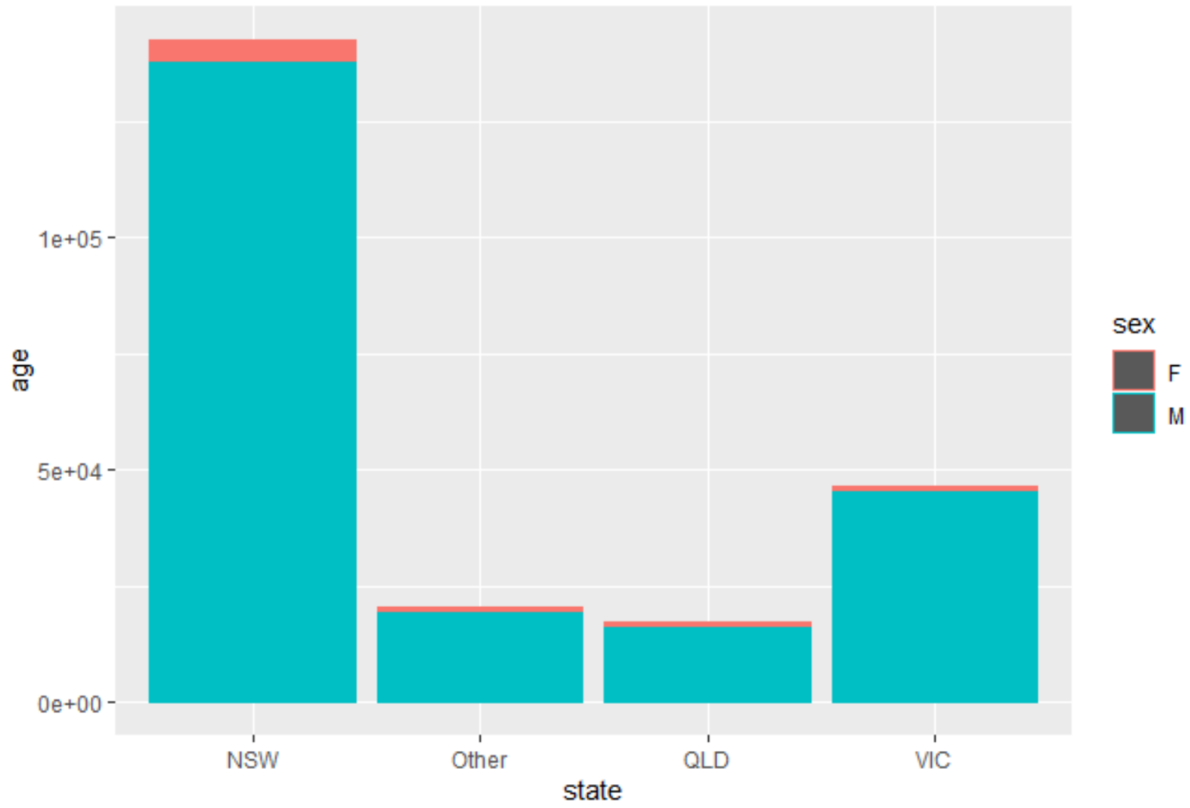


Figure 1.3.4: Histogram illustrating the relationship between state, age and sex with ggplot

```
> #histogram and boxplot of age across different categories
> plot1<- ggplot(Aids2ann,
+ aes(y = age, x = state,color= sex)) +
+ geom_histogram(stat="identity")
> plot1
```

Figure 1.3.5 R output for histogram illustrating the relationship between state, age and sex with ggplot

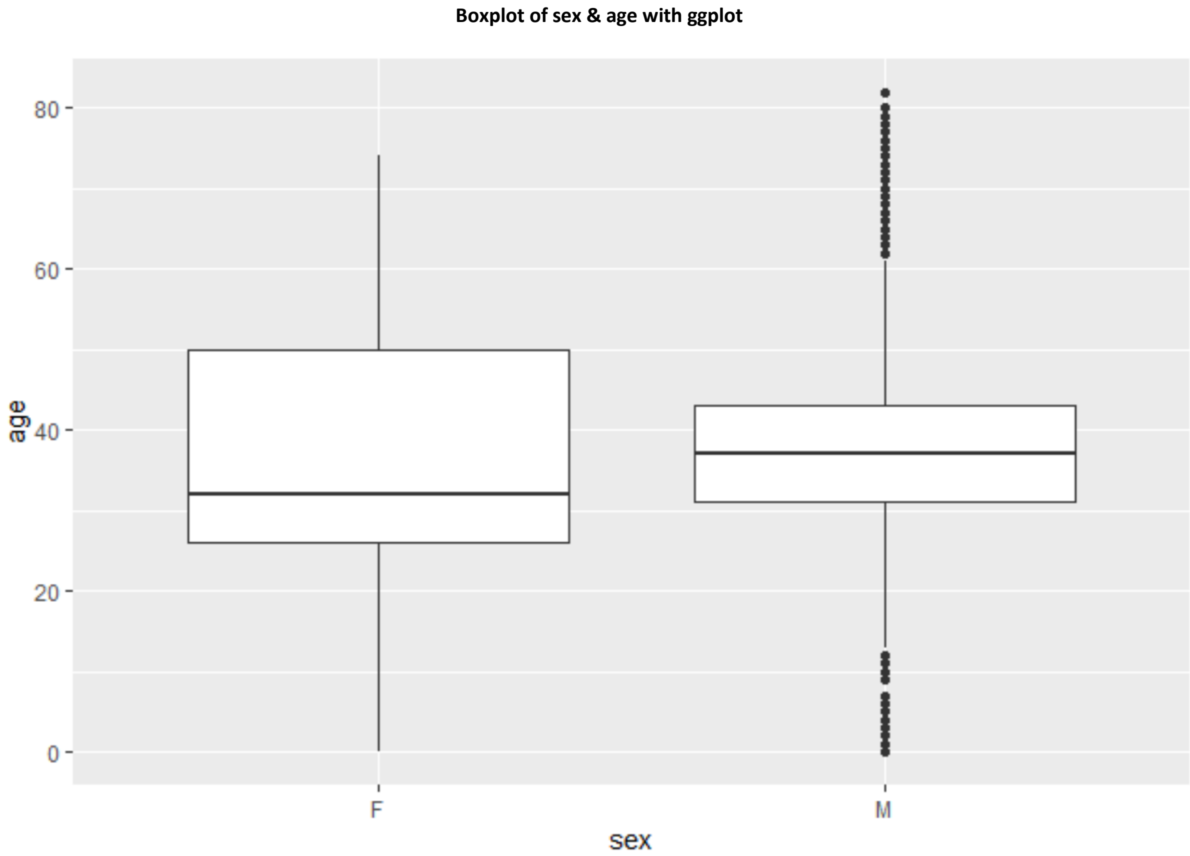


Figure 1.3.6: Boxplot illustrating the outliers and relationship between sex and age with ggplot

```
> plot2 <- ggplot(Aids2ann, aes(x = sex, y = age)) + geom_boxplot()
> plot2
```

Figure 1.3.7 R output for boxplot illustrating the outliers and relationship between sex and age with ggplot

Histogram of diagnosis minus death with ggplot

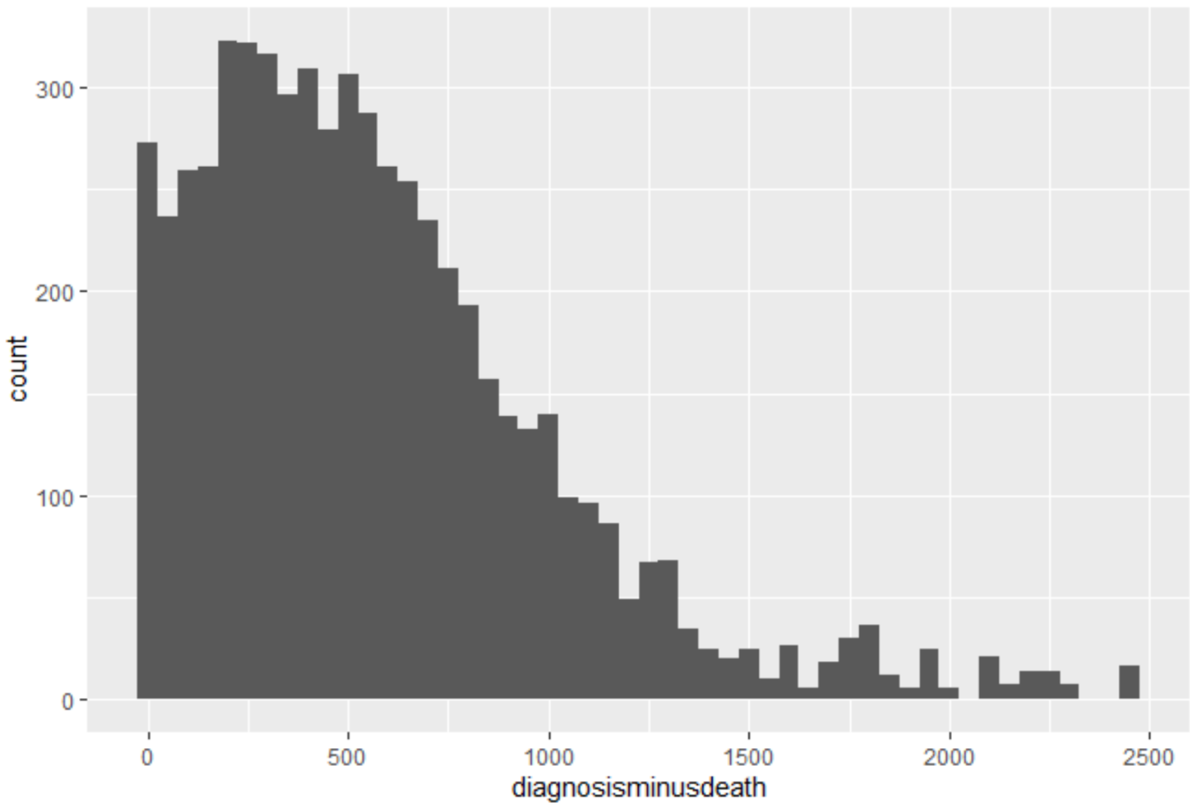


Figure 1.3.8: Histogram representing the distribution of diagnosis- minus-death with ggplot

```
> plot3<- ggplot(Aids2ann, aes(x = diagnosisminusdeath)) + geom_histogram(stat="bin",binwidth
= 50)
> plot3
```

Figure 1.3.9 R output for histogram representing the distribution of diagnosis minus death with ggplot

Histograms of age and sex (Female &Male) with ggplot

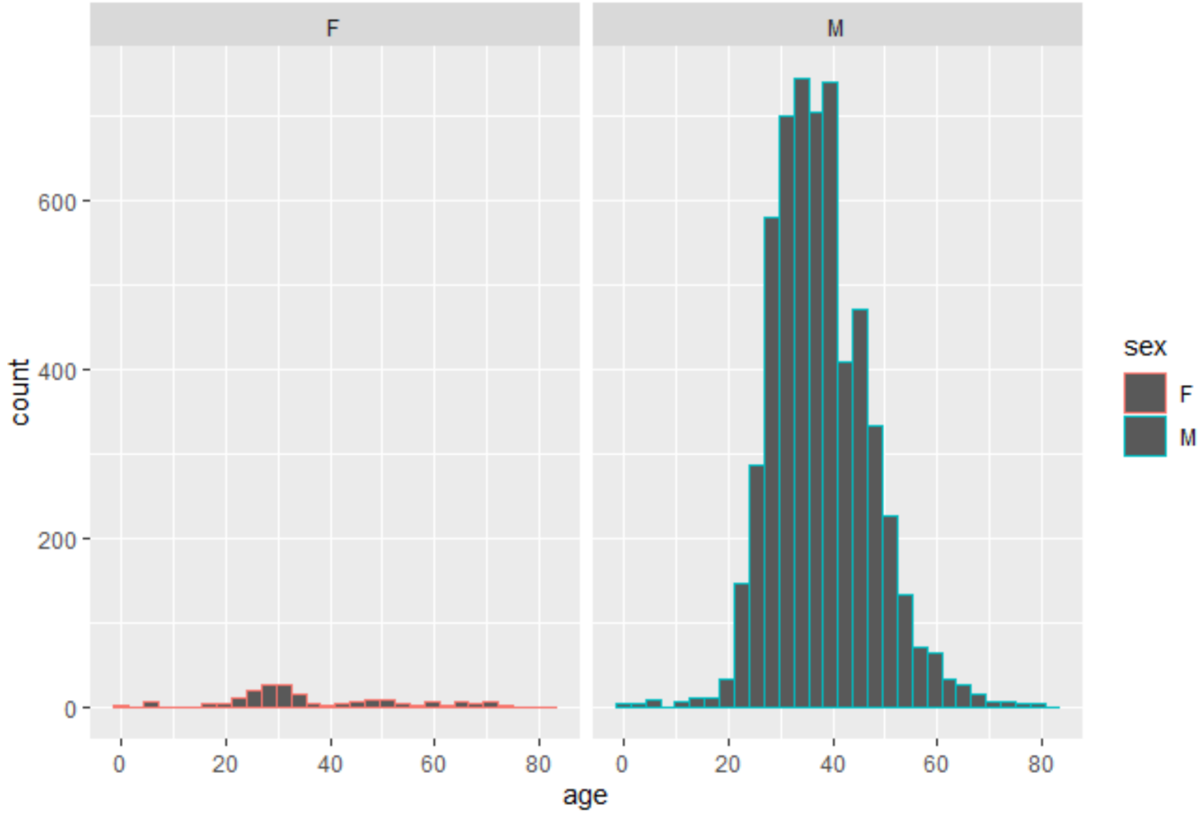


Figure 1.3.10: Histograms representing the distribution of age and sex (Female & Male) with ggplot

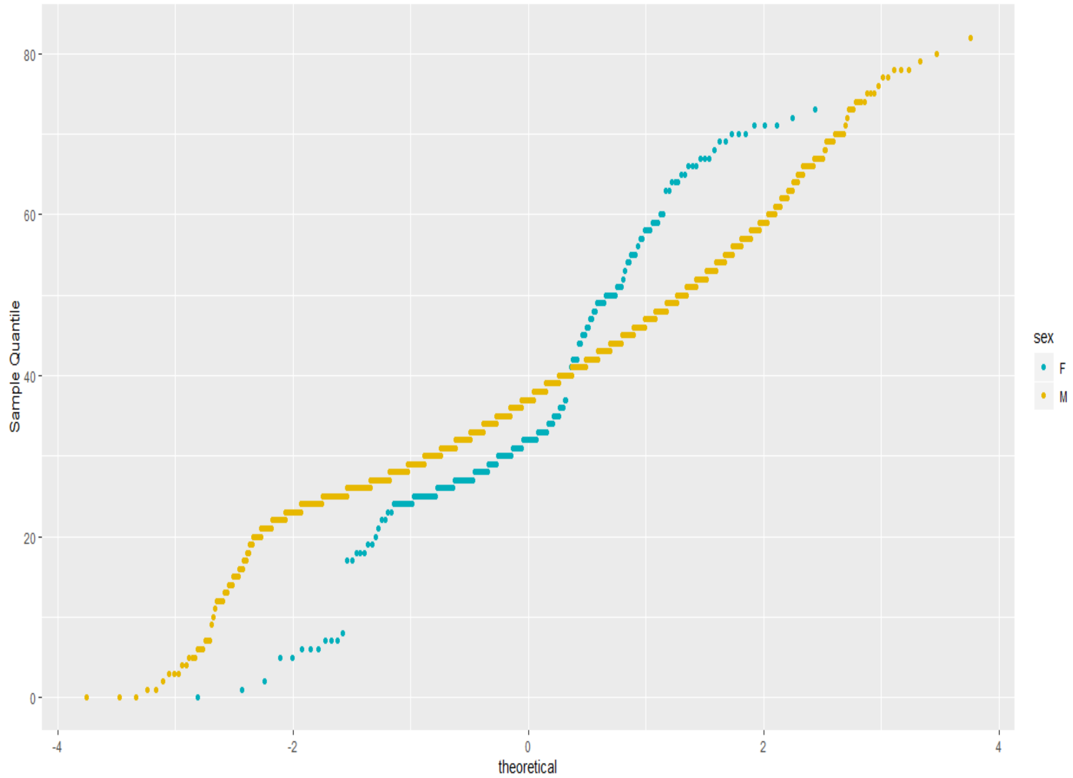
```
> #distribution of age
> plot4<- ggplot(Aids2ann, aes(age,color=sex))+ facet_wrap(Aids2ann$sex) +
+   geom_histogram(bins=30)
> plot4
```

Figure 1.3.11 R output for histograms representing the distribution of age and sex (Female & Male) with ggplot

Appendix 2 – R Output for Task 2

```
> #Is Age normally distributed?
> shapiro.test(Aids2ann$age)
Error in shapiro.test(Aids2ann$age) :
  sample size must be between 3 and 5000
> library(nortest)
> normalitytest <- ad.test(Aids2ann$age)
> normalitytest

	Anderson-Darling normality test

data:  Aids2ann$age
A = 33.047, p-value < 2.2e-16

> #	Anderson-Darling normality test - null hypothesis of normal
ity is rejected at 5% level.
```

Figure 2.1 R output for Anderson-Darling normality test




```
#qqplot of age
plot5 <-ggplot(Aids2ann, aes(sample=age)) +
  stat_qq(aes(color = sex)) +
  scale_color_manual(values = c("#00AFBB", "#E7B800"))+
  labs(y = "Sample Quantile")
plot5
```

Figure 2.2 R output and the qqplot of the age variable for both sexes

```
> #Are the average age of male and female equal?
> t.test(Aids2ann$age[Aids2ann$sex=="M"], Aids2ann$age[Aids2ann$sex=="F"])

        Welch Two Sample t-test

data:  Aids2ann$age[Aids2ann$sex == "M"] and Aids2ann$age[Aids2ann$sex == "F"]
t = 0.52872, df = 205.32, p-value = 0.5976
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.735381  3.007205
sample estimates:
mean of x mean of y 
 37.76462  37.12871 
```

Figure 2.3 R output for t-test

```
> #Is the average age same across both outcomes?
> t.test(Aids2ann$age[Aids2ann$outcome=="Survived"], Aids2ann$age[Aids2ann$outcome=="Died"])

        Welch Two Sample t-test

data:  Aids2ann$age[Aids2ann$outcome == "Survived"] and Aids2ann$age[Aids2ann$outcome == "Died"]
t = -5.2399, df = 3077.3, p-value = 1.715e-07
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -2.0525517 -0.9347363
sample estimates:
mean of x mean of y 
 37.30590  38.79955 
```

Figure 2.4 R output for t-test

```
> #Is the average age same across states?
> fit = lm(age~state,Aids2ann)
> anova(fit)
Analysis of Variance Table

Response: age
            Df Sum Sq Mean Sq F value  Pr(>F)  
state        3    619  206.32    2.16 0.09058 .
Residuals 6010 574077   95.52                  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 2.5 R output for analysis of variance table

```
> # test of independence between outcome and state
> chisq.test(Aids2ann$outcome,Aids2ann$state)

        Pearson's Chi-squared test

data:  Aids2ann$outcome and Aids2ann$state
X-squared = 6.5233, df = 3, p-value = 0.08875

> #Ho: Status and state are independent
> #We will reject Ho at 5% level.
```

Figure 2.6 R output for Chi-squared test

```
> #test of independence between state and sex
> chisq.test(Aids2ann$state,Aids2ann$sex)

        Pearson's Chi-squared test

data:  Aids2ann$state and Aids2ann$sex
X-squared = 19.583, df = 3, p-value = 0.0002071

> #Ho: State and sex are independent
> #We reject Ho at 5% level.
```

Figure 2.7 R output for Chi-squared test

```
> #correlation between number of days one surives and their age
> round(cor(Aids2ann$diagnosisminusdeath,Aids2ann$age),2)
[1] -0.03
```

Figure 2.8 R output for correlation of age and diagnosisminusdeath

Appendix 3 – R Output for Task 3

```
> mylogit <- glm(outcome~sex+age+age*sex+state*age+state,data = Aids2ann, family = "binomial")
> summary(mylogit)

Call:
glm(formula = outcome ~ sex + age + age * sex + state * age + 
    state, family = "binomial", data = Aids2ann)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-1.2024  -0.8444  -0.7910   1.4772   2.0931  

Coefficients:
                Estimate Std. Error z value Pr(>|z|)    
(Intercept)    -2.295634   0.432538  -5.307 1.11e-07 ***
sexM            0.727685   0.438926   1.658 0.097342 .  
age             0.032250   0.009790   3.294 0.000987 ***
stateOther      0.791010   0.450544   1.756 0.079144 .  
stateQLD        0.815970   0.399205   2.044 0.040955 *  
stateVIC        0.170568   0.301631   0.565 0.571742    
sexM:age       -0.013790   0.010045  -1.373 0.169825    
age:stateOther -0.024985   0.011583  -2.157 0.030996 *  
age:stateQLD   -0.016918   0.009938  -1.702 0.088689 .  
age:stateVIC   -0.005694   0.007720  -0.737 0.460840    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

We can also look at the odds ratios coefficients of these regressors –

```
exp(coef(mylogit))
  (Intercept)           sexM            age     stateOther       stateQLD       stateVIC 
    0.1006976      2.0702823      1.0327756      2.2056222      2.2613676      1.1859787 
     sexM:age age:stateOther   age:stateQLD   age:stateVIC 
    0.9863050      0.9753246      0.9832241      0.9943226 
```

Figure 3.1 R output for odds ratio in logistic regression

Below is the 95% confidence interval for these regression coefficients:

```
> round(confint(mylogit),2)

               2.5 % 97.5 %
(Intercept)    -3.18  -1.48
sexM           -0.11   1.62
age             0.01   0.05
stateOther     -0.10   1.67
stateQLD        0.02   1.59
stateVIC       -0.42   0.76
sexM:age       -0.03   0.01
age:stateOther -0.05   0.00
age:stateQLD   -0.04   0.00
age:stateVIC   -0.02   0.01
```

Figure 3.2 R output for confidence interval of the coefficients

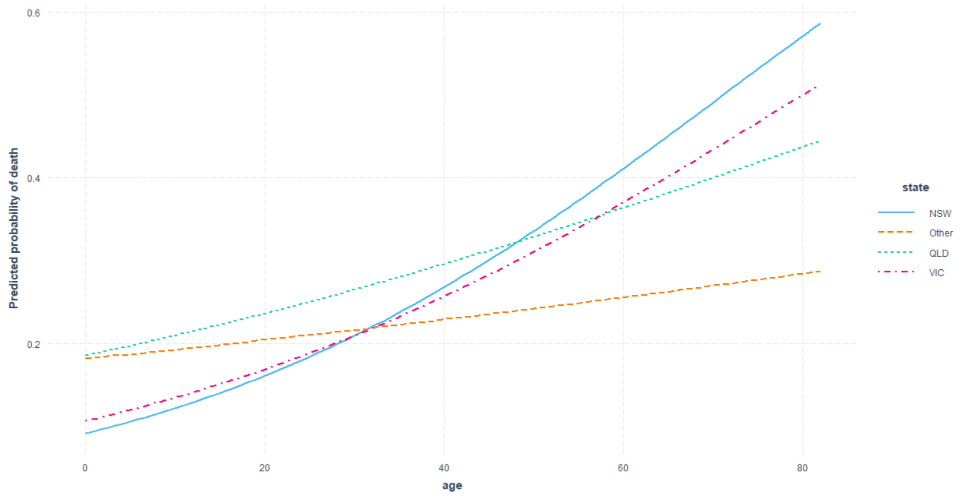


Figure 3.3 Relationship between predicted probability of death and age across different states

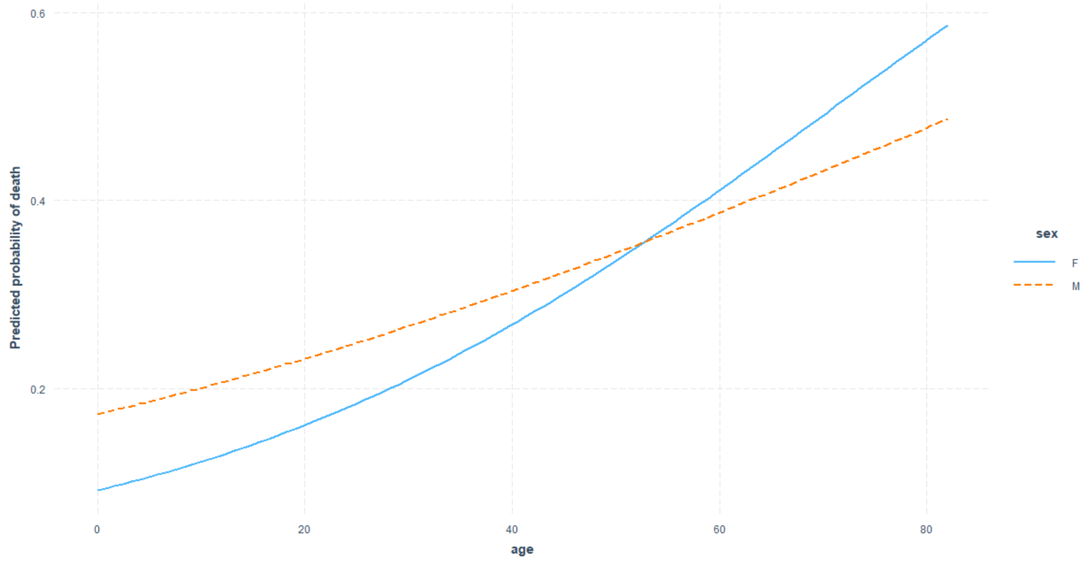


Figure 3.4 Relationship between predicted probability of death and age for both genders

```
interact_plot(mylogit, pred = age, modx = state,y.label = "Predicted probability of death")
interact_plot(mylogit, pred = age, modx = sex, y.label = "Predicted probability of death")
```

R output for Figure 3.3 & 3.4

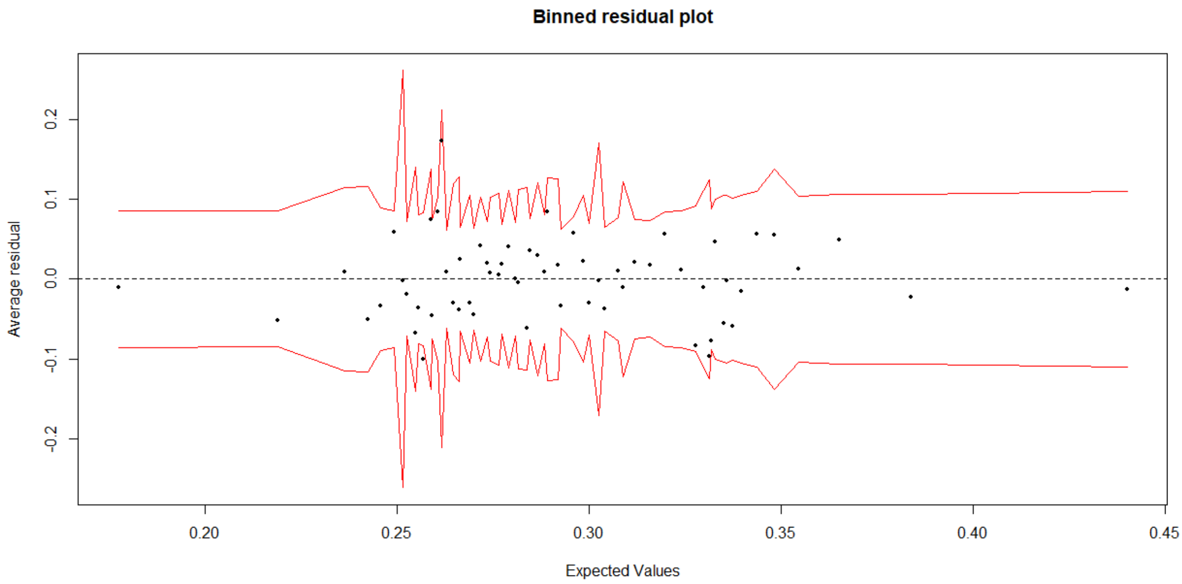


Figure 3.5 Binned residual plot

R output for Figure 3.5 :

```
binnedplot(fitted(mylogit),
           residuals(mylogit, type = "response"),
           nclass = NULL,
           xlab = "Expected Values",
           ylab = "Average residual",
           cex.pts = 0.6,
           col.pts = 1,
           col.int = "red")
```

```
> anova(mylogitconstant,mylogit)
Analysis of Deviance Table

Model 1: outcome ~ 1
Model 2: outcome ~ sex + age + state + age * sex + state * age
  Resid. Df Resid. Dev Df Deviance
1      6013     7272.8
2      6004     7228.4  9   44.418
> #p-value is less than 1% and 5% => We reject H0.
> 1-pchisq(35.673,df=9)
[1] 4.530792e-05
```

Figure 3.6 Likelihood ratio test

Appendix 4 – All R Output

Task 1

```
####################################################################################################
#
#  CS5606 - Quantitative Data Analysis
#   Emma Luk
#   1830215@brunel.ac.uk
#
##########################################################################################################

#uploading data set
rm(list=ls())
library(aod)
library(ggplot2)
library(plyr)
library(forcats)
library(interactions)
library(arm)

###################
###PART1
###################
#set the working directory
setwd("D:/back-up/brunel/CS5606 - Quantitative Data Analysis")

Aids2ann <- read.csv("Aids2ann.csv")
#converting the date from julian format to standard format
Aids2ann$diag <- as.Date(Aids2ann$diag,origin="1970-01-01")
Aids2ann$death <- as.Date(Aids2ann$death,origin="1970-01-01")
#number of days he/she was alive after diagnosis
Aids2ann$diagnosisminusdeath <- Aids2ann$death- Aids2ann$diag
#convert them into number of days
Aids2ann$diagnosisminusdeath<-as.numeric(Aids2ann$diagnosisminusdeath)

#convert outcome variable into factor variable
#Aids2ann$outcome<-factor(Aids2ann$outcome,labels=c("Survived","Died"))
levels(Aids2ann$outcome)
table(Aids2ann$outcome)
#data structure
str(Aids2ann)
#tabulating the cateogorical variable
table(Aids2ann$state)
table(Aids2ann$sex)
table(Aids2ann$outcome)
table(Aids2ann$T.categ)

table1 = table(Aids2ann$outcome,Aids2ann$state)
round(prop.table(table1,2),2)

table2 = table(Aids2ann$outcome,Aids2ann$sex)
round(prop.table(table2,2),2)

table3 = table(Aids2ann$state,Aids2ann$sex)
round(prop.table(table3,2),2)
```

```r
#histogram and boxplot of age across different categories
plot1<- ggplot(Aids2ann,
       aes(y = age, x = state,color= sex)) +
  geom_histogram(stat="identity")
plot1
plot2 <- ggplot(Aids2ann, aes(x = sex, y = age)) + geom_boxplot()
plot2

plot3<- ggplot(Aids2ann, aes(x = diagnosisminusdeath)) + geom_histogram(stat="bin",binwidth = 50)
plot3

#distribution of age
plot4<- ggplot(Aids2ann, aes(age,color=sex))+ facet_wrap(Aids2ann$sex) +
  geom_histogram(bins=30)
plot4

#qqplot of age
plot5 <-ggplot(Aids2ann, aes(sample=age)) +
  stat_qq(aes(color = sex)) +
  scale_color_manual(values = c("#00AFBB", "#E7B800"))+
  labs(y = "Sample Quantile")
plot5


#summary stats
summary(Aids2ann)

#summary stats
Aids2ann<-Aids2ann[,2:11]
summary(Aids2ann)
########################
splom(Aids2ann)
library(lattice)
splom(~Aids2ann[ ,2:8])
splom(~Aids2ann[ ,3:8])
splom(~Aids2ann[ ,2:10])
plotmatrix(Aids2ann[ ,2:10])

#$diag
summary(Aids2ann$diag)
length(Aids2ann$diag)
quantile(Aids2ann$diag)
range_Diag <- max(Aids2ann$diag) - min(Aids2ann$diag)
range_Diag
IQR_Diag <- quantile(Aids2ann$diag, .75) - quantile(Aids2ann$diag,.25)
IQR_Diag
sd(Aids2ann$diag)
```

```
#$death
summary(Aids2ann$death)
length(Aids2ann$death)
quantile(Aids2ann$death)
range_Death <- max(Aids2ann$death) - min(Aids2ann$death)
range_Death
IQR_Death <- quantile(Aids2ann$death, .75) - quantile(Aids2ann$death,.25)
IQR_Death
sd(Aids2ann$death)
mode(Aids2ann$death)

result <- getmode(Aids2ann$death)
print(result)

#$Age
summary(Aids2ann$age)
length(Aids2ann$age)
quantile(Aids2ann$age)
range_Age <- max(Aids2ann$age) - min(Aids2ann$age)
range_Age
IQR_Age <- quantile(Aids2ann$age, .75) - quantile(Aids2ann$age,.25)
IQR_Age
sd(Aids2ann$age)
la
mfv(Aids2ann$age)
mode(Aids2ann$age)

#$year
summary(Aids2ann$year)
length(Aids2ann$year)
quantile(Aids2ann$year)
range_Year <- max(Aids2ann$year) - min(Aids2ann$year)
range_Year
IQR_Year <- quantile(Aids2ann$year, .75) - quantile(Aids2ann$year,.25)
IQR_Year
sd(Aids2ann$year)
mode(Aids2ann$year)
mfv(Aids2ann$outcome)

#$diagnosisminusdeath
summary(Aids2ann$diagnosisminusdeath)
length(Aids2ann$diagnosisminusdeath)
quantile(Aids2ann$diagnosisminusdeath)
range_diagnosisminusdeath <- max(Aids2ann$diagnosisminusdeath) - min(Aids2ann$diagnosisminusdeath)
range_diagnosisminusdeath
IQR_diagnosisminusdeath <- quantile(Aids2ann$diagnosisminusdeath, .75) - quantile(Aids2ann$diagnosisminusdeath,.25)
IQR_diagnosisminusdeath
sd(Aids2ann$diagnosisminusdeath)
mode(Aids2ann$diagnosisminusdeath)
```

```
# Categorical variables
length(Aids2ann$state)
length(Aids2ann$sex)
length(Aids2ann$status)
length(Aids2ann$T.categ)
length(Aids2ann$outcome)


summary(Aids2ann)
str(Aids2ann)
```

Task 2

```
################################
###****#PART2**
########################
#Is Age normally distributed?
shapiro.test(Aids2ann$age)
library(nortest)
normalitytest <- ad.test(Aids2ann$age)
# Anderson-Darling normality test - null hypothesis of normality is rejected at 5% level.

#Are the average age of male and female equal?
t.test(Aids2ann$age[Aids2ann$sex=="M"], Aids2ann$age[Aids2ann$sex=="F"])

#Is the average age same across both outcomes?
t.test(Aids2ann$age[Aids2ann$outcome=="Survived"], Aids2ann$age[Aids2ann$outcome=="Died"])

#Is the average age same across states?
fit = lm(age~state,Aids2ann)
anova(fit)
#conclusion: The null hypothesis that average age is same across all states cannot be rejected at 5% level.
#one of assumptions of t-test and anova test is that continous variable (age) is normality distributed. However, that assumption seems to be violated as per Anderson-Darling normality test

# test of independence between status and state
chisq.test(Aids2ann$outcome,Aids2ann$state)
#Ho: Status and state are independent
#We will reject Ho at 5% level.

#test of independence between status and sex
chisq.test(Aids2ann$outcome,Aids2ann$sex)
#Ho: Status and sex are independent
#We are unable to reject Ho at 5% level but the same can be done at 10% level.

#test of independence between state and sex
chisq.test(Aids2ann$state,Aids2ann$sex)
#Ho: State and sex are independent
#We reject Ho at 5% level.

#correlation between number of days one surives and their age
round(cor(Aids2ann$diagnosisminusdeath,Aids2ann$age),2)
```

Task 3

```
############################
#Part 3
############################
mylogit <- glm(outcome~sex+age+state+age*sex+state*age,data = Aids2ann, family = "binomial")
summary(mylogit)
interact_plot(mylogit, pred = age, modx = state,y.label = "Predicted probability of death")
interact_plot(mylogit, pred = age, modx = sex, y.label = "Predicted probability of death")

#log odds of the regression coefficients
exp(coef(mylogit))

#confidence interval of the coefficients - didn't work on my R
round(confint(mylogit),2)


#residual plot
plot(predict(mylogit),residuals(mylogit))

#Does model as a whole fits better than intercept only model? Likelihood Ratio Test

mylogitconstant <- glm(outcome~1,data = Aids2ann, family = "binomial")
summary(mylogitconstant)
#compute the test statistic which will be 35.673 with 5 degrees of freedom
anova(mylogitconstant,mylogit)
#p-value is less than 1% and 5% => We reject H0.
1-pchisq(35.673,df=9)
#Aliter: with(mylogit, pchisq(null.deviance - deviance, df.null - df.residual, lower.tail = FALSE))


binnedplot(fitted(mylogit),
           residuals(mylogit, type = "response"),
           nclass = NULL,
           xlab = "Expected Values",
           ylab = "Average residual",
           cex.pts = 0.6,
           col.pts = 1,
           col.int = "red")
```

References

Ripley, B. D. and Solomon, P. J. (1992) *A note on Australian AIDS survival*. Available at: https://pdfs.semanticscholar.org/7d23/36da875505e66ae983a271ee6cd83ce42677.pdf (Accessed: 26 December 2019).