Statistics 100 Summer 2021 Final Exam Solutions

Kartik Srikumar

Useful Formatting Notes.

It is best to enclose any in-line equations, including math operators, within two \$ symbols, e.g. $0.40 + 0.02 = 0.42$. The following operators may be useful: $\times$, $\cdot$, $\cap$, $\cup$, $\neq$, $\leq$, $\geq$, μ, σ, α, β, χ, and δ. To create a superscript, A^C. To create a subscript, A_C. To use the square root symbol, $\sqrt{x}$. To create a "hat", use $\hat{y}$ or $\widehat{caries}$.

To typeset fractions, use the command $\frac{numerator}{denominator}$.

For your convenience, the following syntax is given:

$$P(D|T^+) = \frac{P(D) \cdot P(T^+|D)}{P(T^+)} = \frac{P(T^+|D) \cdot P(D)}{[P(T^+|D) \cdot P(D)] + [P(T^+|D^C) \cdot P(D^C)]}$$

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|A^C)P(A^C)}$$

$$P(X = k) = \frac{e^{-\lambda} \cdot \lambda^k}{k!}$$

$$t = \frac{\overline{x} - \mu_0}{s/\sqrt{n}}$$

$$\overline{x} \pm \left(t_{df,\ 1-(\alpha/2)} \times \frac{s}{\sqrt{n}}\right)$$

$$t = \frac{(\overline{x}_1 - \overline{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

$$(\overline{x}_1 - \overline{x}_2) \pm \left(t_{df,\ 1-(\alpha/2)} \times \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}\right)$$

$$t = \frac{\overline{d} - \delta_0}{s_d/\sqrt{n}}$$

$$\overline{d} \pm \left(t_{df,1-(\alpha/2)} \times \frac{s_d}{\sqrt{n}}\right)$$

$$\log\left[\frac{\hat{p}(\text{status} = \text{lived}|\text{age, cpr, cre})}{1 - \hat{p}(\text{status} = \text{lived}|\text{age, cpr, cre})}\right] = b_0 + b_1(age) + b_2(cpr_{yes}) + b_3(cre_{>2.0})$$

To make large brackets or parentheses around a fraction...

$$\left[\frac{num}{denom}\right]$$
$$\left(\frac{num}{denom}\right)$$

There are three ways to define a math environment:

- Using the $$ syntax is useful for short expressions within a text explanation.
- The
  $$yourmathhere$$
  syntax is useful for entering centered, single-line equations.
- An align* environment is useful for a series of equations, such as showing the steps for Bayes' Rule. Remember to place an & symbol where the equations line up, such as before or after the = sign in each line. Use \ to denote a new line.

Sample align* environment:

$$\begin{aligned} yourmathhere &= yourmathhere \\ &= yourmathhere \\ &= yourmathhere \end{aligned}$$

Problem Scoring		
Problem	Point Value	Points Scored
1	60	
2	60	
3	80	
Total	200	

I confirm that I have worked independently on this take-home exam, except for any assistance I may have received from the teaching staff with technical issues. All the work is my own, and I have not collaborated in any way with fellow students.

Signature: Kartik K Srikumar

PROBLEM 1: SHORT ANSWER

PART A)

i. The p-value represents the probability of observing the data given the null hypothesis is true, rather than the probability that the null hypothesis is true given the observed data. Similarly (1-p value; 0.99 in this case) is not the probability of the null hypothesis being false.
The p-value of 0.01 means that there is a very low probability of observing the results the researchers did, if the pro-social behaviors of the two SES groups were identical.
Explanation for an general audience:
It is misleading to state that "there is a 99% chance that the rate of returning mis-delivered mail is different between high and low SES households in NYC". The study conducted shows that there is enough evidence that people of higher SES on average are more likely to return the mis-delivered postcard as compared to people from lower SES. More specifically, the researchers have a very low chance of observing the results they did, had the rate of returning mis-delivered mail been the same among the two groups.

ii.

- Confounding factors: Things like busy schedules caused by working many jobs, may cause people of low SES to not return the postcard. Similarly, given that the people defined as high SES by researchers were much higher than the average wealth amount (2.5MM compared to $126,000), it is possible that those people may have domestic help or other resources to return the postcard, making it *easier* for them to do so.

- Generalizing behavior by one Act: Returning one mis-delivered postcard cannot be a sole measure of likelihood of engaging in pro-social behavior. It is possible that the people of low SES, who did not return the postcard are involved in many other pro-social activities, but did not return the postcard.

PART B)

- Let us consider a discrete random variable X, which can take on the values of the number of cards picked by the dealer that match the players picks.
- So, X can take on the values 0, 1, 2, 3, 4, 5, 6

- Let us calculate P(X=4), P(X=5), P(X=6) and P(X= 0 or 1 or 2 or 3)
- The dealer can pick the cards above in a variety of permutations that we have to account for as well using $\frac{n!}{x!(n-x)!}$

$$P(X=4) = \left(\frac{12!}{12!(12-4)!}\right) * \frac{6}{52} * \frac{5}{51} * \frac{4}{50} * \frac{3}{49} * \frac{45}{48} * \frac{44}{47} * \frac{43}{46} * \frac{42}{45} * \frac{41}{44} * \frac{40}{43} * \frac{39}{42} * \frac{38}{41}$$

```
#Calculating P(X=4)
p4=(factorial(12)/(factorial(4)*factorial(8)))*
  (6/52)*(5/51)*(4/50)*(3/49)*(46/48)*(45/47)*(44/46)*
  (43/45)*(42/44)*(41/43)*(40/42)*(39/41)
p4
```

```
## [1] 0.01896503
```

Similarly, we can calculate the probabilities of X taking on the values 5 and 6:

```
#Calculating P(X=5)
p5=(factorial(12)/(factorial(5)*factorial(7)))*
  (6/52)*(5/51)*(4/50)*(3/49)*(2/48)*(46/47)*(45/46)*
  (44/45)*(43/44)*(42/43)*(41/42)*(40/41)
p5
```

```
## [1] 0.001556105
```

```
#Calculating P(X=6)
p6=(factorial(12)/(factorial(6)*factorial(6)))*
  (6/52)*(5/51)*(4/50)*(3/49)*(2/48)*(1/47)
p6
```

```
## [1] 4.53864e-05
```

Now, the P(X= 0 or 1 or 2 or 3) is $1-(P(X=4)+P(X=5)+P(X=6))$

```
#Calculating P(X= 0 or 1 or 2 or 3)
pelse=1-(p4+p5+p6)
pelse
```

```
## [1] 0.9794335
```

Let's put the information in a table:

$X=x$	$P(X=x)$
4	0.01896503
5	0.001556105
6	$4.53864e-05$
0, 1, 2, 3	0.9794335

Now, we know that the profits when the dealer picks 4, 5, 6 and 'all else' cards, so we can think of a new random variable P, which will have identical probabilities of taking on values of the profit corresponding to X as seen below:

$P=x$	$P(P=x)$
95 Dollars	0.01896503
995 Dollars	0.001556105
9995 Dollars	$4.53864e-05$
-5 Dollars	0.9794335

The Expected value of P, or E(P) = P(X=95) × 95 + P(X=995) × 995 + P(X=9995) × 9995 - P(X=-5) × -5 =

```
(0.01896503*95)+(0.001556105*995)+(4.53864e-05*9995)+(0.9794335*-5)
```

```
## [1] -1.093528
```

From the Expected value of P, we can see that in general, on average, my friend would lose $1.09 and therefore I would ask him to avoid playing.

PART C)

- Let Event A be the event that an applicant is in the 95th percentile for academics.
- Let Event B be the event that an applicant is in the 95th percentile for athletics.
- Information given in the problem:
  P(A)=0.05
  P(B)=0.05
  Event C = A ∪ B
- Therefore:
  P(C)=P(A ∪ B)=P(A)+P(B)-P(A ∩ B)
  So, P(C)=$0.05 + 0.05 - (0.05 * 0.05)\ldots$since A and B are independent, P(A ∩ B)=P(A)*P(B)
  Computing the above, we get P(C)=0.0975
- Using Conditional Probability, we know that:
  $P(A|C) = \dfrac{P(A \cap C)}{P(C)}$
  Since $A \cap C = A \cap (A \cup B) = A\ldots$since A and B are independent
  $P(A|C) = \dfrac{P(A)}{P(C)} = \dfrac{0.05}{0.0975}$=**0.5128205**
  Similarly, we can compute the P(B|C), which comes out to **0.5128205**
- Also, P(A∩B|C)=$\dfrac{P((A \cap B) \cap (A \cup B))}{P(C)} = \dfrac{P(A \cap B)}{P(C)} = \dfrac{0.0025}{0.0975}$=**0.02564103**

As seen above, $P(A \cap B|C) \neq P(A|C) * P(B|C)\ldots$since $0.02564103 \neq 0.2629849$
We can conclude that conditional on C, are A and B **NOT** independent and academic ability and athletic ability are **NOT** independent among the admitted applicants.

PART D)

i.

```
#load the data
load("datasets/spam.Rdata")

#define error_rates function
error_rates = function(model, test_data, train_data, test_y, train_y){
#classify train set using model
phat_train = predict(model, newdata = train_data ,type = 'response')
yhat_train = (phat_train >= 0.50)
#classify test set using model
phat_test = predict(model, newdata = test_data, type = 'response')
yhat_test = (phat_test >= 0.50)
out = c(train = mean(yhat_train != train_y),
test = mean(yhat_test != test_y))
return(out)
}
```

```r
#Assign to random groups
set.seed(2021)
test.index = sample(1:5, size = nrow(spam), replace = TRUE)

#Test error array
test_errors = array(0, dim= c(4, 5))

for(k in 1:5){

  #Defining test/train sets
  test.set = spam[test.index == k, ]
  train.set = spam[test.index != k, ]

  #define models
  lm.reply = glm(y ~ reply + word_count + caps + exclaim, data = train.set,
                 family = binomial(link = "logit"))
  lm.buy = glm(y ~ buy + word_count + caps + exclaim, data = train.set,
                 family = binomial(link = "logit"))
  lm.win = glm(y ~ win + word_count + caps + exclaim, data = train.set,
                 family = binomial(link = "logit"))
  lm.send = glm(y ~ send + word_count + caps + exclaim, data = train.set,
                 family = binomial(link = "logit"))

  #compute error rates and store in array
  test_errors[1, k] = error_rates(lm.reply, test.set, train.set,
  test.set$y, train.set$y)[2]
  test_errors[2, k] = error_rates(lm.buy, test.set, train.set,
  test.set$y, train.set$y)[2]
  test_errors[3, k] = error_rates(lm.win, test.set, train.set,
  test.set$y, train.set$y)[2]
  test_errors[4, k] = error_rates(lm.send, test.set, train.set,
  test.set$y, train.set$y)[2]

}

test_errors = cbind(test_errors, rowMeans(test_errors))
rownames(test_errors) = c("reply", "buy", "win", "send")
colnames(test_errors) = c("k = 1", "k = 2", "k = 3", "k = 4", "k = 5", "Avg")
test_errors
```

```
##           k = 1     k = 2     k = 3     k = 4     k = 5       Avg
## reply 0.1409052 0.1217472 0.1528777 0.1427328 0.1484962 0.1413518
## buy   0.1511529 0.1291822 0.1663669 0.1488251 0.1513158 0.1493686
## win   0.1485909 0.1217472 0.1582734 0.1436031 0.1475564 0.1439542
## send  0.1468830 0.1263941 0.1654676 0.1479547 0.1522556 0.1477910
```

Using the 5-fold cross validation approach, we use a fifth of the data as the test data (80-20 split),

but use a different 20% to test every time. This enables us to get a better understanding of the test error of a certain set of predictors.
We perform 5-fold CV by looping through the 5 folds of the data splits, applying each of the four models we are interested in comparing (adjusting for word count, caps and exclamation points every time) and then creating a matrix of test erors and corresponding average test errors.
We observe that the model using "reply" has the lowest average test error amongst the 4 models, with an average test error of 0.141. It is worth noting that the others' average test errors were not drastically different, indicating that they are all similarly predictive.
The AIC scores below also indicate that the model using "reply", relatively has the most parsimonious fit, with AIC=2980.66.

```
lm.reply$aic
```

```
## [1] 2980.661
```

```
lm.buy$aic
```

```
## [1] 3058.322
```

```
lm.win$aic
```

```
## [1] 3012.244
```

```
lm.send$aic
```

```
## [1] 3054.362
```

ii. Here, the Type-I error or False Positives are those text messages incorrectly classified as spam when in reality they are not spam. The Type-II error or False Negatives are those text messages incorrectly classified as not-spam when in reality they are spam.

iii.

```
#random group assignment
set.seed(2021)
test.index = sample(0:1, prob = c(0.80, 0.20),
                    size = nrow(spam), replace = TRUE)
#define test and train sets
test.set = spam[test.index == 0, ]
train.set = spam[test.index == 1, ]
#Fit model
lm.reply_2 = glm(y ~ reply + word_count + caps +
                   exclaim, data = train.set,
                family = binomial(link = "logit"))

#Compute Type I & II errors at 0.15 cutoff
p.hat = predict(lm.reply_2, newdata = test.set,
                type = "response")
y.hat = (p.hat > 0.15)
error.table = table(Predict = y.hat, Observe = test.set$y)
rownames(error.table) = c("Not Spam", "Spam")
prop.table(error.table, 2)
```

```
##           Observe
## Predict             0         1
##   Not Spam 0.7965661 0.1815068
##   Spam     0.2034339 0.8184932
```

```
#Compute Type I & II errors at 0.20 cutoff
p.hat = predict(lm.reply_2, newdata = test.set, type = "response")
y.hat = (p.hat > 0.20)
error.table = table(Predict = y.hat, Observe = test.set$y)
rownames(error.table) = c("Not Spam", "Spam")
prop.table(error.table, 2)
```

```
##           Observe
## Predict           0        1
##   Not Spam 0.864204 0.369863
##   Spam     0.135796 0.630137
```

```
#Compute Type I & II errors at 0.10 cutoff
p.hat = predict(lm.reply_2, newdata = test.set, type = "response")
y.hat = (p.hat > 0.10)
error.table = table(Predict = y.hat, Observe = test.set$y)
rownames(error.table) = c("Not Spam", "Spam")
prop.table(error.table, 2)
```

```
##           Observe
## Predict              0          1
##   Not Spam 0.67976067 0.08732877
##   Spam     0.32023933 0.91267123
```

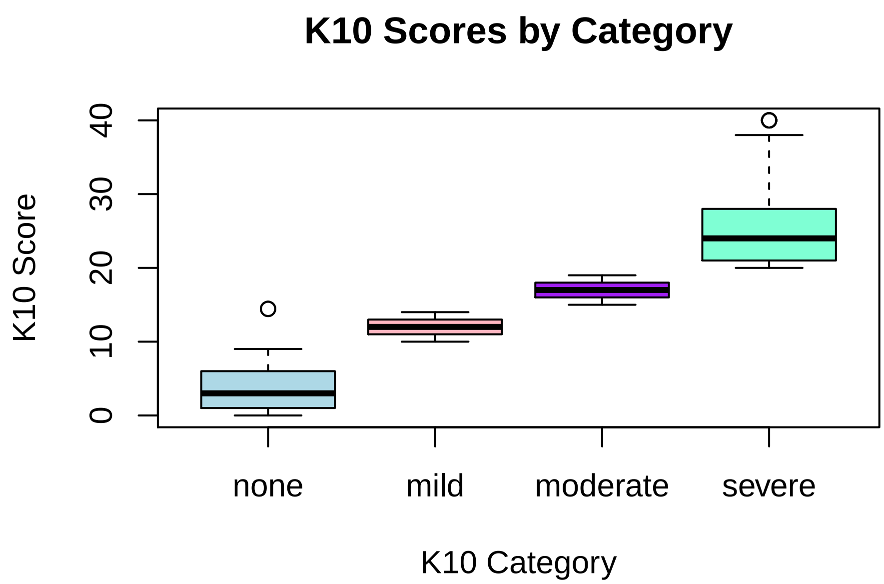
With a 0.15 cutoff the Type I error is ~20.34% and Type-II error is ~18.15%.
With a 0.20 cutoff the Type I error is ~13.57% and Type-II error is ~37%.
With a 0.10 cutoff the Type I error is ~32% and Type-II error is ~8.7%.

Based on these data, it would be preferable to use a 0.20 cutoff since we are more concerned in this setting with minimizing Type-I errors. In other words we want to minimize text messages incorrectly classified as spam when in reality they are not spam since this may lead to users not reading important text messages.

PROBLEM 2: PSYCHOLOGICAL HEALTH

PART A)

```
#load the data
load("datasets/wellbeing.Rdata")

#Plotting K10 distribution
boxplot(wellbeing$k10.score~wellbeing$k10.cat,
        main="K10 Scores by Category", xlab="K10 Category",
        ylab="K10 Score", col=c("light blue", "light pink",
                                "purple", "aquamarine"))
```




```
#Understand number of people in each K10 Category
table(wellbeing$k10.cat)
```

```
##
##     none     mild moderate   severe
##     1251      272      209      267
```

```
#proportion with k10 indicating anxiety/depression(>12)
nrow(subset(wellbeing, k10.score>12))/nrow(wellbeing)
```

```
## [1] 0.2835821
```

As seen above, a majority of the people's scores (more than 50%) reported not having any distress in turn no anxiety or depression. 272 people fell under the "Mild" actegory and 209 and 267 people fell under "moderate" and "severe" respectively. The median score for people who were categorized as "mild" is about 13, for "moderate" approximately 15, and "severe" approximately 25.
The proportion of patients with a K10 score indicative of an anxiety or depressive disorder was 0.283, indicating that about 28% of the sample being studied reported some level of distress in their scores.

PART B)

- H_0 : the proportion of NZ adults experiencing low psychological wellbeing is 0.25.
- H_A : the proportion of NZ adults experiencing low psychological wellbeing differs from 0.25.
- Let $\alpha = 0.05$

```
whoLow=nrow(subset(wellbeing, who.score<13))
total=nrow(wellbeing)

binom.test(x=whoLow, n=total, p = 0.25, alternative = "two.sided")
```

```
##
##  Exact binomial test
##
## data:  whoLow and total
## number of successes = 776, number of trials = 2010, p-value < 2.2e-16
## alternative hypothesis: true probability of success is not equal to 0.25
## 95 percent confidence interval:
##  0.3647167 0.4077555
## sample estimates:
## probability of success
##              0.3860697
```

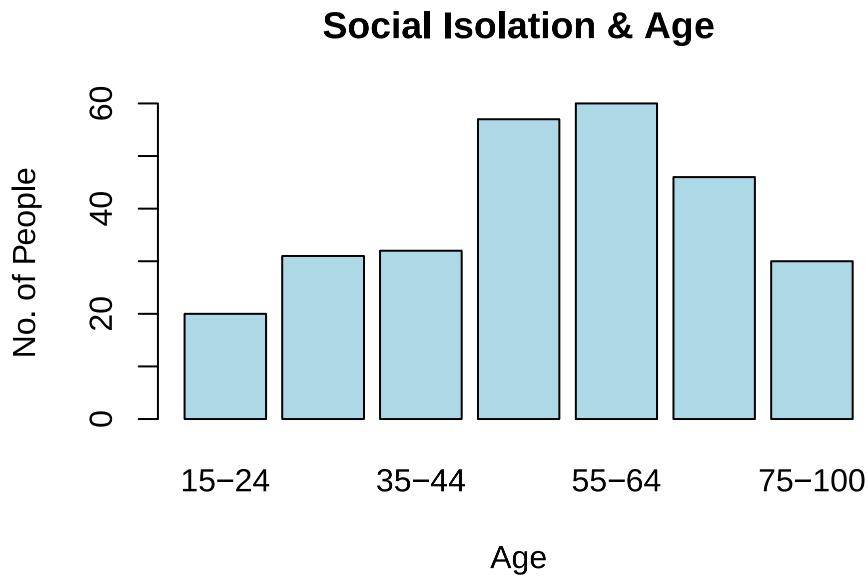
The p-value above is less than $\alpha = 0.05$. Hence there is sufficient evidence to reject the null hypothesis, indicating that the proportion of NZ adults experiencing low psychological wellbeing differs from 0.25.
Further, we are 95% confident that the interval (0.364, 0.407), contains the population proportion of NZ adults experiencing low psychological wellbeing.

Assumptions:
- Independence: It is reasonable to assuem that the observations in the data are independent from one another. The scores of one person do not seem to have an dependence on another based on the data given to us.
- The data being studied in the test are simple random variable from the population. The person can either have a score of less than 13 on the WHO scale or not.
- The data are binomially distributed.

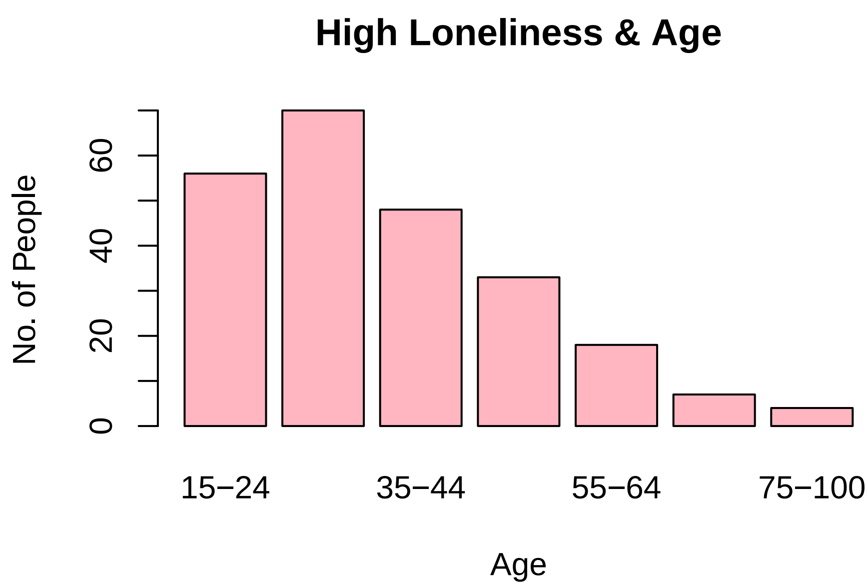
PART C)

i.

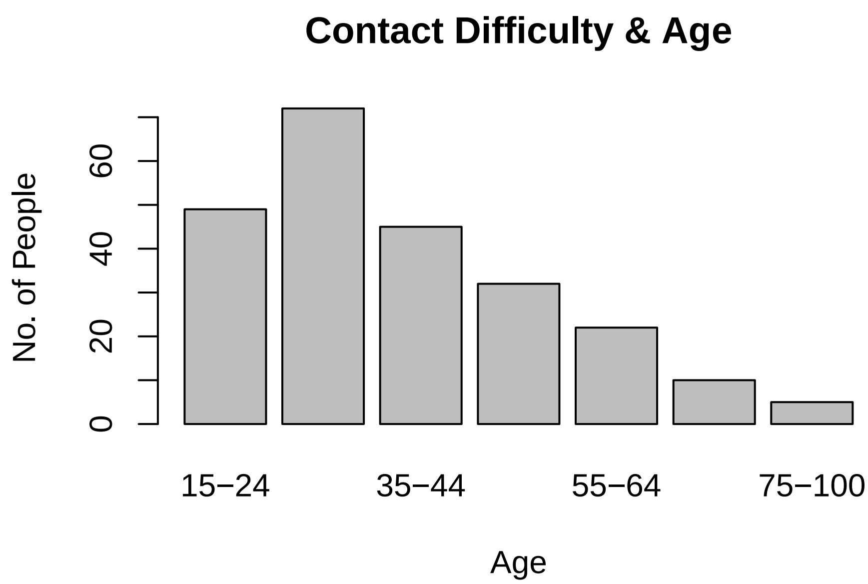
```
#Plotting social isolation and age
wellbeing.alone=subset(wellbeing,
                       lives.alone=="Live by myself")
plot(wellbeing.alone$age, main="Social Isolation & Age",
     xlab="Age", ylab="No. of People", col="light blue")
```




```r
#Plotting loneliness and age
wellbeing.lonely=subset(wellbeing, loneliness=="All/most of the time")
plot(wellbeing.lonely$age, main="High Loneliness & Age",
     xlab="Age", ylab="No. of People", col="light pink")
```




```r
#Plotting contact and age
wellbeing.contact=subset(wellbeing, easy.contact=="Hard")
plot(wellbeing.contact$age, main="Contact Difficulty & Age",
     xlab="Age", ylab="No. of People")
```



The plots above show that for this particular dataset, on average younger people (below the age oif 44) havea greater chance of living with people. People greater than the age of 45 tend to live alone apto the age of around 65 when they again tend to live with people (old-age homes presumably).
The trend for loneliness and age seems to be negative from the plot for this dataset. People of higher ages tend to report lower loneliness.
The trend for contactability and age seems to also be negative from the plot for this dataset. People of higher ages tend to report less difficulaty in perceived ease of maintaining contact with friends and family.

ii.

```
#Create new feature with k10.score > 11
wellbeing$k10.binary = ifelse(wellbeing$k10.score > 11,1,0)
#Convert it to a factor type
wellbeing$k10.binary = factor(wellbeing$k10.binary,
                              levels = c(1,0), labels = c("TRUE", "FALSE"))
#Exploring data
table(wellbeing$k10.binary, wellbeing$age)
```

```
##
##         15-24 25-34 35-44 45-54 55-64 65-74 75-100
##   TRUE    133   191   137    88    48    24      8
##   FALSE   132   213   211   237   256   214    107
```

Independence is reasonable to assume and the grid above shows that none of the counts are extremely small.

```
#Fitting a simple logistic Regression model
glm(k10.binary~age, data = wellbeing, family = binomial(link ="logit"))
```

```
##
## Call:  glm(formula = k10.binary ~ age, family = binomial(link = "logit"),
##     data = wellbeing)
##
```

```
## Coefficients:
## (Intercept)     age25-34     age35-44     age45-54     age55-64     age65-74
##   -0.007547     0.116566     0.439424     0.998271     1.681524     2.195469
##   age75-100
##    2.600934
##
## Degrees of Freedom: 1998 Total (i.e. Null);  1992 Residual
##   (11 observations deleted due to missingness)
## Null Deviance:       2490
## Residual Deviance: 2251  AIC: 2265
```

The statistical analysis indicates that as we move to a higher age bracket the estimated odds of a person having a K10 score greater than 11 increases by a certain amount assuming nothing else (like loneliness, contactability etc.) changes. The estimated increase in odds of a person having a K10 score greater than 11 from the baseline, by age group are:
* Age 25-30: 0.116
* Age 35-44: 0.439
* Age 45-54: 0.998
* Age 55-64: 1.681
* Age 65-74: 2.195
* Age 75-100: 2.60

iii.

```
glm(k10.binary~age+lives.alone+easy.contact+loneliness,
    data = wellbeing,family =binomial(link ="logit"))
```

```
##
## Call:  glm(formula = k10.binary ~ age + lives.alone + easy.contact +
##     loneliness, family = binomial(link = "logit"), data = wellbeing)
##
## Coefficients:
##                         (Intercept)                             age25-34
##                            -1.36342                             -0.06179
##                            age35-44                             age45-54
##                             0.08917                              0.56247
##                            age55-64                             age65-74
##                             1.13435                              1.56281
##                           age75-100           lives.aloneLive with others
##                             2.13143                             -0.39687
##                   easy.contactHard            easy.contactHave not tried
##                            -0.58071                             -0.16795
##   easy.contactNeither easy nor hard            lonelinessNone of the time
##                            -0.62611                              3.69735
## lonelinessSome/a little of the time
##                             1.94760
##
## Degrees of Freedom: 1997 Total (i.e. Null);  1985 Residual
##   (12 observations deleted due to missingness)
```

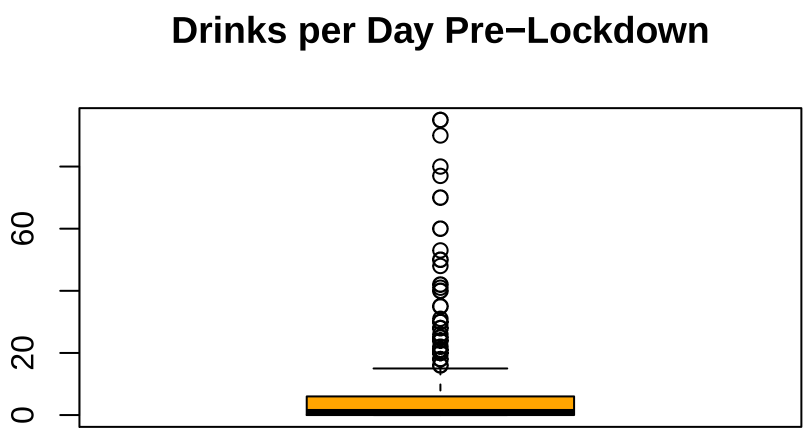
```
## Null Deviance:       2489
## Residual Deviance: 1805  AIC: 1831
```

The model adjusting for loneliness, social isolation and contactability has an AIC score of 1831 as compared to the previous model that has an AIC score of 2265. This indicates that this model is more parsimonious.Despite the penalty that Akaike's Information Criterion applies to the additional predictors, the model still is more parsimonious.

PART D)

i.

```
#Plotting Drinks per Day Pre-Lockdown
boxplot(wellbeing$pre.alcohol, main="Drinks per Day Pre-Lockdown",
        col="orange")
```




```
nrow(subset(wellbeing, pre.alcohol>20))
```

```
## [1] 96
```

```
nrow(subset(wellbeing, pre.alcohol==0))
```

```
## [1] 791
```

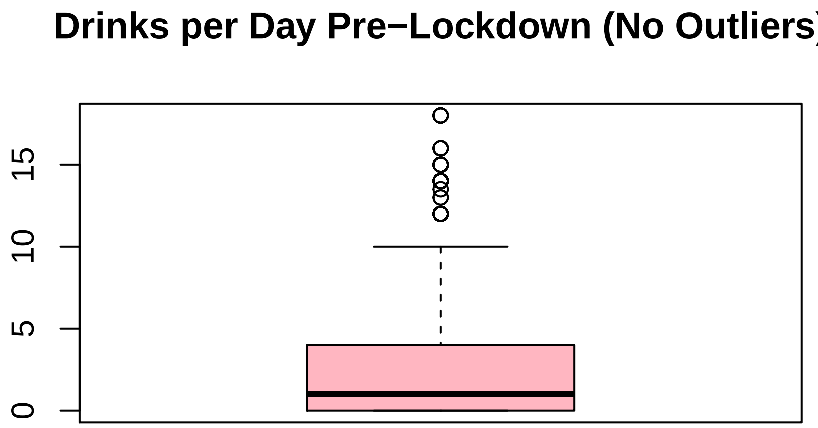
```
nrow(subset(wellbeing, pre.alcohol>15 & pre.alcohol<20))
```

```
## [1] 12
```

```
nrow(subset(wellbeing, pre.alcohol>10 & pre.alcohol<15))
```

```
## [1] 85
```

```
#Subsetting data removing people who consume more than 20/day
wellbeing.noOutliers=subset(wellbeing, pre.alcohol<20)

#Plotting Drinks per Day Pre-Lockdown w/o outliers
boxplot(wellbeing.noOutliers$pre.alcohol,
        main="Drinks per Day Pre-Lockdown (No Outliers)",
        col="light pink")
```



The number of drinks consumed per day varies greatly with 96 people consuming more than 20 drinks a day. Nearly a third of the sample do not consume alcohol. The first boxplot shows the distribution before outliers were removed. Outliers have been defined as more than 20 drinks per day. The second boxplot depicts a more interpretable visual. We see that most people in the dataset consume 0-5 drinks per day, with the median being 3 drinks per day.

ii. It is reasonable to remove outliers from this dataset. One reason for this is that people who have a very high consumption of daily alcohol will likely not change their intake during lockdown. In other words, lockdown will likely not have a significant impact on their consumption. It is reasonable to define outliers here as people consuming greater than 20 drinks a day. This is a very high amount of alcohol by most standards.

iii.

- Let δ be the population mean of the difference in alcohol consumption for people before and after lockdown.
- $H_0 : \delta = 0$, there is no difference in mean alcohol consumption for people before and after lockdown.
- $H_A : \delta \neq 0$, there is a difference in mean alcohol consumption for people before and after lockdown.
- Let $\alpha = 0.05$

```
#Conducting paired t-test since the same people are
##being measured before and after event
t.test(wellbeing.noOutliers$pre.alcohol, wellbeing.noOutliers$during.alcohol,
       alternative ="two.sided",paired =TRUE)
```

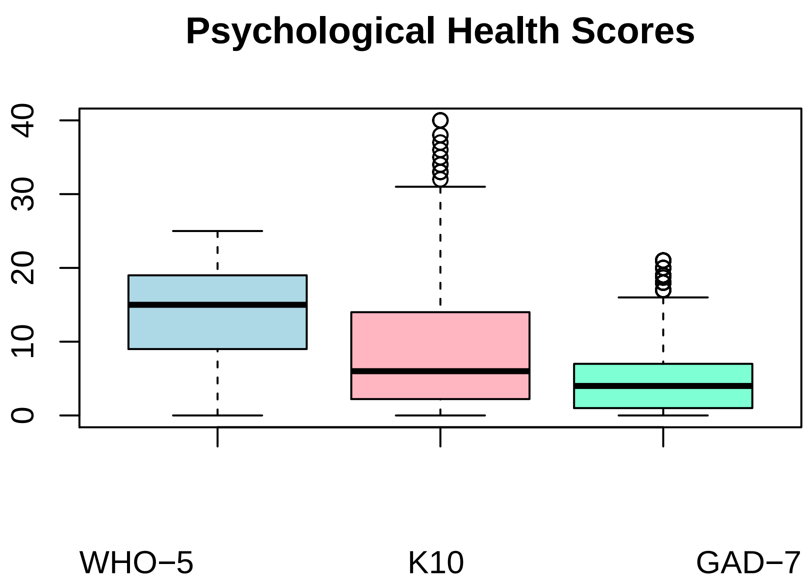
```
##
##  Paired t-test
##
## data:  wellbeing.noOutliers$pre.alcohol and wellbeing.noOutliers$during.alcohol
## t = -5.5884, df = 1884, p-value = 2.626e-08
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.9480988 -0.4555087
```

```
## sample estimates:
## mean of the differences
##              -0.7018037
```

- As we can see p-value=2.626e-08, which is less than $\alpha = 0.05$, and so there is sufficient evidence to reject the null hypothesis, indicating that there is sufficient evidence that the mean alcohol consumption for people before and after lockdown is not equal and on average people consumed more alcohol during lockdown than before it.

PART E)

```
boxplot(wellbeing$who.score, wellbeing$k10.score,
        wellbeing$gad.score, main="Psychological Health Scores",
        xlab=c("WHO-5                        K10                          GAD-7"),
        col=c("light blue", "light pink","aquamarine"))
```




```
summary(wellbeing)
```

```
##      age                 gender                ethnicity
##  15-24 :269   Female         :1063   Asian          : 256
##  25-34 :407   Gender diverse:    6   European/Other:1231
##  35-44 :349   Male           : 941   Maori          : 408
##  45-54 :327                          Pacific        : 115
##  55-64 :304
##  65-74 :239
##  75-100:115
##            lives.alone                    happiness.bubble
##  Live by myself   : 276   Dissatisfied or neither: 374
##  Live with others:1734   Satisfied              :1636
##
##
##
```

```
##
##
##                easy.contact                      loneliness
##  Easy                 :1476   All/most of the time     :236
##  Hard                 : 235   None of the time         :776
##  Have not tried       :  36   Some/a little of the time:997
##  Neither easy nor hard: 263   NA's                     :  1
##
##
##
##                  employment                 work.type
##  I am a business owner :  31   Not essential worker:772
##  I am retired          : 329   Not in workforce    :845
##  I am self-employed    : 112   Yes essential worker:393
##  I have never had a job:  65
##  No, I don't have a job: 451
##  Yes, I have a job     :1022
##
##               positive     pre.alcohol     during.alcohol      who.score
##  Awaiting results:   8   Min.   : 0.000   Min.   :  0.000   Min.   : 0.00
##  Negative        :  75   1st Qu.: 0.000   1st Qu.:  0.000   1st Qu.: 9.00
##  Positive        :   9   Median : 1.000   Median :  1.000   Median :15.00
##  NA's            :1918   Mean   : 4.635   Mean   :  5.167   Mean   :13.92
##                          3rd Qu.: 6.000   3rd Qu.:  6.000   3rd Qu.:19.00
##                          Max.   :95.000   Max.   :120.000   Max.   :25.00
##                          NA's   :4        NA's   :4         NA's   :10
##    gad.score        k10.score         k10.cat     silver.personal
##  Min.   : 0.000   Min.   : 0.000   none    :1251   No  :1107
##  1st Qu.: 1.000   1st Qu.: 2.222   mild    : 272   Yes : 898
##  Median : 4.000   Median : 6.000   moderate: 209   NA's:   5
##  Mean   : 5.079   Mean   : 8.994   severe  : 267
##  3rd Qu.: 7.000   3rd Qu.:14.000   NA's    :  11
##  Max.   :21.000   Max.   :40.000
##  NA's   :2        NA's   :11
##  silver.society k10.binary
##  No  :1249      TRUE : 629
##  Yes : 756      FALSE:1370
##  NA's:   5      NA's :  11
##
##
##
##
```

- The first step is to see if our **sample data is reasonably representative** of the population. Looking at the summary of the dataset above, we see that basic demographic variables are well distributed and seem reasonably representative of the population.

```r
nrow(subset(wellbeing, wellbeing$who.score<13))/nrow(wellbeing)
```

```
## [1] 0.3860697
```

```r
nrow(subset(wellbeing, wellbeing$k10.score>12))/nrow(wellbeing)
```

```
## [1] 0.2835821
```

```r
nrow(subset(wellbeing, wellbeing$gad.score>15))/nrow(wellbeing)
```

```
## [1] 0.03034826
```

- Additionally, all the proportions above are in a fairly narrow range(0.28-0.38), indicating that even using different rating criterion of the three different scores, the proportion of people reporting anxiety/depression is fairly consistent in the dataset. This further indicates that at a high level, the sample is representative of the population.
- Further, when we look at the numbers above; the proportion of people having anxiety and/or depression, calculated for the three scores seem appropriate and representative of the population. According to many articles online, including the one cited below, 50-80% of New Zealanders experience "mental distress or addiction challenges". While the numbers calculated above are slightly lower, they indicate the general prevalence of anxiety and depression among the population of New Zealand.

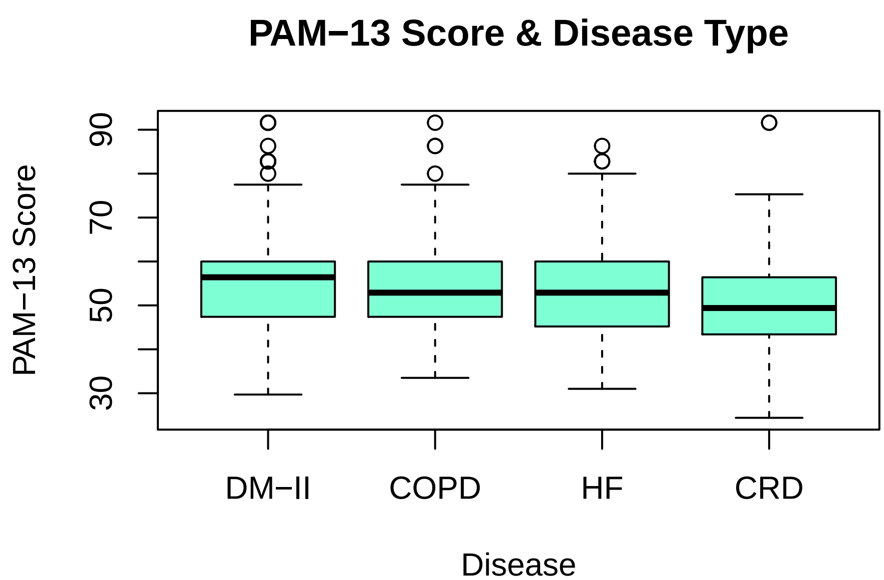
[Link] (https://www.theguardian.com/world/2018/dec/04/crisis-in-new-zealand-health-services-as-depression-and-anxiety-soar)

PROBLEM 3: SELF-MANAGEMENT

PART A)

i.

```r
#load the data
load("datasets/self_manage.Rdata")

plot(self.manage$pam.score~self.manage$disease,
     main="PAM-13 Score & Disease Type", xlab="Disease",
     ylab="PAM-13 Score", col="aquamarine")
```



From the plot above we see that the majority of participants have a PAM-13 score between 45 and 60 for all 4 disease types with the median being between 50 and 60. No significant difference is discernible from the plot for scores between different disease types. Few participants with CRD had the lowest scores overall.

ii.

- Hypothesis:
  H0: There is no association between PAM-13 score and disease type.
  HA: There is an association between PAM-13 score and disease type.

- $\alpha = 0.05$

- Model:

```r
summary(lm(pam.score~disease, data=self.manage))
```

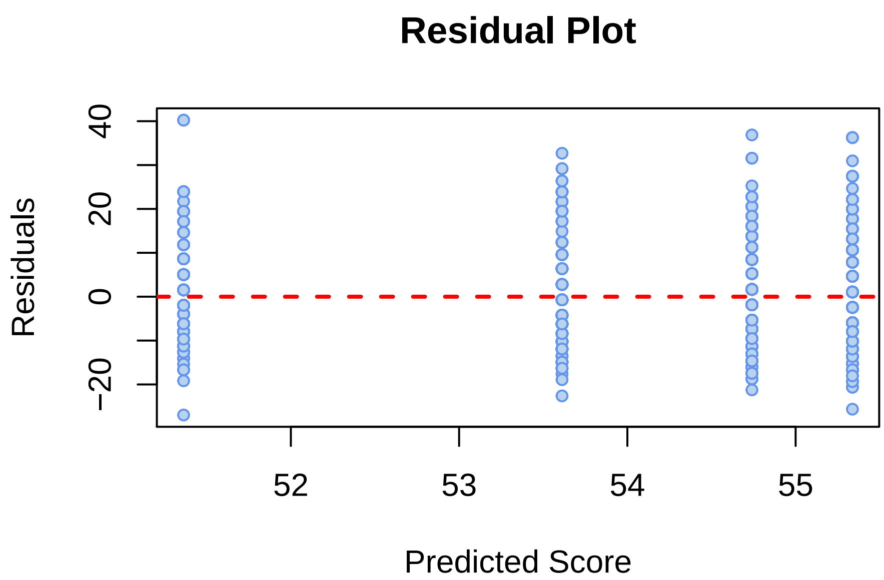
```
##
## Call:
## lm(formula = pam.score ~ disease, data = self.manage)
##
## Residuals:
```

```
##      Min       1Q   Median       3Q      Max
## -26.963  -7.938  -1.840   5.260  40.237
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   55.3377     0.5216 106.101  < 2e-16 ***
## diseaseCOPD   -0.5973     0.8172  -0.731   0.4650
## diseaseHF     -1.7260     0.8870  -1.946   0.0519 .
## diseaseCRD    -3.9751     0.8923  -4.455 9.21e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.71 on 1150 degrees of freedom
## Multiple R-squared:  0.01829,    Adjusted R-squared:  0.01573
## F-statistic: 7.142 on 3 and 1150 DF,  p-value: 9.391e-05
```

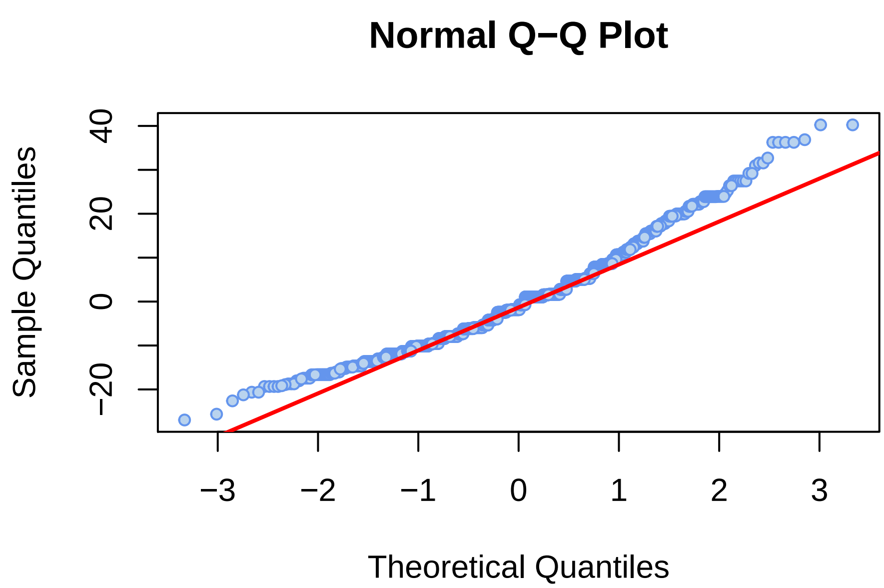
As we see above the overall p-value for the model is lower than $\alpha = 0.05$, so there is enough evidence to reject the null hypothesis, indicating that there is an association between PAM-13 score and disease type.
The p-values for COPD and HF are greater than 0.05 indicating that these two diseases likely have no or low association to the PAM-13 score.
The model shows that compared to the mean baseline PAM-13 score for DM-II patients, which is 55.337, the mean PAM-13 score for COPD patients is on average 0.597 less, for HF patients is on average 1.726 less and for CRD patients is 0.892 less.

- Assumptions:

```
model1=lm(pam.score~disease, data=self.manage)

#Constant Variance
plot(residuals(model1)~predict(model1),main = "Residual Plot",
     xlab = "Predicted Score", ylab = "Residuals",
     pch = 21, col = "cornflowerblue", bg = "slategray2",cex = 0.75)
abline(h = 0, lty = 2, lwd = 2, col = "red")
```




```
#Normality of Residuals
qqnorm(residuals(model1),pch = 21, col = "cornflowerblue",
       bg = "slategray2", cex = 0.75)
qqline(residuals(model1),col = "red", lwd = 2)
```



From the plots above we see that there is constant variance, but the residuals deviate from normality at the lower and upper tails in the upward direction. This indicates that the linear model is not the best choice in thsi problem.
It is also reasonable to assume independence of the observations. No patient's data is dependent on another's as far as we know.

iii.

```
#participants by disease
d1=subset(self.manage, disease=="DM-II")
d2=subset(self.manage, disease=="COPD")
```

```r
d3=subset(self.manage, disease=="HF")
d4=subset(self.manage, disease=="CRD")

print("DM-II")
```

```
## [1] "DM-II"
```

```r
table(d1$pam.cat)
```

```
##
## Level 1 Level 2 Level 3 Level 4
##      96     107     163      56
```

```r
print("COPD")
```

```
## [1] "COPD"
```

```r
table(d2$pam.cat)
```

```
##
## Level 1 Level 2 Level 3 Level 4
##      66      86      96      42
```

```r
print("HF")
```

```
## [1] "HF"
```

```r
table(d3$pam.cat)
```

```
##
## Level 1 Level 2 Level 3 Level 4
##      67      67      61      28
```

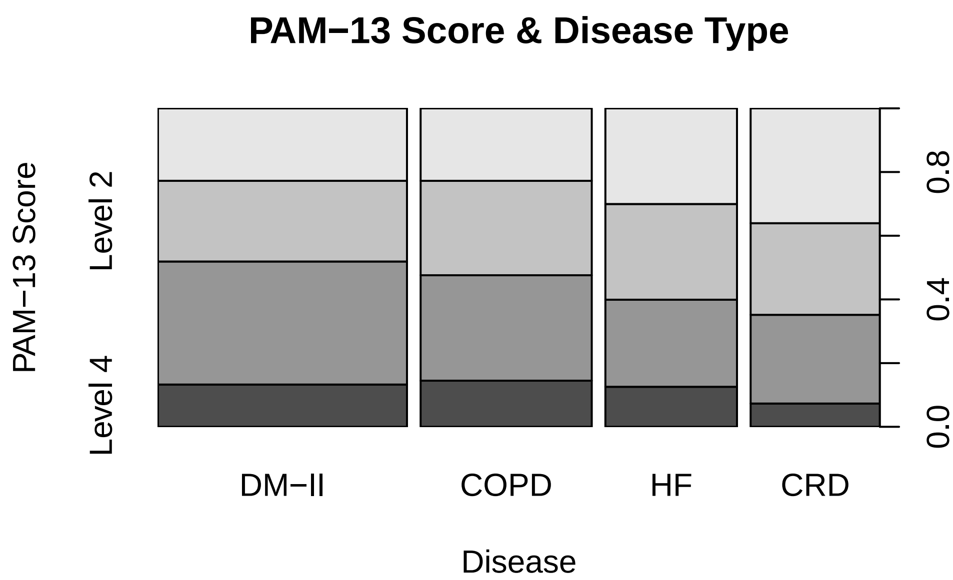
```r
print("CRD")
```

```
## [1] "CRD"
```

```r
table(d4$pam.cat)
```

```
##
## Level 1 Level 2 Level 3 Level 4
##      79      63      61      16
```

```r
#Plotting the summary
plot(self.manage$pam.cat~self.manage$disease,
     main="PAM-13 Score & Disease Type", xlab="Disease",
     ylab="PAM-13 Score")
```



We see that the proportion of participants in the Level 1 bracket is higher for DM-II and COPD and lowest for CRD. The CRD group seem to have the highest proportion of people in the Level 1 bracket, followed by HF. Overall there seems to be a visible downward trend in proportion of people in higher PAM-13 levels as we move from DM-II to COPD to HF and to CRD.

iv.

- Hypothesis:
  H0: There is no association between PAM-13 level and disease type.
  HA: There is an association between PAM-13 level and disease type.
- $\alpha = 0.05$

- Assumptions for chi-square test:
  There is independence between observations since it is reasonable to assume independence of the observations. No patient's data is dependent on another's as far as we know.
  We can see below that each expected cell count must be greater than or equal to 10. For tables larger than 2X2, it is appropriate to use the test if no more than 1/5 of the expected counts are less than 5, and all expected counts are greater than 1.

```
table(self.manage$disease, self.manage$pam.cat)
```

```
##
##          Level 1 Level 2 Level 3 Level 4
##   DM-II       96     107     163      56
##   COPD        66      86      96      42
##   HF          67      67      61      28
##   CRD         79      63      61      16
```

- Chi Square Test:

```
#Test for association using a chi-sq-test
chisq.test(self.manage$pam.cat, self.manage$disease)
```

```
##
```

```
##  Pearson's Chi-squared test
##
## data:  self.manage$pam.cat and self.manage$disease
## X-squared = 27.869, df = 9, p-value = 0.001003
```

We see that the p-value $= 0.001$, which is less than $\alpha = 0.05$, and so there is sufficient evidence to reject the null hypothesis, indicating that there is an association between PAM-13 level and disease type.

v. The analysis from part ii is more informative as it gives us extent of the association between PAM-13 score and disease type. The approach from part iv only tells us that there is evidence to indicate an association exists. Moreover, in the part ii analysis we use actual scores and not categories of scores which enables us to be more prcise with our conclusions relative to the analysis from part iv.

PART B)

```
summary(lm(pam.score~supp.total, data=self.manage))
```

```
##
## Call:
## lm(formula = pam.score ~ supp.total, data = self.manage)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -24.930  -8.049  -1.084   6.261  37.738
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 46.39041    1.29478  35.829  < 2e-16 ***
## supp.total   0.12248    0.01999   6.128 1.22e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.63 on 1135 degrees of freedom
##   (17 observations deleted due to missingness)
## Multiple R-squared:  0.03203,    Adjusted R-squared:  0.03118
## F-statistic: 37.56 on 1 and 1135 DF,  p-value: 1.223e-09
```

```
confint(lm(pam.score~supp.total, data=self.manage))
```

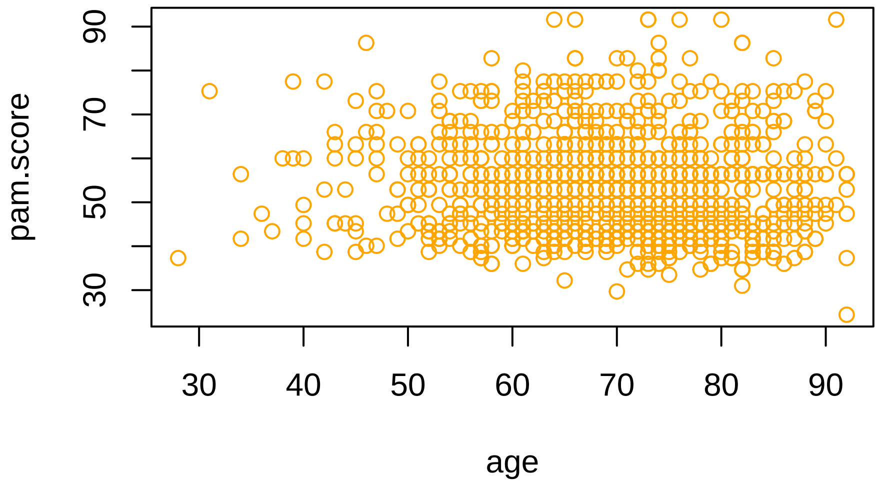
```
##                   2.5 %     97.5 %
## (Intercept) 43.84998104 48.9308485
## supp.total   0.08326591  0.1616923
```

The analysis shows that there is a significant association between PAM-13 score and perceived level of social support. On average the mean PAM-13 Score is higher by 0.122 for every additional point on the Multidimensional Scale of Perceived Support (MSPSS).
We are 95% confident that the interval (0.08326591, 0.1616923) contains the increment in mean PAM-13 score for each additional point on the MSPSS.

PART C)

i.

```r
#Plot Pam score and age
plot(pam.score~age, data=self.manage, col="orange")
```




```r
#Fit Linear Model
summary(lm(pam.score~age, data=self.manage))
```

```
##
## Call:
## lm(formula = pam.score ~ age, data = self.manage)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -28.201  -8.282  -1.056   5.772  38.931
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 58.83676    2.05215  28.671   <2e-16 ***
## age         -0.06778    0.02911  -2.328   0.0201 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.78 on 1149 degrees of freedom
##   (3 observations deleted due to missingness)
## Multiple R-squared:  0.004696,  Adjusted R-squared:  0.00383
## F-statistic: 5.422 on 1 and 1149 DF,  p-value: 0.02006
```

By looking at the plot of Pam score and age, no apparent trend is discernible, but running a linear model shows a significant association between PAM-13 score and age, with the mean PAM-13 score decreasing by 0.067 for every increase in 1 year of age.

ii.

```r
#Create new feature with PAM-13 score either < or > 55.2
self.manage$pam.binary = ifelse(self.manage$pam.score < 55.2,1,0)
#Convert it to a factor type
self.manage$pam.binary = factor(self.manage$pam.binary,
                             levels = c(1,0), labels = c("TRUE", "FALSE"))
```

```r
summary(glm(pam.binary ~ age + edu, data = self.manage,
family = binomial(link = "logit")))
```

```
##
## Call:
## glm(formula = pam.binary ~ age + edu, family = binomial(link = "logit"),
##     data = self.manage)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.4103  -1.0823  -0.9946   1.2611   1.4304
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.517778   0.402366   1.287   0.1982
## age         -0.010075   0.005541  -1.818   0.0690 .
## edumiddle   -0.168427   0.132278  -1.273   0.2029
## eduhigh      0.458167   0.179463   2.553   0.0107 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1552.1  on 1125  degrees of freedom
## Residual deviance: 1536.9  on 1122  degrees of freedom
##   (28 observations deleted due to missingness)
## AIC: 1544.9
##
## Number of Fisher Scoring iterations: 4
```

We see from the above Logistic Regression model that the estimated log odds of a person having a PAM-13 score lower than 55.2 decreases by 0.01, for every additional year of age, while adjusting for education level.
That said, the p-value associated with the slope for age is 0.069, which is greater than an $\alpha = 0.05$, indicating that the association between PAM-13 score being lower than 55.2, and age is not significant.

iii.

```r
#Adding an interactyion term age*edu in the model
summary(glm(pam.binary ~ age + edu + age*edu, data = self.manage,
family = binomial(link = "logit")))
```

```
##
```

```
## Call:
## glm(formula = pam.binary ~ age + edu + age * edu, family = binomial(link = "logit"),
##     data = self.manage)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5453  -1.0982  -0.9166   1.2371   1.6315
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -1.024474   0.607739  -1.686 0.091851 .
## age            0.011649   0.008456   1.377 0.168360
## edumiddle      2.648939   0.858452   3.086 0.002031 **
## eduhigh        2.839807   1.187913   2.391 0.016822 *
## age:edumiddle -0.040438   0.012201  -3.314 0.000919 ***
## age:eduhigh   -0.033975   0.016955  -2.004 0.045087 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1552.1  on 1125  degrees of freedom
## Residual deviance: 1524.9  on 1120  degrees of freedom
##   (28 observations deleted due to missingness)
## AIC: 1536.9
##
## Number of Fisher Scoring iterations: 4
```

The interaction term(age*edu) indicates the difference in the slope coefficient of Age between the education level of the participants. For middle education level, an increase in 1 year of age is associated with a lower predicted log odds of a PAM-13 score lower than 55.2, by 0.040438 (0.011649 - 0.040438 = -0.028789).The difference is large enough that although PAM-13 score log odds and age is positively as-associated, in middle level of education, they are negatively associated.
Similarly, for high education level, an increase in 1 year of age is associated with a lower predicted log odds of a PAM-13 score lower than 55.2, by 0.033975 (0.011649 - 0.033975 = -0.022326).The difference is large enough that although PAM-13 score log odds and age is positively as-associated, in high level of education, they are negatively associated.
Note: The AIC for this model with the interaction term is slightly lower, indicating a slightly more parsimonious fit.

iv. The statistical analysis conducted above does not conclusively indicate that older individuals generally tend to have a lower level of activation for self-management. **However**, when adjusting for other aspects of a persons life, such as their education level, we do see that the odds of a person having a lower level of activation for self-management increase slightly.

PART D)

i. The relative risk (RR)=$\dfrac{PAMLevel1InGroupWithAnxiety > 11}{PAMLevel1InGroupWithAnxiety <= 11} =$

```r
anxious=subset(self.manage, hads.anxiety>11)
notanxious=subset(self.manage, hads.anxiety<=11)

riskanxious=nrow(subset(anxious, pam.cat=="Level 1"))/nrow(anxious)

risknotanxious=nrow(subset(notanxious, pam.cat=="Level 1"))/nrow(notanxious)

RR=(riskanxious)/(risknotanxious)
RR
```

```
## [1] 1.565154
```

So the Relative risk of being classified as PAM Level 1 for individuals with an anxiety disorder versus individuals without an anxiety disorder is 1.565. In other words, individuals with anxiety are 1.565 times more likely to have a PAM-13 level of 1.

ii. Yes, Relative Risk has an interpretable meaning in this case:
"Individuals with anxiety are 1.565 times more likely to have a PAM-13 level of 1 compared to individuals without anxiety."

The relative risk cannot be used in studies that use outcome-dependent sampling. In our case here, the researchers did not sample a certain number of participants with and without anxiety and so it is reasonable to assume that the sample proportion of individuals with PAM-13 Level being 1 represents the estimated population proportion.

PART E)

i.

```r
thismodel=lm(pam.score~age+bmi+edu+financial+
               disease+supp.total+hads.anxiety, data=self.manage)
summary(thismodel)
```

```
##
## Call:
## lm(formula = pam.score ~ age + bmi + edu + financial + disease +
##     supp.total + hads.anxiety, data = self.manage)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -21.932  -7.521  -1.330   5.798  38.175
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  64.0639450  3.6180799  17.707  < 2e-16 ***
## age          -0.0743743  0.0306907  -2.423 0.015551 *
## bmi          -0.2854002  0.0719840  -3.965 7.86e-05 ***
## edumiddle     0.0007191  0.7008846   0.001 0.999182
## eduhigh       2.4906715  0.9626195   2.587 0.009808 **
## financiallow -2.1592488  0.6824268  -3.164 0.001602 **
```

```
## financialhigh -1.6077919  1.2729926  -1.263 0.206877
## diseaseCOPD   -0.2115228  0.8337272  -0.254 0.799773
## diseaseHF     -1.3062815  0.9308654  -1.403 0.160832
## diseaseCRD    -3.4468910  0.9006020  -3.827 0.000137 ***
## supp.total     0.1043092  0.0208150   5.011 6.37e-07 ***
## hads.anxiety  -0.3739119  0.0866497  -4.315 1.75e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.1 on 1022 degrees of freedom
##   (120 observations deleted due to missingness)
## Multiple R-squared:  0.1199, Adjusted R-squared:  0.1104
## F-statistic: 12.65 on 11 and 1022 DF,  p-value: < 2.2e-16
```

The model summary above shows the slope of the coeffcients for the variables included in the model. The provide useful information in terms of the predicted change in the mean value of PAM-13 score. For example, while controlling for all other variables, a high education level has a large positive effect on the PAM-13 score, with the expected change in score being 2.49 higher for individuals who have a 'high' level of education. Similarly, while controlling for all other variables, the presence of CRD, is associated with a predicted decrease of 3.446 on average to the mean PAM-13 score. These results are useful in understanding how each variable affects the predicted mean PAM-13 scores for individuals while controlling for other factors. One further step to help potentially inprove the model would be to compare models using different combinations of predictor variables and evaluating the adjusted R^2 values of the models. This might help us get to a more parsimonious model.

ii.

```
#Fit and Confidence interval for individual
predict(thismodel,data =self.manage,newdata =data.frame(age =63,
  disease ="DM-II", financial="none",edu="high", bmi=30, supp.total=75,
              hads.anxiety=4),level =0.95,interval ="confidence")
```

```
##        fit     lwr      upr
## 1 59.63458 57.7004 61.56875
```

```
#Fit and Prediction interval for individual(***Only for exploration***)
predict(thismodel,data =self.manage,newdata =data.frame(age =63,
  disease ="DM-II", financial="none",edu="high", bmi=30, supp.total=75,
  hads.anxiety=4),level =0.95,interval ="prediction")
```

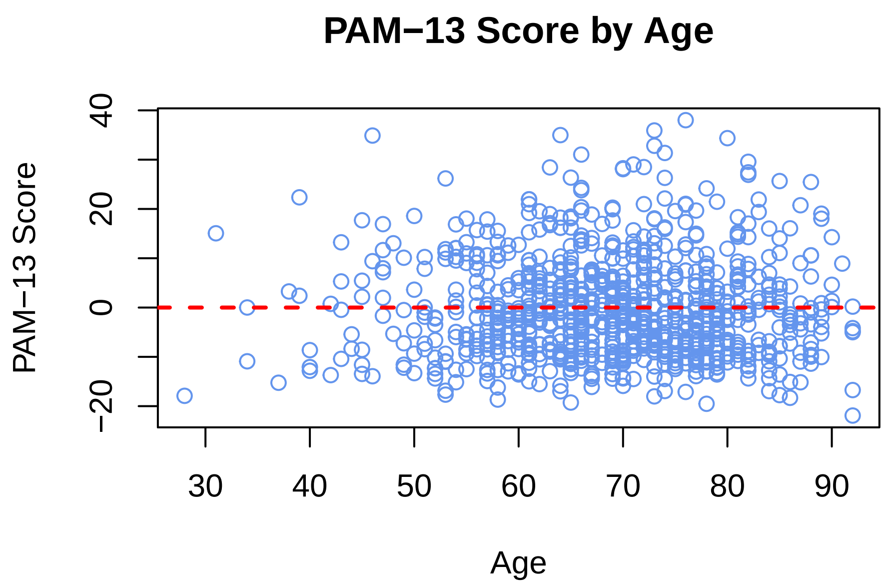
```
##        fit     lwr      upr
## 1 59.63458 39.7171 79.55206
```

We can see that the predicted mean value of PAM-13 score for the individual in question is 59.634. Additionally, we are 95% confident that the interval (57.7004, 61.56875), includes the mean Pam-13 score for an individual with those characteristics.

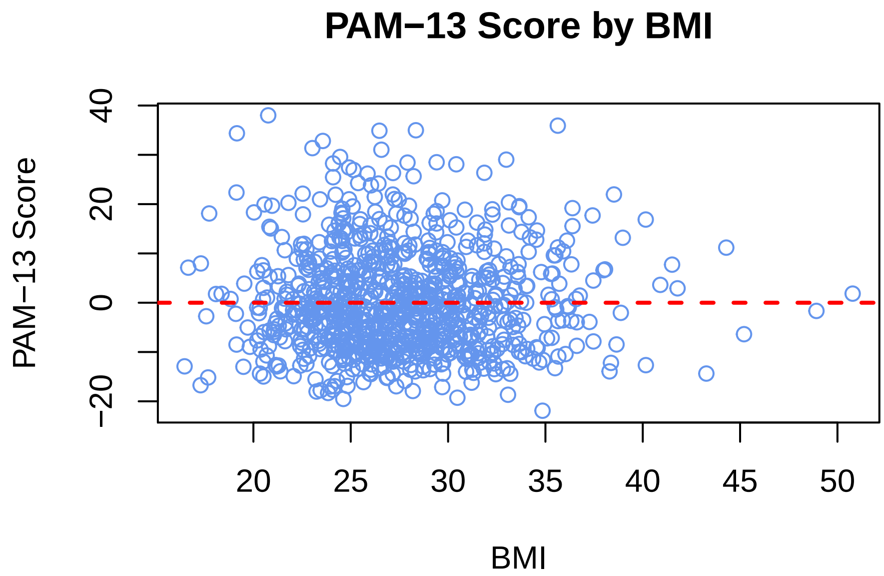
iii.

```
#Check Assumptions
self.manage=na.omit(self.manage)
```

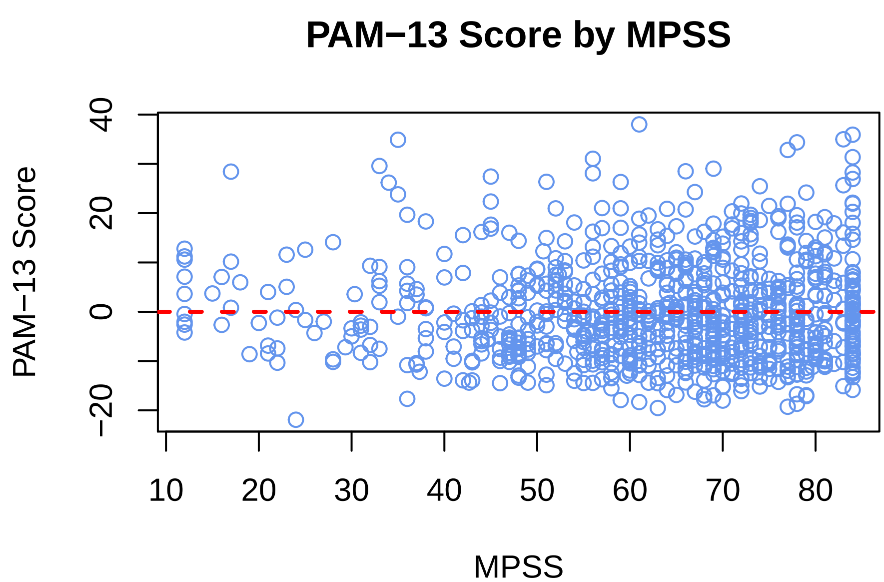
```r
thatmodel=lm(pam.score~age+bmi+edu+financial+
               disease+supp.total+hads.anxiety, data=self.manage)

###############
###Linearity###
###############

#AGE
dataPlot=plot(residuals(thatmodel)~self.manage$age,
              main = "PAM-13 Score by Age",
              xlab = "Age", ylab = "PAM-13 Score",
              col="cornflowerblue")
abline(lm(residuals(thatmodel)~self.manage$age),
       col = "red", lty = 2, lwd = 2)
```



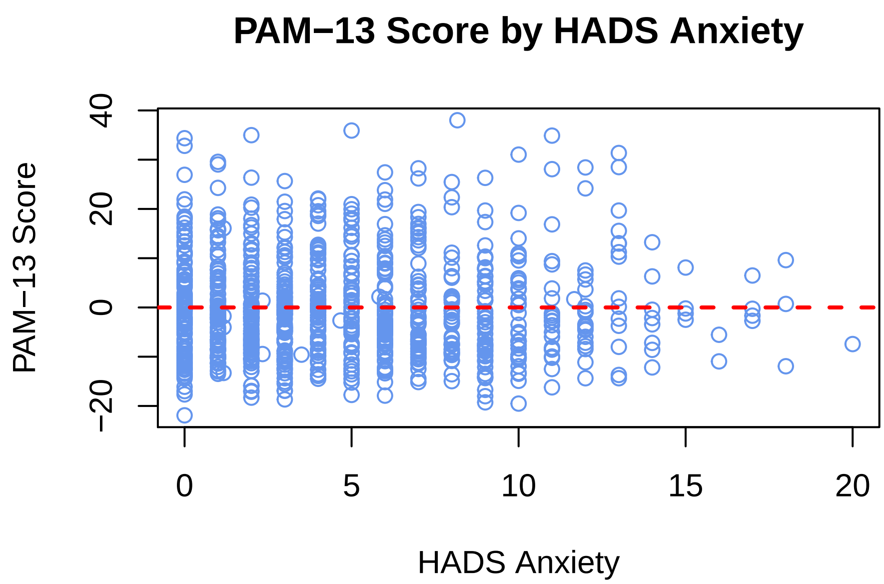

```r
#BMI
dataPlot=plot(residuals(thatmodel)~self.manage$bmi,
              main = "PAM-13 Score by BMI",
              xlab = "BMI", ylab = "PAM-13 Score",
              col="cornflowerblue")
abline(lm(residuals(thatmodel)~self.manage$bmi),
       col = "red", lty = 2, lwd = 2)
```



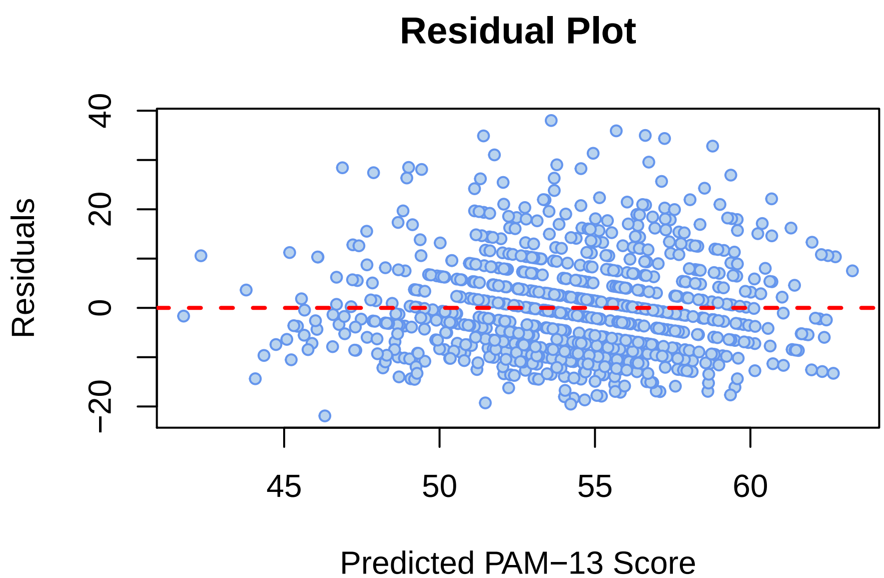

```
#MPSS
dataPlot=plot(residuals(thatmodel)~self.manage$supp.total,
              main = "PAM-13 Score by MPSS",
              xlab = "MPSS", ylab = "PAM-13 Score",
              col="cornflowerblue")
abline(lm(residuals(thatmodel)~self.manage$supp.total),
       col = "red", lty = 2, lwd = 2)
```




```
#HADS Anxiety
dataPlot=plot(residuals(thatmodel)~self.manage$hads.anxiety,
              main = "PAM-13 Score by HADS Anxiety",
              xlab = "HADS Anxiety", ylab = "PAM-13 Score",
              col="cornflowerblue")
abline(lm(residuals(thatmodel)~self.manage$hads.anxiety),
       col = "red", lty = 2, lwd = 2)
```

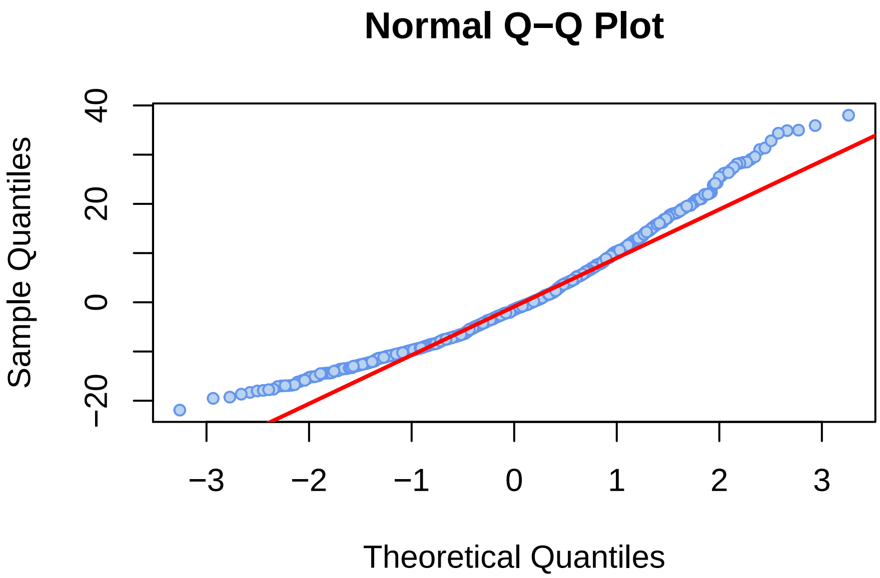



```
######################
###Constant Variance###
######################

plot(residuals(thatmodel)~predict(thatmodel),main = "Residual Plot",
     xlab = "Predicted PAM-13 Score", ylab = "Residuals",
     pch = 21, col = "cornflowerblue", bg = "slategray2",cex = 0.75)
abline(h = 0, lty = 2, lwd = 2, col = "red")
```




```
###########################
###Normality of Residuals###
###########################

qqnorm(residuals(thatmodel),pch = 21, col = "cornflowerblue",
       bg = "slategray2", cex = 0.75)
```

```
qqline(residuals(thatmodel),col = "red", lwd = 2)
```



Assumptions:
* LINEARITY: We can see that when we check for linearity of numerical predictors- age, bmi, MPSS Score and HADS Anxiety score, there is reasonable linearity to proceed.
* CONSTANT VARIANCE: A residual plot depicts that there is constant variance.
* NORMALITY OF RESIDUALS: There is approximate normality of residuals, although there is slight deviations from normality in the upward direction in both the lower as well as the upper tails.
* INDEPENDENCE: It is reasonable to assume independence of observations since we are given no information about dependence of data for one individual on another in the dataset.

iv. When we look at the p-values associated with the slope coefficients in the model, we notice that the p-values for 'edu middle', 'disease COPD“,”disease HF' and 'financial High' are greater than our α value of 0.05. This would suggest that these associations are not as statistically significant.