ProblemSet8

Johnny Magdaleno

2025-03-08

PROBLEM 1. Self-management. *Question A. Write an informative summary describing features of the study participants with respect to the variables age, sex, and type of chronic disease. Reference appropriate numerical and graphical summaries as needed. No more than five sentences.*

R code:

```
load("/Users/jm/Desktop/MPH/Spring 25 - Biostatistics in Public Health/Datasets/self_manage.Rdata")

table(self.manage$sex)
```

```
##
##   male female
##    694    458
```

```
summary(self.manage$age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   28.00   63.00   70.00   69.64   78.00   92.00       3
```

```
self.manage.male = subset(self.manage, self.manage$sex == "male")
self.manage.female = subset(self.manage, self.manage$sex == "female")
summary(self.manage.male$age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   28.00   63.00   70.00   69.64   77.00   92.00       1
```

```
summary(self.manage.female$age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   34.00   62.00   70.00   69.64   78.00   92.00
```

```
table(self.manage$disease)
```

```
##
## DM-II  COPD    HF   CRD
##   422   290   223   219
```

```
plot(self.manage$age ~ self.manage$sex)
```


```
prop.table(table(self.manage$disease, self.manage$sex), 2)
```

```
##
##              male    female
##   DM-II 0.3472622 0.3951965
##   COPD  0.2651297 0.2292576
##   HF    0.1930836 0.1921397
##   CRD   0.1945245 0.1834061
```

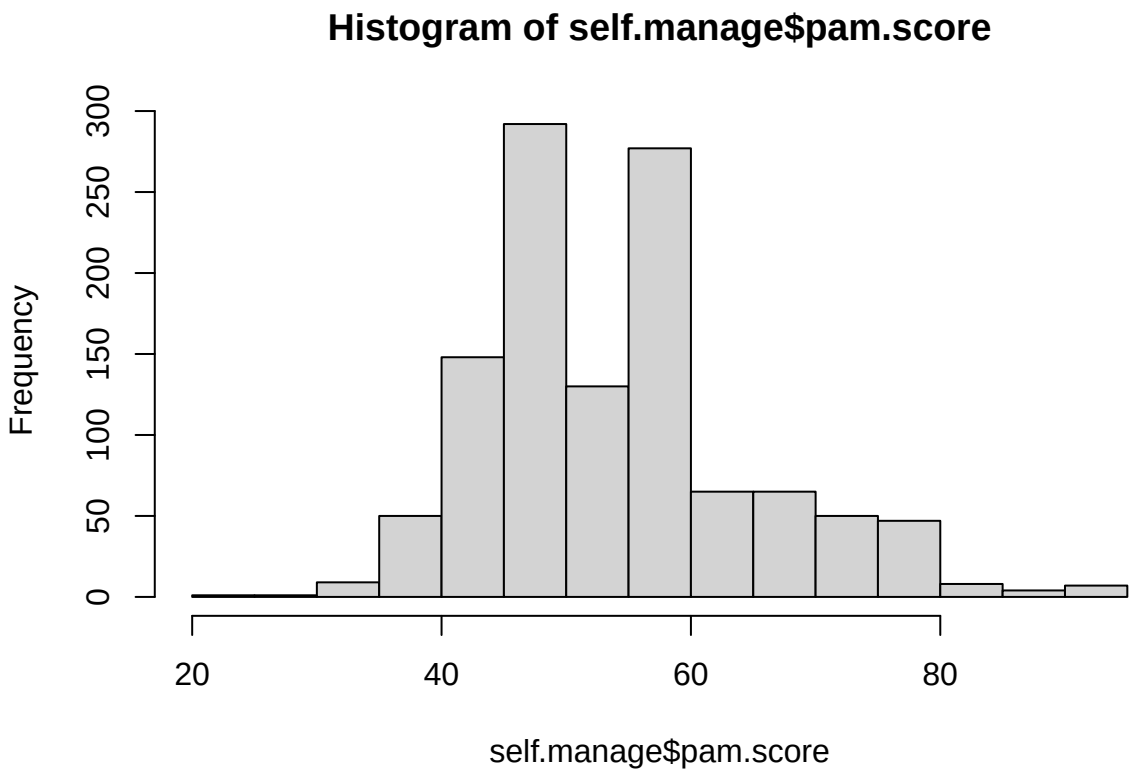
```
hist(self.manage$age)
```

Answer: The sample has a mean and median age of about 70 years old for males as well as females. Sixty percent (60%) of the sample is male. Type 2 diabetes is the most common disease, followed by COPD, chronic heart failure and chronic renal disease. Females in the sample experience a higher rate of type 2 diabetes than males, while males experience a higher rate of COPD and chronic renal disease than females.

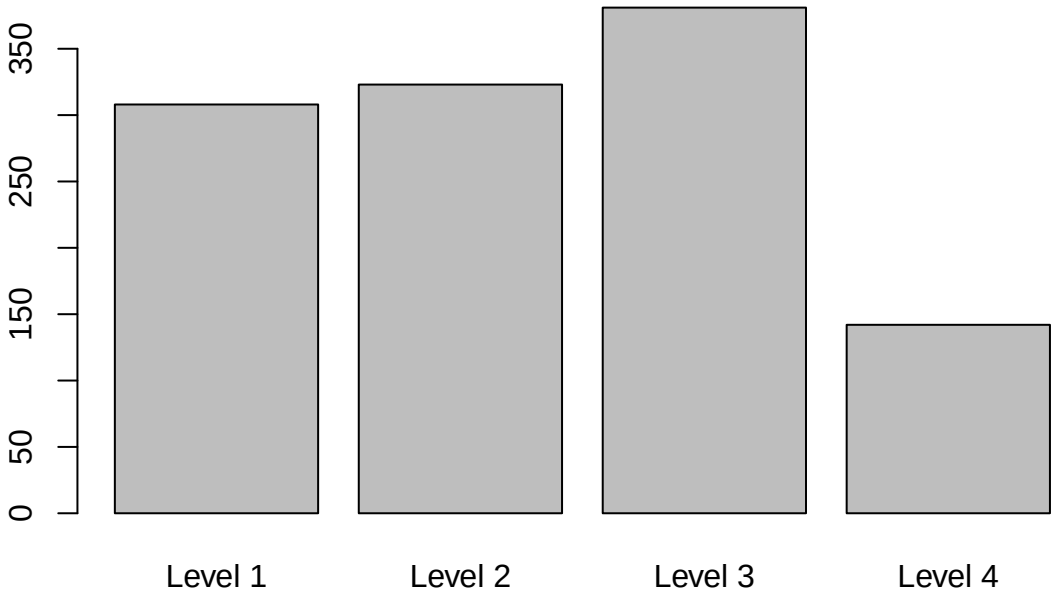
Question B.The main measurement of interest is activation for self-management. Describe the distribution of activation for self-management within the study participants, both in terms of PAM-13 score and PAM level. Reference appropriate numerical and graphical summaries as needed. Five sentence maximum.

R code:

```
hist(self.manage$pam.score)
```




```
plot(self.manage$pam.cat)
```




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

```
##
## Level 1 Level 2 Level 3 Level 4
##     308     323     381     142
```

```
prop.table(table(self.manage$pam.cat, self.manage$sex), 2)
```

```
##
##                male    female
```

```
##   Level 1 0.2449568 0.2969432
##   Level 2 0.2896254 0.2663755
##   Level 3 0.3371758 0.3209607
##   Level 4 0.1282421 0.1157205
```

```r
prop.table(table(self.manage$pam.cat, self.manage$disease), 2)
```

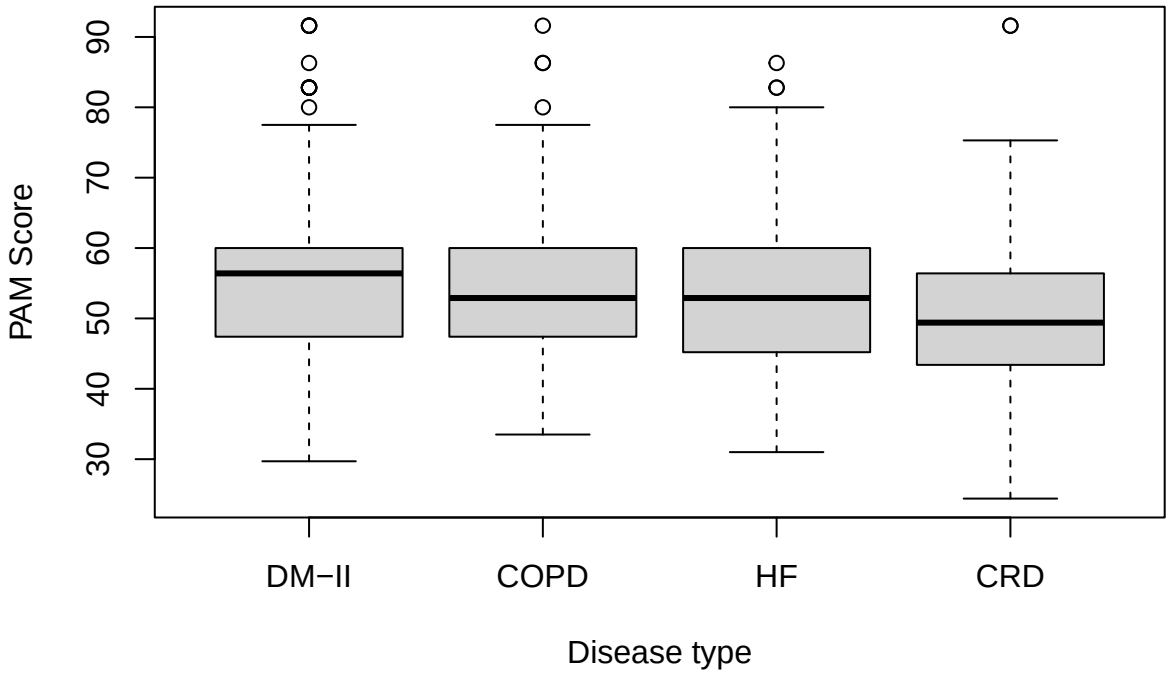
```
##
##                DM-II       COPD         HF        CRD
##   Level 1 0.22748815 0.22758621 0.30044843 0.36073059
##   Level 2 0.25355450 0.29655172 0.30044843 0.28767123
##   Level 3 0.38625592 0.33103448 0.27354260 0.27853881
##   Level 4 0.13270142 0.14482759 0.12556054 0.07305936
```

Answer: The distribution of PAM scores is bimodal, therefore it is not normally distributed. The predominant PAM-13 score level of the sample is 3, followed by level 2, level 1 and then level 4. High level 3 frequency indicates most of the sample participants are "goal-oriented and actively working to achieve best practice behaviors." Men experience a higher rate of level 2, level 3 and level 4 PAM scores than women, while women experience a higher rate of level 1 scores level than men. Patients with chronic renal disease experience the highest rate of level 1 scores; heart failure, the highest rate of level 2 scores; diabetes, the highest rate of level 3 scores; and COPD, the highest rate of level 4 scores.

Question C. Explore the relationship between activation for self-management and disease type.
C. i. Create a plot to graphically show the association between disease type and PAM-13 score. Describe what you see.

R code:

```r
plot(self.manage$disease, self.manage$pam.score, xlab = "Disease type", ylab = "PAM Score")
```



Answer: Generally speaking, patients with type 2 diabetes appear to produce the highest average PAM scores as indicated by the high median relative to the other three data types. The type 2 diabetes patient group also has the largest amount of upper outliers. COPD and HF had similar median scores, and CRD had the lowest median scores and overall score range.

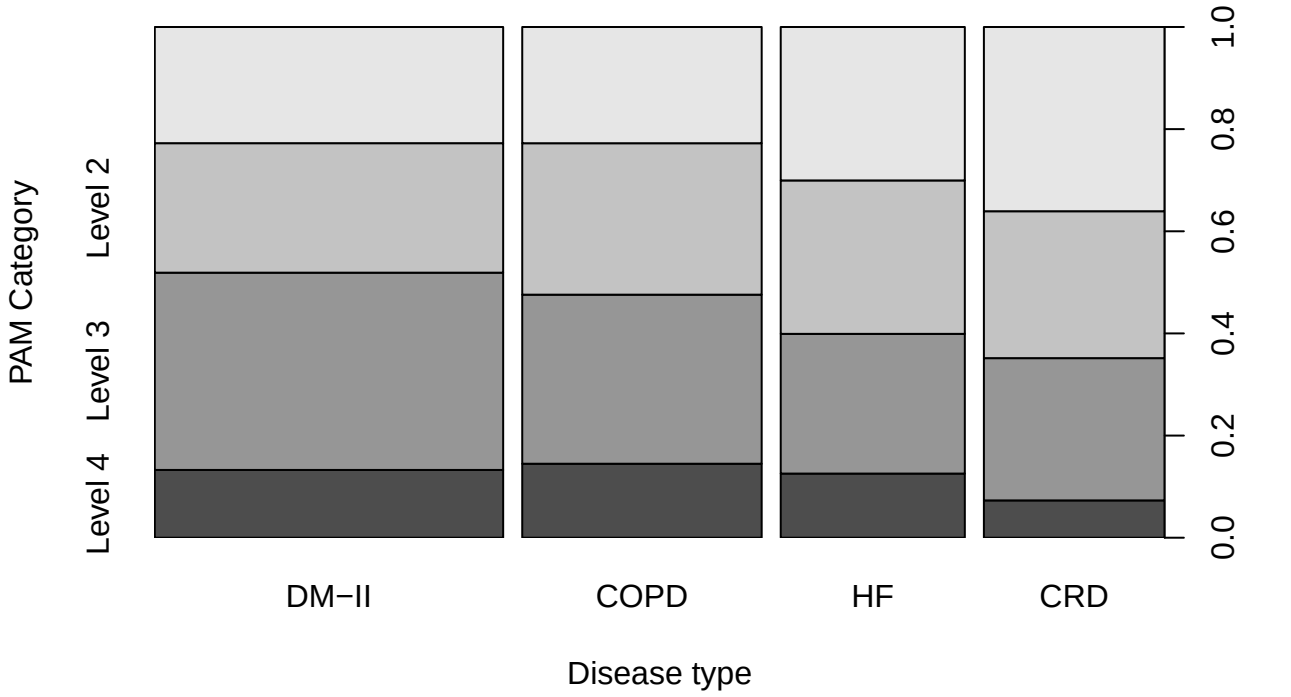
C. ii. Create a summary that shows how the distribution of PAM level differs between disease type. Describe what you see.

R code:

```
prop.table(table(self.manage$pam.cat, self.manage$disease), 2)
```

```
##
##                DM-II       COPD         HF        CRD
##   Level 1 0.22748815 0.22758621 0.30044843 0.36073059
##   Level 2 0.25355450 0.29655172 0.30044843 0.28767123
##   Level 3 0.38625592 0.33103448 0.27354260 0.27853881
##   Level 4 0.13270142 0.14482759 0.12556054 0.07305936
```

```
plot(self.manage$disease, self.manage$pam.cat, xlab = "Disease type", ylab = "PAM Category")
```



Answer: The summary graphic and table both show that CRD and HF patients experience the greatest proportions of level 1 PAM scores. HF and COPD patients experience the greatest proportions of level 2 PAM scores. DM-II and COPD patients experience the greatest proportions of both level 3 and level 4 scores.

C. iii. Do you find the summary from part i. or the summary from part ii. more informative with regards to understanding the relationship between activation for self-management and disease type? Explain your answer.

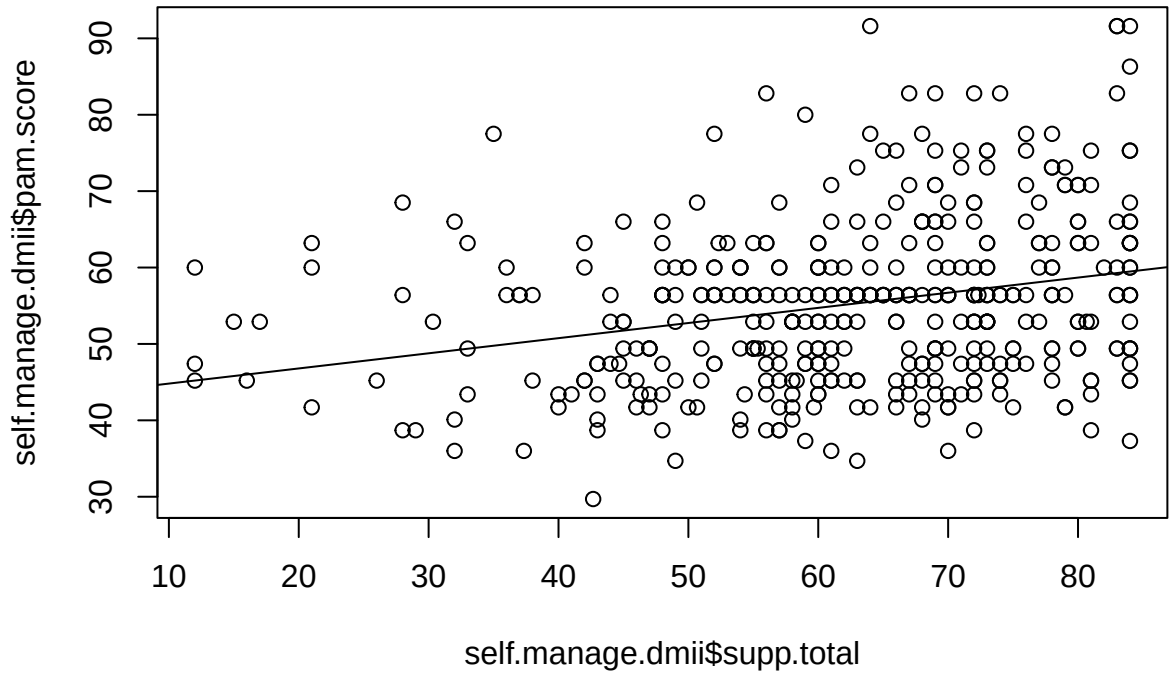
Answer: I find the summary from part ii. more informative. From a public health standpoint, I am interested in understanding this relationship with the goal of identifying patients with disease types that are disproportionately represented in the level 1 and level 2 score groups, indicating low self-management. The graphic in part ii. clearly shows the largest proportions of patients in the level 1 score category are those diagnosed with HF or CRD. This will help me understand which patient groups need support for improving self-management.

Question D. Is PAM-13 score associated with perceived level of social support?

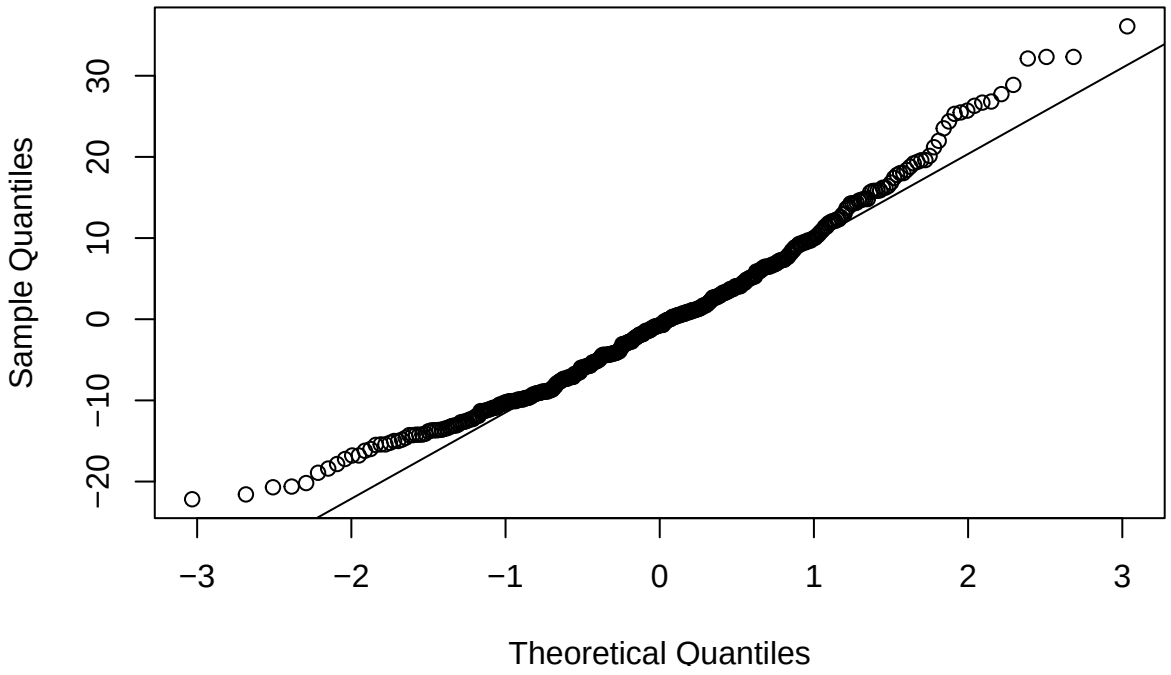
D. i. Using graphical and numerical methods, investigate this question separately within disease types. Summarize your findings, with particular focus on whether the association between PAM-13 score and perceived level of social support seems to differ between types of chronic disease.

R code:

```
self.manage.dmii = subset(self.manage, self.manage$disease == "DM-II")
self.manage.copd = subset(self.manage, self.manage$disease == "COPD")
self.manage.hf = subset(self.manage, self.manage$disease == "HF")
self.manage.crd = subset(self.manage, self.manage$disease == "CRD")
options(scipen=0)

#Plot, analysis for DM-II.
plot(self.manage.dmii$pam.score ~ self.manage.dmii$supp.total)
abline(lm(self.manage.dmii$pam.score ~ self.manage.dmii$supp.total))
```




```
dmii.pam.supp.model = lm(self.manage.dmii$pam.score ~ self.manage.dmii$supp.total)
qqnorm(resid(dmii.pam.supp.model), main = "DM-II PAM Score by Social Support Q-Q Plot")
qqline(resid(dmii.pam.supp.model))
```




```
coef(summary(dmii.pam.supp.model))
```

```
##                               Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)                 42.8258632 2.23727836 19.141947 8.104204e-59
## self.manage.dmii$supp.total  0.1982561 0.03475468  5.704443 2.238397e-08
```

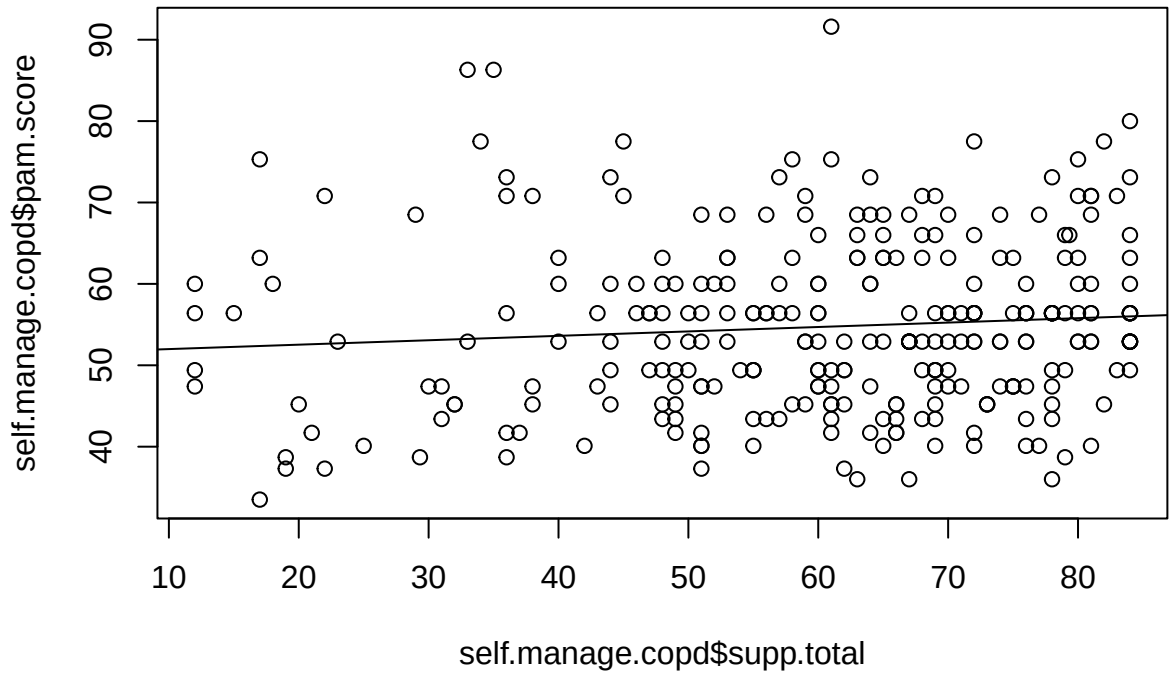
```
format(2.238397e-08, scientific = FALSE)
```

```
## [1] "0.00000002238397"
```

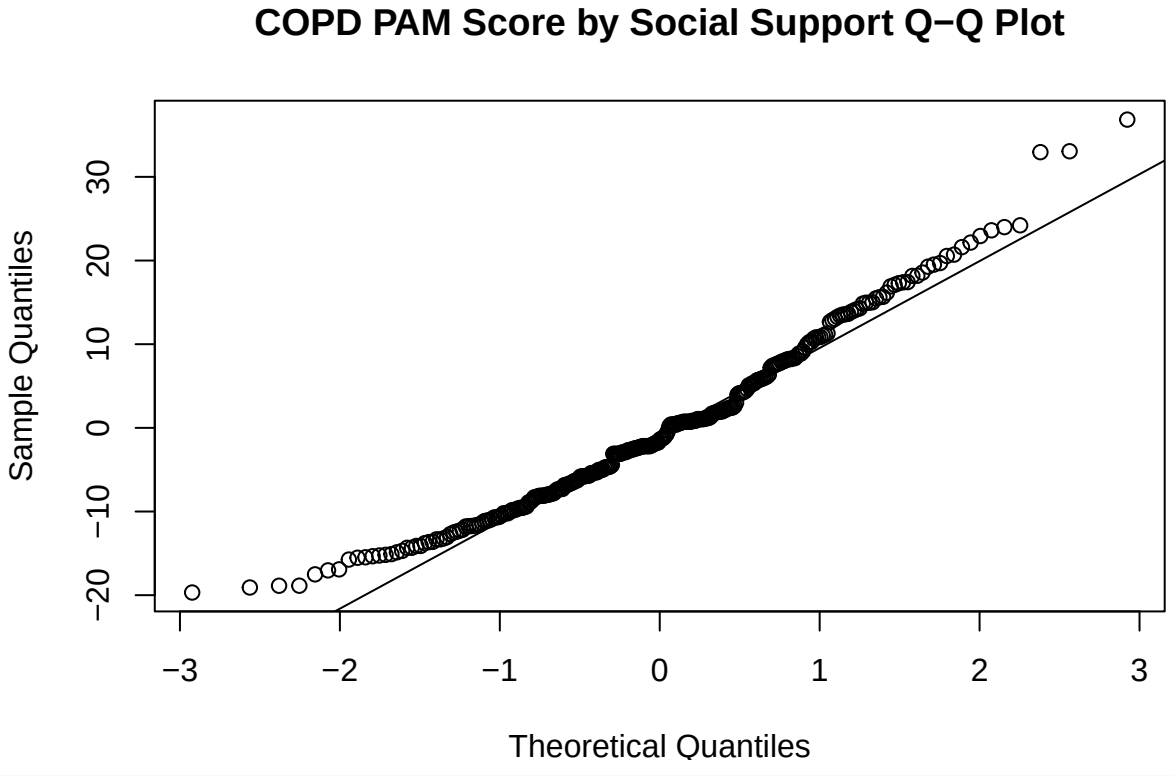
```
summary(dmii.pam.supp.model)$r.squared
```

```
## [1] 0.07353147
```

```
#Plot, analysis for COPD.
plot(self.manage.copd$pam.score ~ self.manage.copd$supp.total)
abline(lm(self.manage.copd$pam.score ~ self.manage.copd$supp.total))
```




```
copd.pam.supp.model = lm(self.manage.copd$pam.score ~ self.manage.copd$supp.total)
qqnorm(resid(copd.pam.supp.model), main = "COPD PAM Score by Social Support Q-Q Plot")
qqline(resid(copd.pam.supp.model))
```




```
coef(summary(copd.pam.supp.model))
```

```
##                                Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)                51.44676443 2.22179681 23.155477 1.252433e-67
## self.manage.copd$supp.total  0.05425702 0.03523856  1.539706 1.247338e-01
```

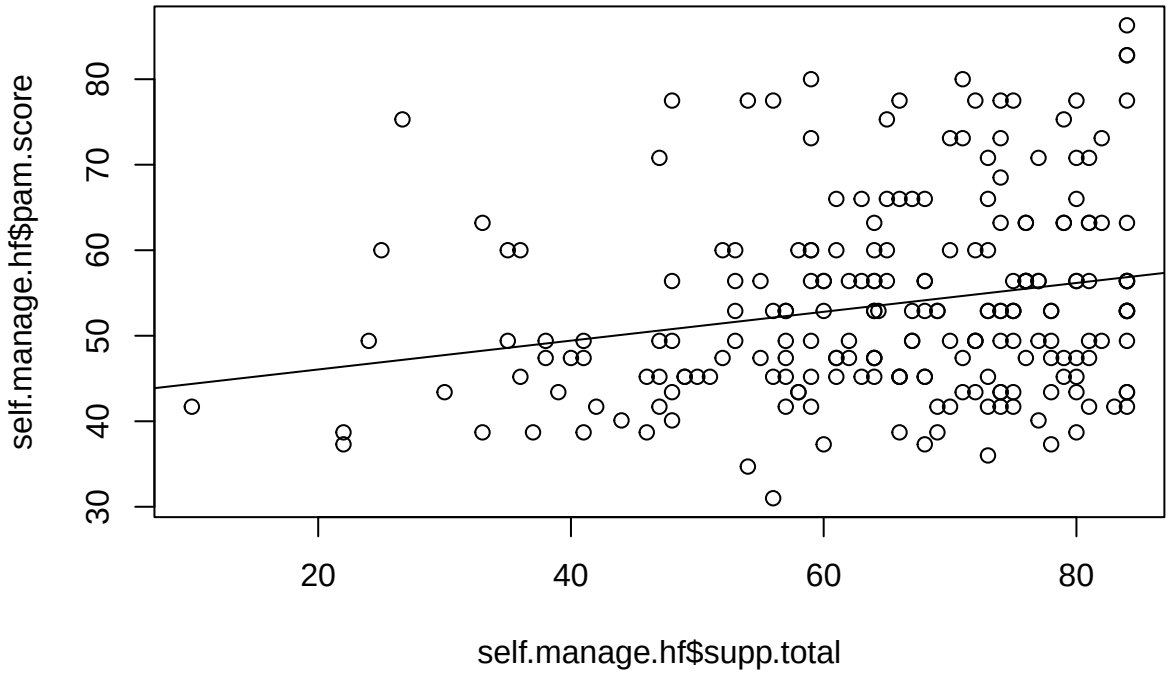
```r
format(1.247338e-01, scientific = FALSE)
```

```
## [1] "0.1247338"
```

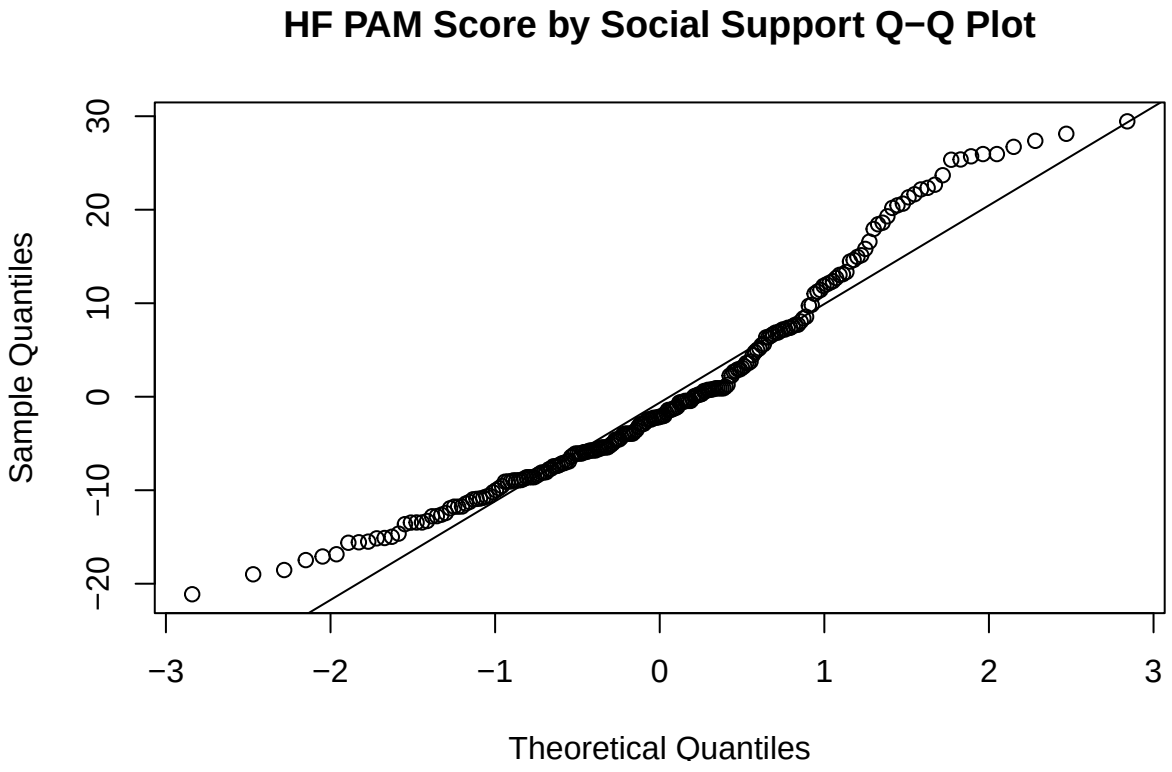
```r
summary(copd.pam.supp.model)$r.squared
```

```
## [1] 0.008192586
```

```r
#Plot, analysis for HF.
plot(self.manage.hf$pam.score ~ self.manage.hf$supp.total)
abline(lm(self.manage.hf$pam.score ~ self.manage.hf$supp.total))
```




```r
hf.pam.supp.model = lm(self.manage.hf$pam.score ~ self.manage.hf$supp.total)
qqnorm(resid(hf.pam.supp.model), main = "HF PAM Score by Social Support Q-Q Plot")
qqline(resid(hf.pam.supp.model))
```




```
coef(summary(hf.pam.supp.model))
```

```
##                             Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)                42.680050 3.28280176 13.001105 4.756050e-29
## self.manage.hf$supp.total   0.168668 0.04919433  3.428607 7.241739e-04
```

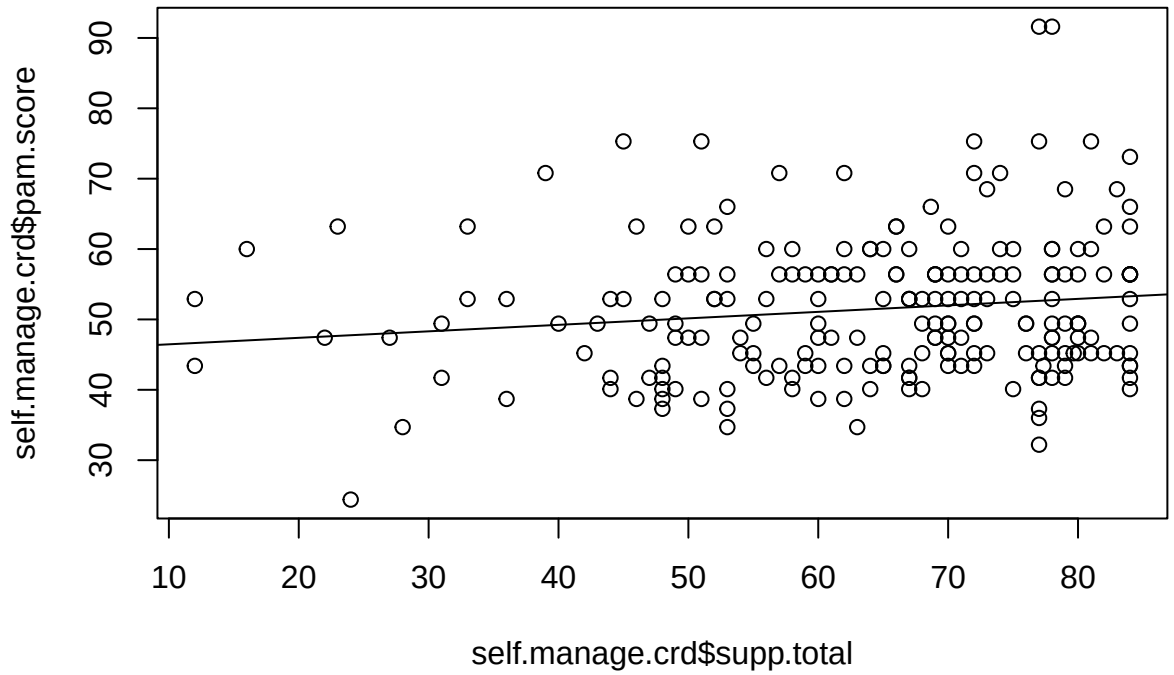
```
format(7.241739e-04, scientific = FALSE)
```

```
## [1] "0.0007241739"
```

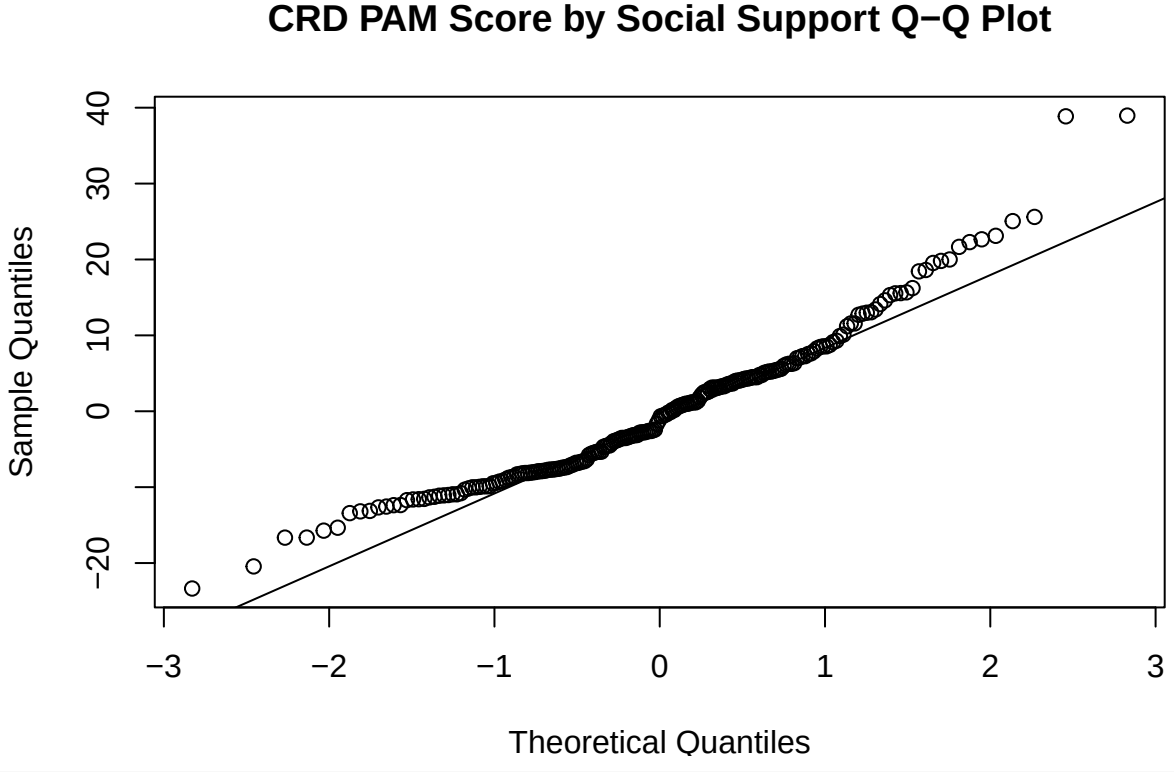
```
summary(hf.pam.supp.model)$r.squared
```

```
## [1] 0.05072309
```

```
#Plot, analsysis for CRD.
plot(self.manage.crd$pam.score ~ self.manage.crd$supp.total)
abline(lm(self.manage.crd$pam.score ~ self.manage.crd$supp.total))
```




```
crd.pam.supp.model = lm(self.manage.crd$pam.score ~ self.manage.crd$supp.total)
qqnorm(resid(crd.pam.supp.model), main = "CRD PAM Score by Social Support Q-Q Plot")
qqline(resid(crd.pam.supp.model))
```




```
coef(summary(crd.pam.supp.model))
```

```
##                              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)                45.54098435 2.93201267 15.532329 7.814352e-37
## self.manage.crd$supp.total  0.09223569 0.04453963  2.070868 3.958205e-02
```

```r
format(3.958205e-02, scientific = FALSE)
```

```
## [1] "0.03958205"
```

```r
summary(crd.pam.supp.model)$r.squared
```

```
## [1] 0.01982766
```

Answer: All distributions analyzed for this answer are approximately normal, according to Q-Q plots. Variation is also consistent around the plotted model lines.

For every disease type except COPD, PAM scores increase as perceived social support increases. At a 95% confidence level, P-values for all disease types indicate statistical significance except COPD, in which the relationship has a p-value of 0.12.

The strongest relationships between the two variables are observed among DM-II patients and HF patients. The R-squared scores for DM-II patients and HF patients are .07 and .05, respectively, and the slope coefficients (increase in PAM score for each 1-point increase in perceived social support score) are 0.2 and 0.17, respectively.

CRD and COPD patient data showed the weakest relationships between perceived social support and PAM scores. The R-squared scores for CRD patients and COPD patients are .02 and .01, respectively, and the slope coefficients are 0.09 and 0.05, respectively.

While the R-squared scores are weak among all patient types, the lowest R-squared score and overall strongest evidence of no association is found among COPD patients.

D. ii. Do these data suggests that an increase in perceived social support leads to better capacity for self-management? Explain your answer in no more than five sentences.

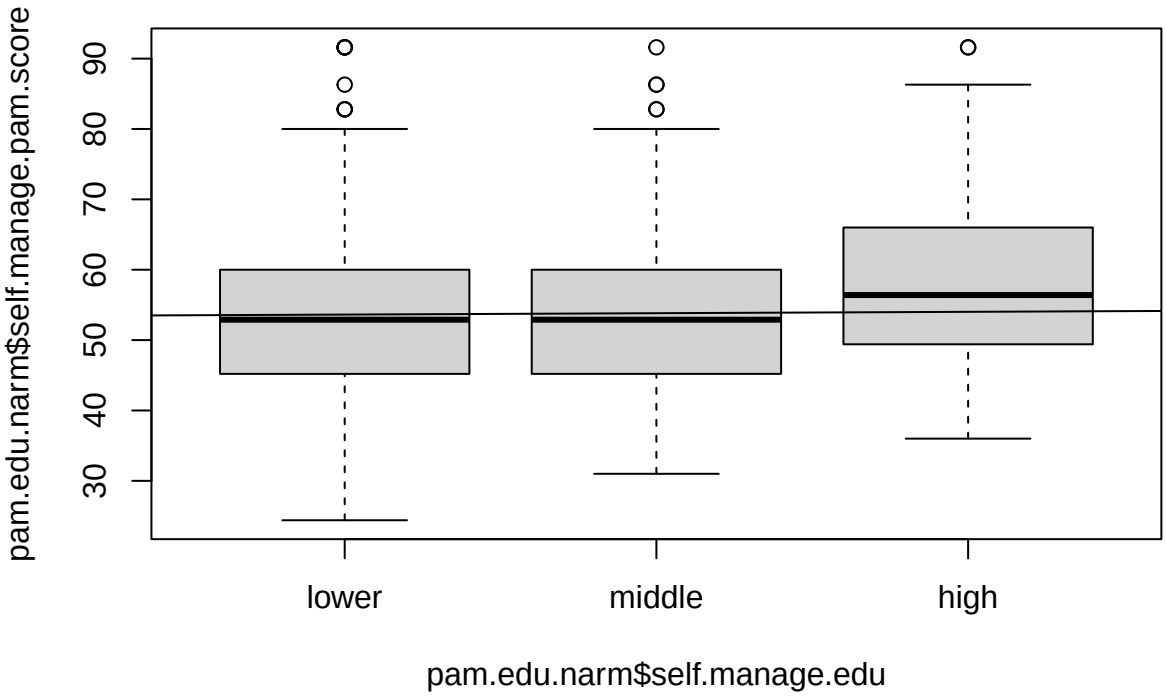
Answer: The relationship between both variables (R^2 scores) observed in each disease is relatively weak, suggesting there may be other variables worth analyzing if the goal is to identify significant contributors to increased PAM score. Additionally, there may be confounding by one or more socio-economic variables that contribute to one's ability to self-manage and one's positive outlook on their social support network. This could be education level, depression or anxiety level or financial security. In short, this analysis isn't enough to conclude that perceived social support on its own contributes to higher PAM scores.

Question E. Investigate the relationship between PAM-13 score and educational level.
E. i. With reference to appropriate numerical and graphical summaries, describe the association between PAM-13 score and educational level.

R code:

```r
load("/Users/jm/Desktop/MPH/Spring 25 - Biostatistics in Public Health/Datasets/self_manage.Rdata")

#Creating data frame of PAM score and education level, to account for NAs.
pam.edu = data.frame(self.manage$pam.score, self.manage$edu)
pam.edu.narm = na.omit(pam.edu)

#Analysis
plot(pam.edu.narm$self.manage.pam.score ~ pam.edu.narm$self.manage.edu)
abline(lm(pam.edu.narm$self.manage.pam.score ~ pam.edu.narm$self.manage.edu))
```

```
## Warning in abline(lm(pam.edu.narm$self.manage.pam.score ~
## pam.edu.narm$self.manage.edu)): only using the first two of 3 regression
## coefficients
```




```
tapply(pam.edu.narm$self.manage.pam.score, pam.edu.narm$self.manage.edu, mean)
```

```
##    lower   middle     high
## 53.42981 53.62611 57.93571
```

```
mean(pam.edu.narm$self.manage.pam.score)
```

```
## [1] 54.17622
```

```
lm(pam.edu.narm$self.manage.pam.score ~ pam.edu.narm$self.manage.edu)
```

```
##
## Call:
## lm(formula = pam.edu.narm$self.manage.pam.score ~ pam.edu.narm$self.manage.edu)
##
## Coefficients:
##                         (Intercept)  pam.edu.narm$self.manage.edumiddle
##                             53.4298                              0.1963
##   pam.edu.narm$self.manage.eduhigh
##                              4.5059
```

```
coef(summary(lm(pam.edu.narm$self.manage.pam.score ~ pam.edu.narm$self.manage.edu)))
```

```
##                                    Estimate Std. Error     t value
## (Intercept)                       53.4298113  0.4662827 114.5867401
## pam.edu.narm$self.manage.edumiddle  0.1962959  0.6971559   0.2815667
## pam.edu.narm$self.manage.eduhigh    4.5059030  0.9504348   4.7408857
##                                       Pr(>|t|)
## (Intercept)                       0.000000e+00
## pam.edu.narm$self.manage.edumiddle 7.783276e-01
## pam.edu.narm$self.manage.eduhigh   2.400644e-06
```

```
summary(lm(pam.edu.narm$self.manage.pam.score ~ pam.edu.narm$self.manage.edu))
```

```
##
```

```
## Call:
## lm(formula = pam.edu.narm$self.manage.pam.score ~ pam.edu.narm$self.manage.edu)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -29.030  -8.426  -0.726   6.374  38.170
##
## Coefficients:
##                                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)                        53.4298     0.4663 114.587  < 2e-16 ***
## pam.edu.narm$self.manage.edumiddle   0.1963     0.6972   0.282    0.778
## pam.edu.narm$self.manage.eduhigh     4.5059     0.9504   4.741  2.4e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.73 on 1124 degrees of freedom
## Multiple R-squared:  0.02116,    Adjusted R-squared:  0.01942
## F-statistic: 12.15 on 2 and 1124 DF,  p-value: 6.034e-06
```

```
confint(lm(pam.edu.narm$self.manage.pam.score ~ pam.edu.narm$self.manage.edu), level = 0.95)
```

```
##                                     2.5 %    97.5 %
## (Intercept)                      52.514929 54.344694
## pam.edu.narm$self.manage.edumiddle -1.171578  1.564169
## pam.edu.narm$self.manage.eduhigh    2.641077  6.370729
```

```
format(6.034e-06, scientific = FALSE)
```

```
## [1] "0.000006034"
```

```
summary(lm(pam.edu.narm$self.manage.pam.score ~ pam.edu.narm$self.manage.edu))$r.squared
```

```
## [1] 0.02115753
```

Answer: There appears to be a relationship between education level and PAM score among patients with higher education levels, but not for patients with lower or middle education levels. The p-value for the hypothesis test of this relationship among highly educated patients is less than .01, indicating high statistical significance. The F-statistic for the overall analysis is 12.15 with a p-value of less than .01, indicating that at least one of the education variables has a statistically significant relationship with PAM score. The model equation produces the highest slope for highly educated patients, with an increase of 4.74 PAM score points compared to lower education status. It is also worth noting that only patients with high education have a mean PAM score higher than the overall sample mean PAM score, while patients with lower to medium education have mean PAM scores less than the sample mean PAM score.

E. ii. Propose one possible explanation for the trends observed in part i. Limit your answer to no more than five sentences.

Answer: High level of education may result in greater awareness of the importance of self-care and the ability to comprehend and prepare for long-term disease outcomes. It also may be associated with an overall higher socio-economic status, which typically means more access to resources (including medical) and support networks and therefore better health outcomes overall. The positive relationship between high education level and better health outcomes has been extensively documented in scientific literature.

Question F. Investigate the relationship between PAM level, age, and education.

F. i. Create a graphical summary that shows the association between age and PAM level. Describe what you see.

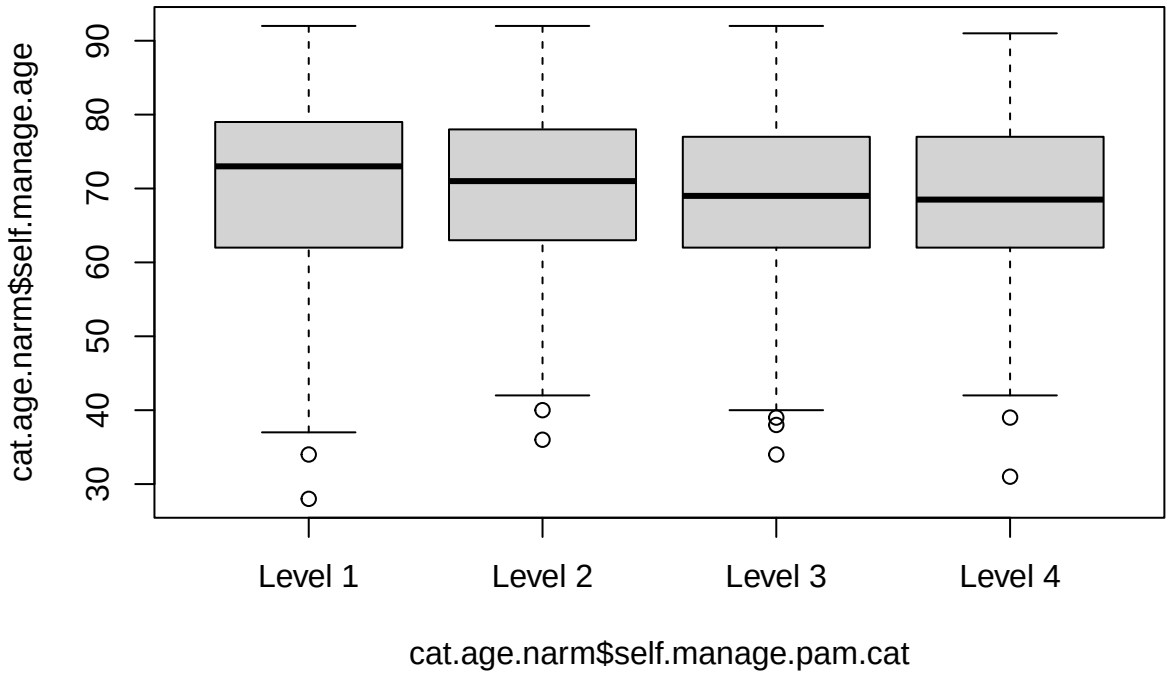
```
summary(self.manage$age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   28.00   63.00   70.00   69.64   78.00   92.00       3
```

```
summary(self.manage$pam.cat)
```

```
## Level 1 Level 2 Level 3 Level 4
##     308     323     381     142
```

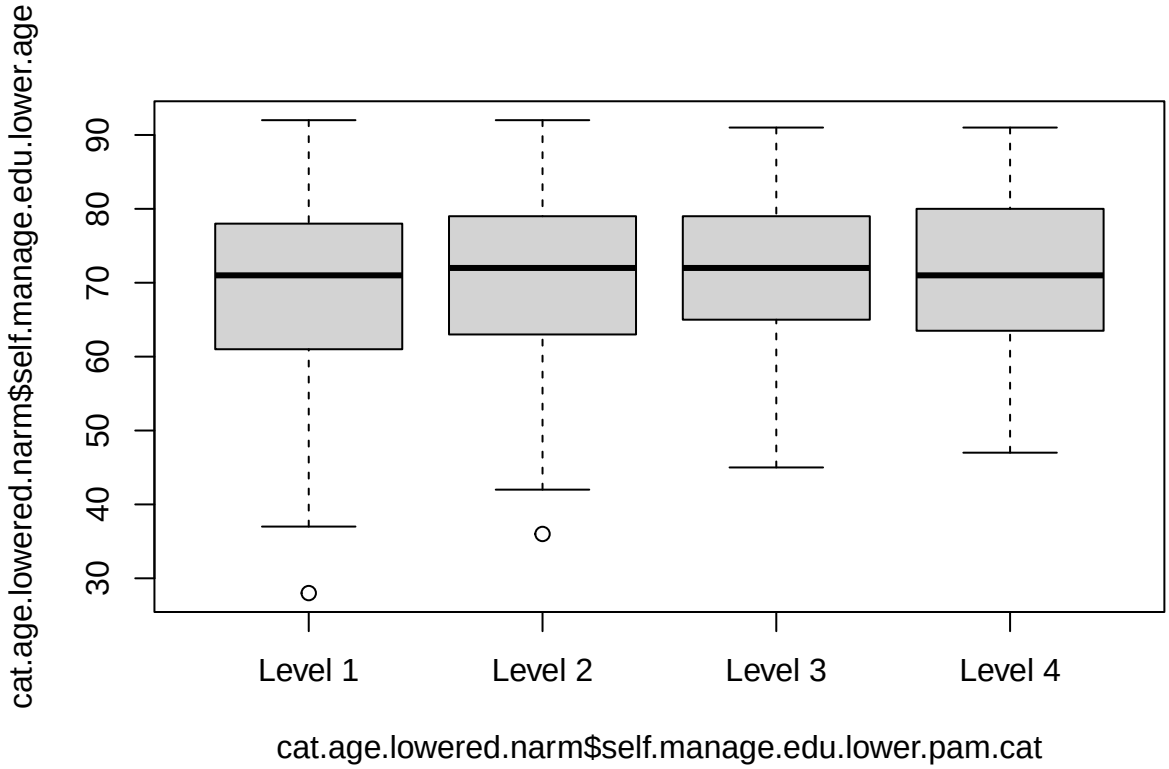
```
cat.age.narm = na.omit(data.frame(self.manage$age, self.manage$pam.cat))
boxplot(cat.age.narm$self.manage.age ~ cat.age.narm$self.manage.pam.cat)
```



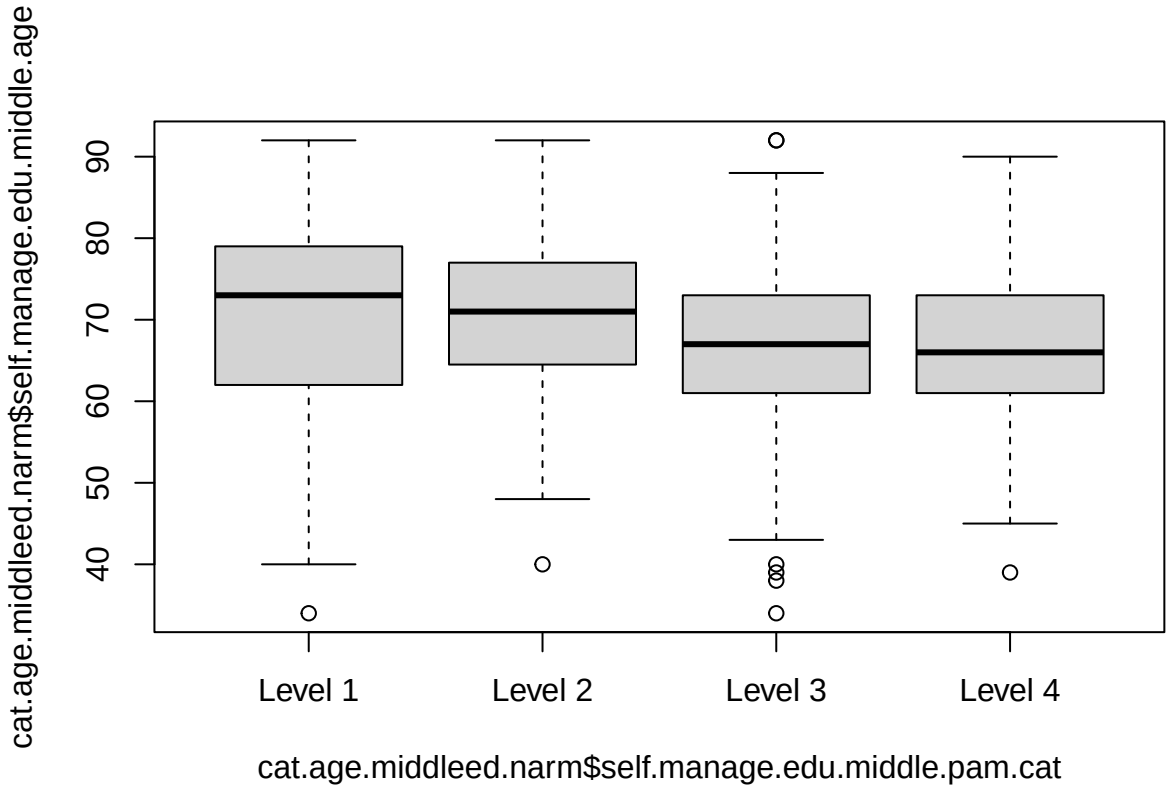
Answer: Based on the boxplot produced above, it looks like PAM score category increases as age decreases up until level 3, where there appears to be little difference between level 3 and level 4. The median age for level 1 scores looks higher than the median age for level 2 scores, and the median age for level 2 scores looks higher than the median age for level 3 scores. The median age for level 4 scores, however, looks about the same as the median age for level 3 scores. The interquartile ranges for age look approximately similar for each PAM score category.

F. ii. Create graphical summaries that show the association between age and PAM level when comparing individuals of the same educational level. Describe what you see.

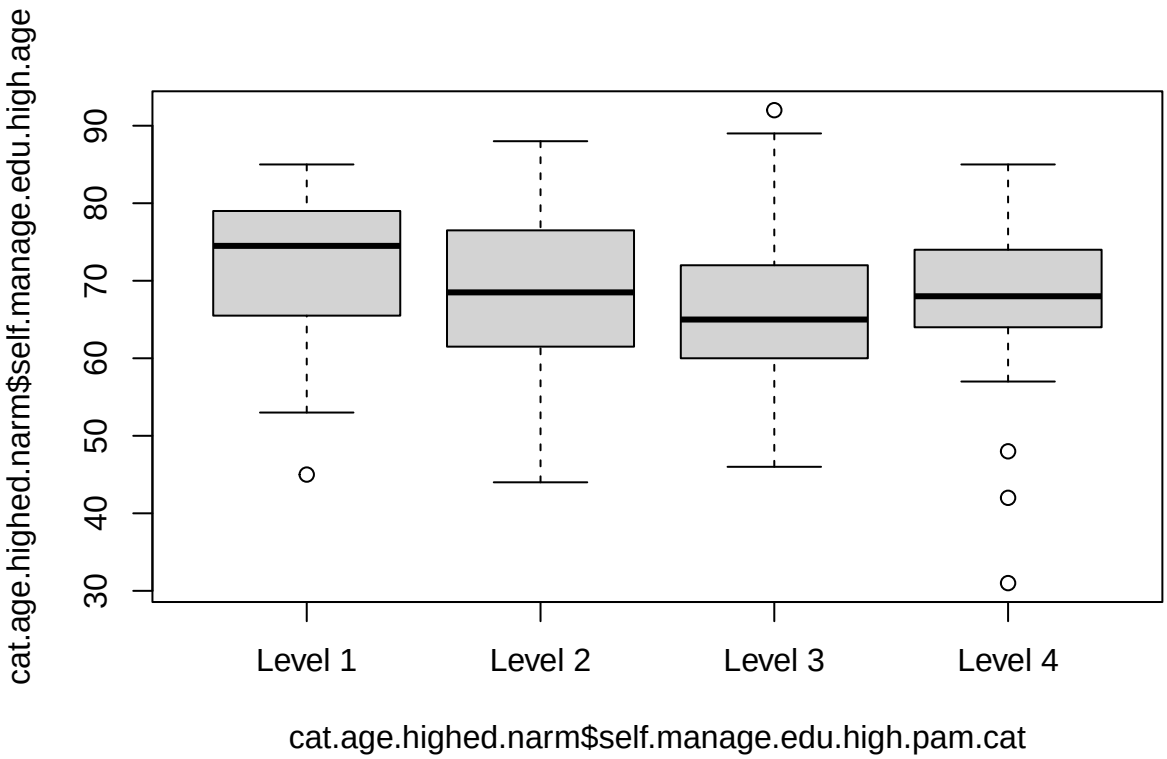
```
#Creating three dataframe subsets for each education category.
self.manage.edu.lower = subset(self.manage, self.manage$edu == "lower")
self.manage.edu.middle = subset(self.manage, self.manage$edu == "middle")
self.manage.edu.high = subset(self.manage, self.manage$edu == "high")

cat.age.lowered.narm = na.omit(data.frame(self.manage.edu.lower$age, self.manage.edu.lower$pam.cat, sel
cat.age.middleed.narm = na.omit(data.frame(self.manage.edu.middle$age, self.manage.edu.middle$pam.cat,
cat.age.highed.narm = na.omit(data.frame(self.manage.edu.high$age, self.manage.edu.high$pam.cat, self.ma

#Creating graphical summaries.
boxplot(cat.age.lowered.narm$self.manage.edu.lower.age ~ cat.age.lowered.narm$self.manage.edu.lower.pam
```




```
boxplot(cat.age.middleed.narm$self.manage.edu.middle.age ~ cat.age.middleed.narm$self.manage.edu.middle
```



```
boxplot(cat.age.highed.narm$self.manage.edu.high.age ~ cat.age.highed.narm$self.manage.edu.high.pam.cat)
```



Answer: When stratified by education level, the relationship between age and PAM score level looks different than analysis of all education levels.

For lower education patients, there appears to be a very minor increase in PAM level as age increases but it may not be statistically significant. Among PAM score groups, the minimum value for age increases as PAM score level increases. The median age for level 2 and level 3 are similar, and appear to be higher than the median age for level 1 and level 4.

The strongest linear relationship looks like it may be observed in the middle education patients, where there appears to be a clear slope indicating PAM score categories worsen as age increases.

There appears to be a negatively correlated relationship between age and PAM score category for high education as well, as the median age and interquartile ranges for age look like they decrease as PAM score category increases - that is, up until PAM level 4, when the median age jumps up again and the interquartile range of age increases. This relationship looks like may not be as strong as that observed in the middle education category.

F. iii. A news outlet is interested in reporting on the study results. You have been asked to address the following question: "Do these data suggest that older individuals tend to have a lower level of activation for self-management?" Address the question with an answer that is understandable to non-statisticians.

Answer: While the graphs suggest older individuals with middle and high education levels may have lower levels of activation for self-management, that doesn't appear to be as true for individuals with lower levels of education. The graph that looks at the entire population lumping in all education levels also seems to show higher age equals lower levels of activation for self-management, more rigorous analysis needs to be performed on these data to understand if those relationships are significant. In any case, even if they are significant, this isn't enough evidence to definitively say that, yes, older individuals have lower levels. More studies should be performed on different populations to see if the same results are found.

Question G. Explore the relationship between depression and activation for self-management.
G. i. From these data, compute the relative risk of being classified as PAM Level 1 comparing individuals considered to have a depressive disorder to those not considered to have a depressive disorder.

R code:

```
depress.pamcat = na.omit(data.frame(self.manage$hads.depress == c(0:10), self.manage$pam.cat))
```

```
## Warning in self.manage$hads.depress == c(0:10): longer object length is not a
## multiple of shorter object length
```

```
depress.pamcat.table = table(depress.pamcat$self.manage.hads.depress....c.0.10., depress.pamcat$self.ma
dimnames(depress.pamcat.table) = list("Depression state" = c("Not depressed", "Depressed"), "PAM catego
addmargins(depress.pamcat.table)
```

```
##                 PAM category
## Depression state Level 1 Level 2 Level 3 Level 4  Sum
##    Not depressed     274     288     335     130 1027
##    Depressed          27      30      33       8   98
##    Sum               301     318     368     138 1125
```

```
not.depressed.lvl1 = 274/1027
depressed.lvl1 = 27/98
depressed.lvl1/not.depressed.lvl1
```

```
## [1] 1.032661
```

Answer: The relative risk of obtaining a level 1 PAM score category is 1.03 when comparing patients with a depressive order to patients without a depressive order. This shows that there may be a slightly increased risk of scoring within level 1 if a patient has a depressive order.

G. ii. Calculate the difference in mean PAM-13 score between individuals classified as having a depressive disorder versus those not classified as having a depressive disorder. In one sentence, report the calculation; use phrasing that is informative for a general audience.

R code:

```
depress.pamscore = na.omit(data.frame(self.manage$hads.depress == c(0:10), self.manage$pam.score))
```

```
## Warning in self.manage$hads.depress == c(0:10): longer object length is not a
## multiple of shorter object length
```

```
tapply(depress.pamscore$self.manage.pam.score, depress.pamscore$self.manage.hads.depress....c.0.10., mea
```

```
##    FALSE     TRUE
## 54.20263 52.60000
```

Answer: On average, patients in the sample with depressive disorder scored 1.6 points lower on the PAM-13 questionnaire than patients without depressive disorder.

G. iii. There are some individuals missing responses for the HADS questionnaire. Does this missingness represent a potential source of bias for the calculations in parts i. and ii? Explain your answer.

Answer: In my analysis I removed any patients with missing responses in either PAM-13 score or HADS score. But without accounting for missing values, yes, they would represent a potential source of bias because missing values produce a less accurate calculation of sample statistics.

PROBLEM 2: Resilience. *Question A. Briefly summarize features of the study participants with respect to the variables age, sex, and train. Reference appropriate graphical and numerical summaries as needed.*

R code:

```
load("/Users/jm/Desktop/MPH/Spring 25 - Biostatistics in Public Health/Datasets/resilience.rdata")

summary(resilience$age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   17.00   21.00   22.00   22.76   24.00   40.00
```

```
summary(resilience$sex)
```

```
## female   male
##    714    636
```

```
summary(resilience$train)
```

```
## pre-clinical     clinical    residency
##          459          491          400
```

```
resilience.male = subset(resilience, resilience$sex == "male")
resilience.female = subset(resilience, resilience$sex == "female")
summary(resilience.male$age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   17.00   21.00   23.00   22.98   25.00   40.00
```

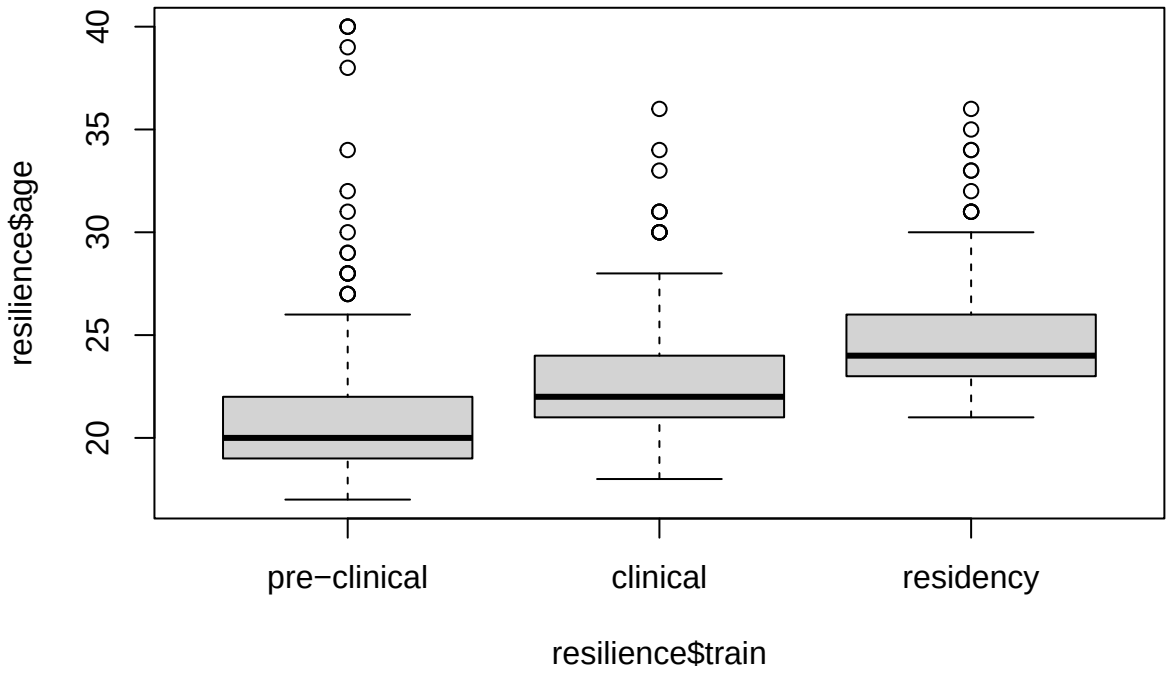
```
summary(resilience.female$age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   17.00   21.00   22.00   22.57   24.00   36.00
```

```
table(resilience$sex, resilience$train)
```

```
##
##          pre-clinical clinical residency
##   female          248      259       207
##   male            211      232       193
```

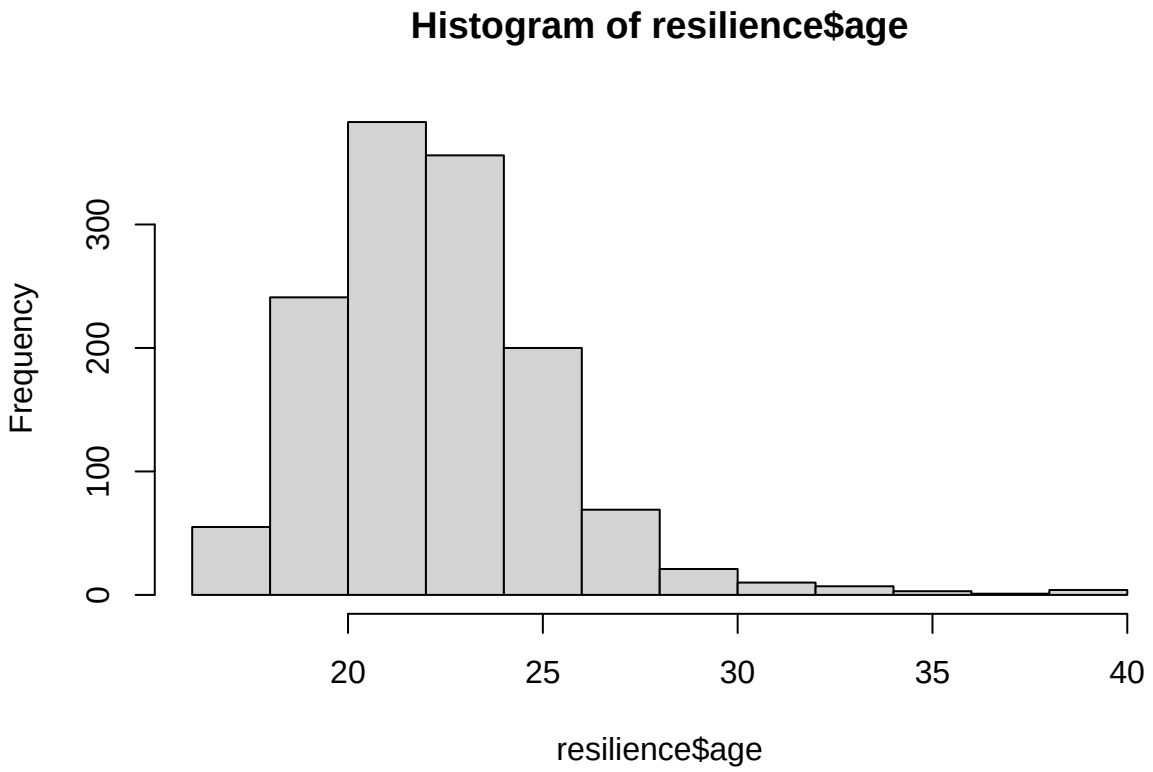
```
plot(resilience$age ~ resilience$train)
```




```
prop.table(table(resilience$train, resilience$sex), 2)
```

```
##
##                    female      male
##   pre-clinical 0.3473389 0.3317610
##   clinical     0.3627451 0.3647799
##   residency    0.2899160 0.3034591
```

```
hist(resilience$age)
```

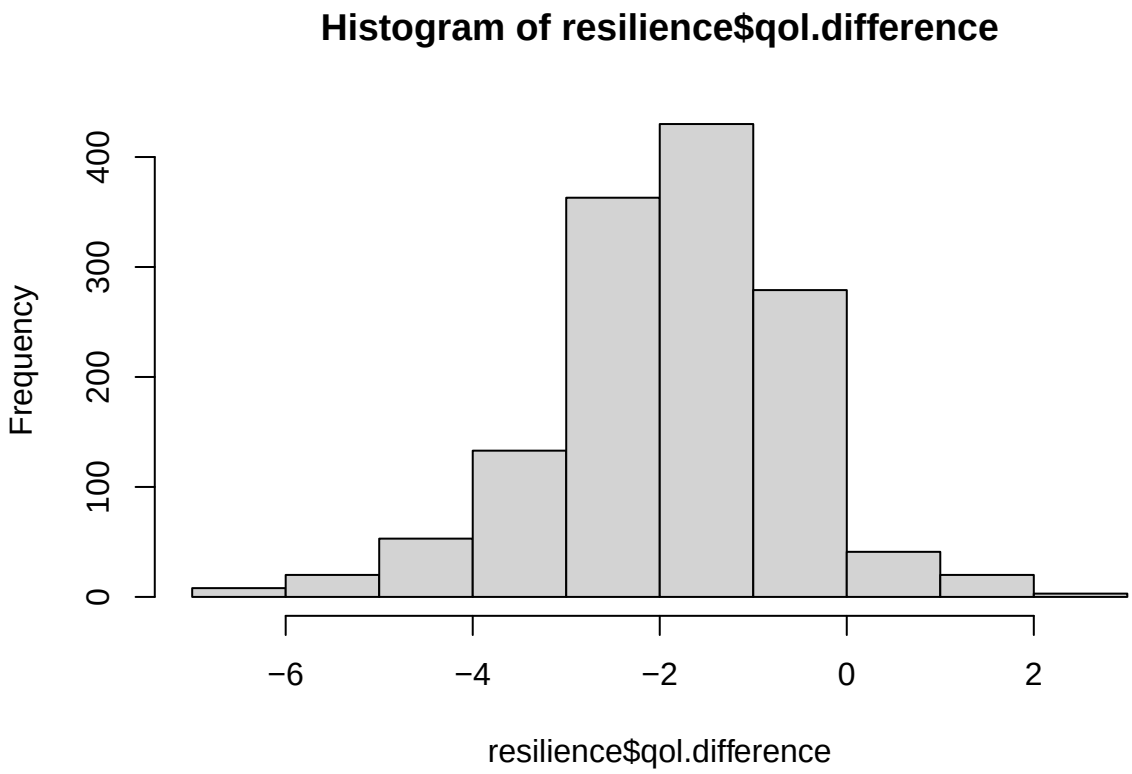


Answer: The sample is about 53% female and 47% male. Males on average tend to be very slightly older than females, with a median of 23 years (compared to 22 years) and a mean of 22.98 years (compared to 22.57 years). A greater proportion of males are in clinical and residency stages of training, while a greater proportion of females are in the pre-clinical stage. Unsurprisingly, sample age appears to increase from pre-clinical to clinical to residency training stages.

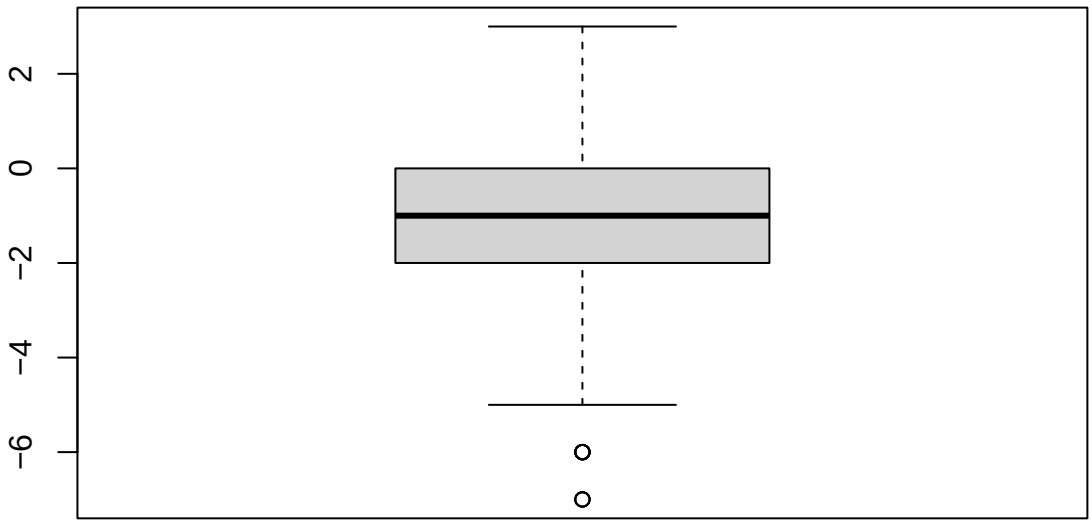
Question B. Participants were asked to rate their overall quality of life and their medical school quality of life, each on a 0-10 point scale.
B. i. Create a plot illustrating the difference between perception of overall QoL and perception of MSQoL. Describe what you see.

R code:

```
resilience$qol.difference <- resilience$qol.medical - resilience$qol.overall
hist(resilience$qol.difference)
```




```
boxplot(resilience$qol.difference)
```



Answer: A histogram displaying the frequency of differences between medical school QoL and overall QoL shows that overall QoL tends to be greater than medical school QoL. This is illustrated by the high frequency of negative differences when data are paired and the calculation of (medical QoL - overall QoL) is performed.

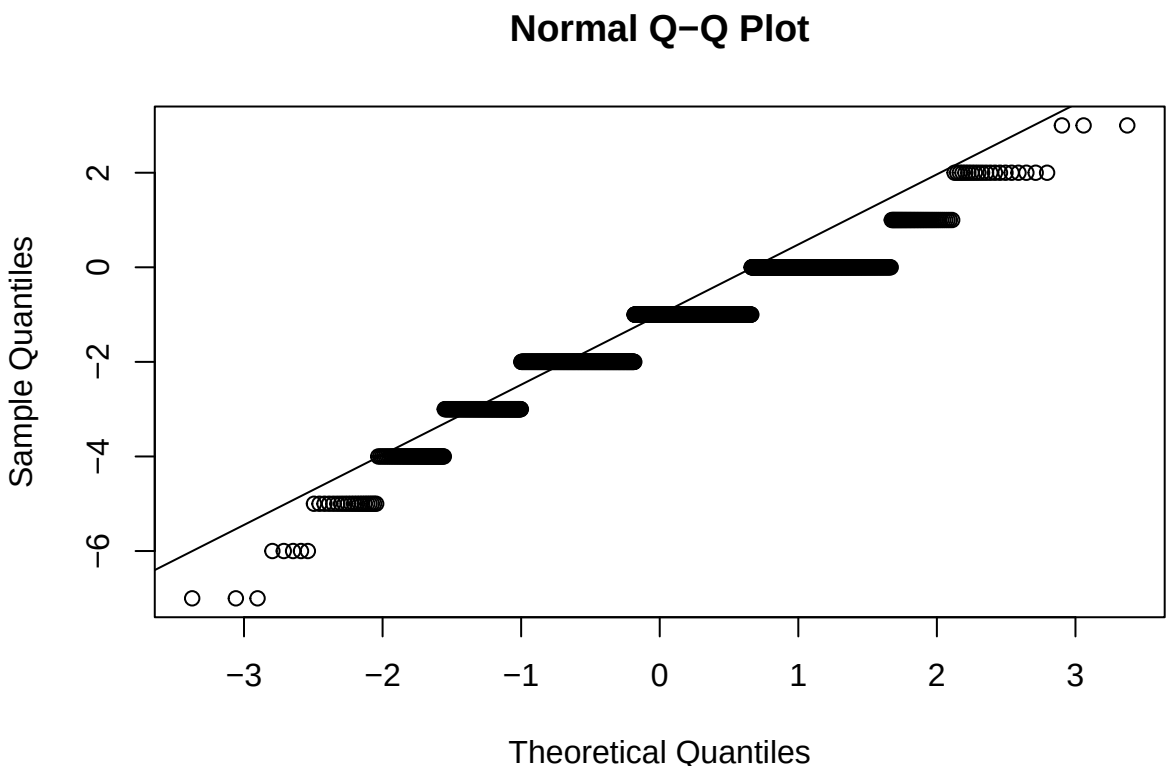
B. ii. Conduct a formal statistical comparison of overall QoL score and MSQoL score. Summarize your findings.

R code:

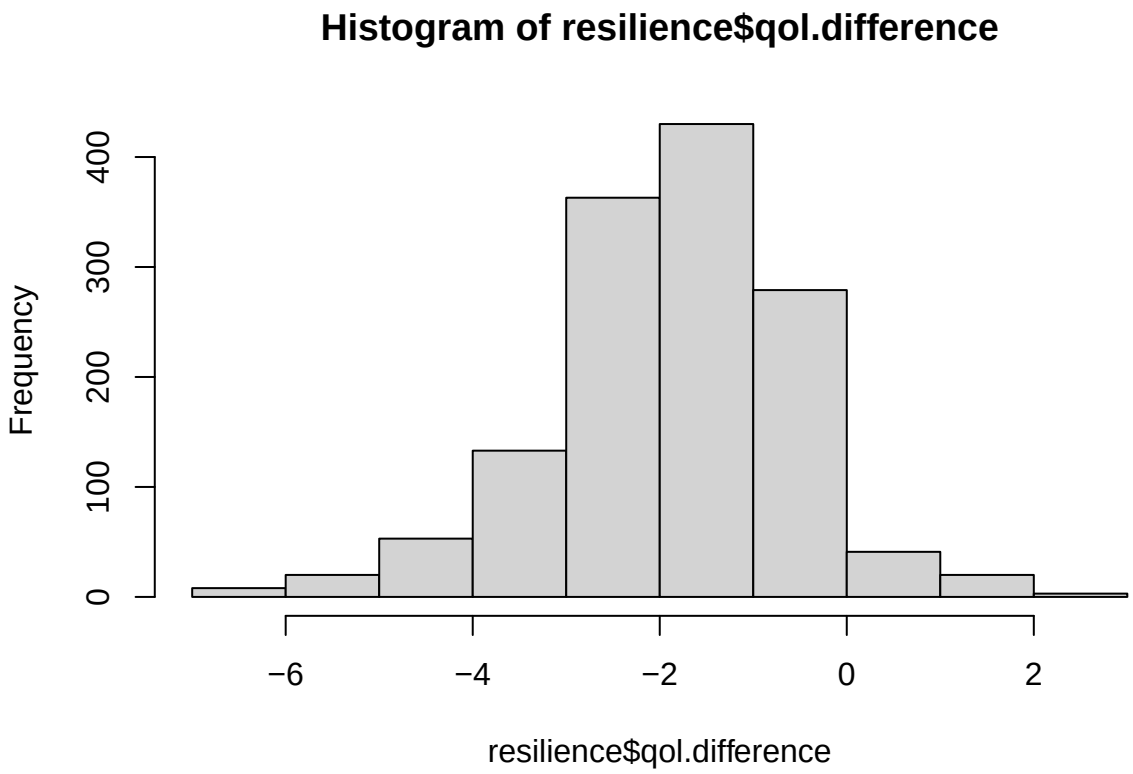
```
qol.model = t.test(resilience$qol.medical, resilience$qol.overall, alternative = "two.sided", paired = 
qol.model
```

```
##
##  Paired t-test
##
## data:  resilience$qol.medical and resilience$qol.overall
## t = -37.094, df = 1349, p-value < 2.2e-16
## alternative hypothesis: true mean difference is not equal to 0
## 95 percent confidence interval:
##  -1.425685 -1.282463
## sample estimates:
## mean difference
##       -1.354074
```

```
qqnorm(resilience$qol.difference)
qqline(resilience$qol.difference)
```




```
hist(resilience$qol.difference)
```



Answer: For this question, a t-test of two paired samples was performed to identify if the population mean of differences between the two scores observed in the sample is equal to 0.

The observations are assumed to be independent given that each participant in the sample is unique. The distribution of differences roughly follows a normal distribution, according to analysis with a histogram plot and a Q-Q plot. Students for the sample were randomly selected from each school, thus it is a random sample.

The t-test identified a t-statistic of -37.094 with a p-value significantly smaller than .01, indicating highly significant evidence that the population mean difference between overall QoL and medical QoL among medical students is not 0. The mean sample difference is -1.35, with a population 95% CI of (-1.43, -1.28), indicating that medical students may perceive a drop in life quality because of the demands of medical school when compared to their lives overall.

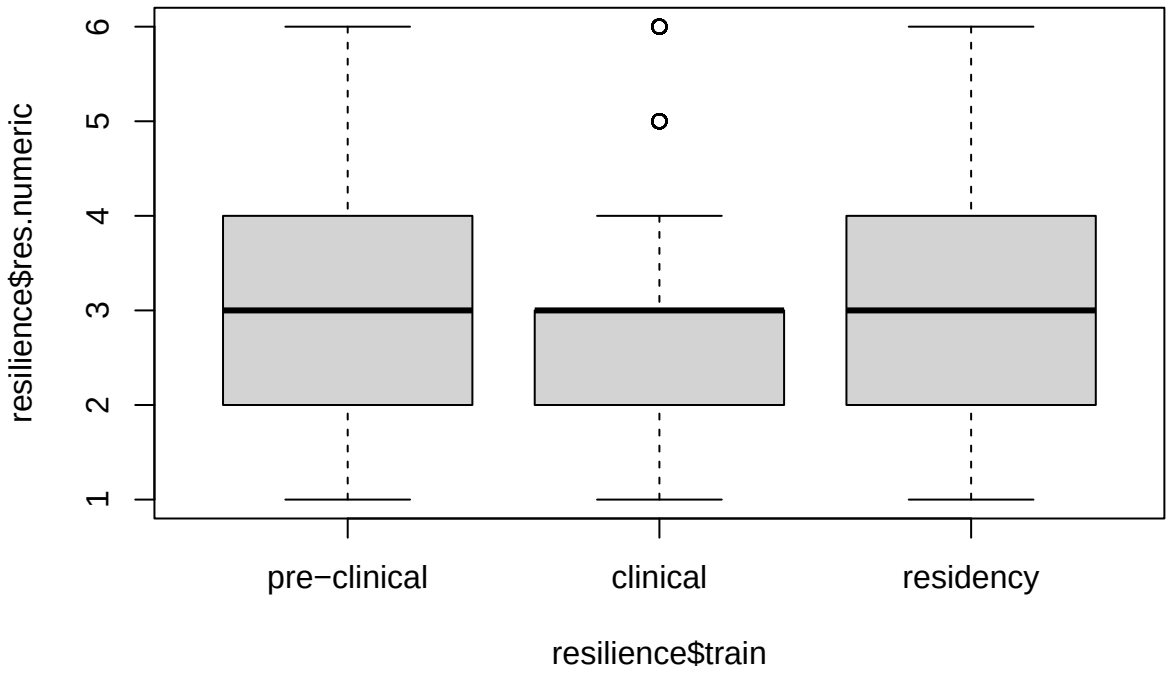
Question C. Investigate the relationship between resilience and level of training.
C. i. Prior to conducting any analysis, comment briefly on whether you think there may or may not be an association between resilience and level of training, and explain your reasoning. Limit your answer to at most five sentences.

Answer: I would assume that there is a positive relationship between resilience and level of training. It would seem that one would need to be more resilient to be able to effectively advance in medical training. On the other hand, it could also be that more advanced training teaches students stronger resiliency skills.

C. ii. Formally assess whether there is evidence of an association between resilience and level of training. Summarize your findings.

R code:

```
resilience$res.numeric <- as.numeric(resilience$res)
resilience$train.numeric <- as.numeric(resilience$train)
plot(resilience$res.numeric ~ resilience$train)
```



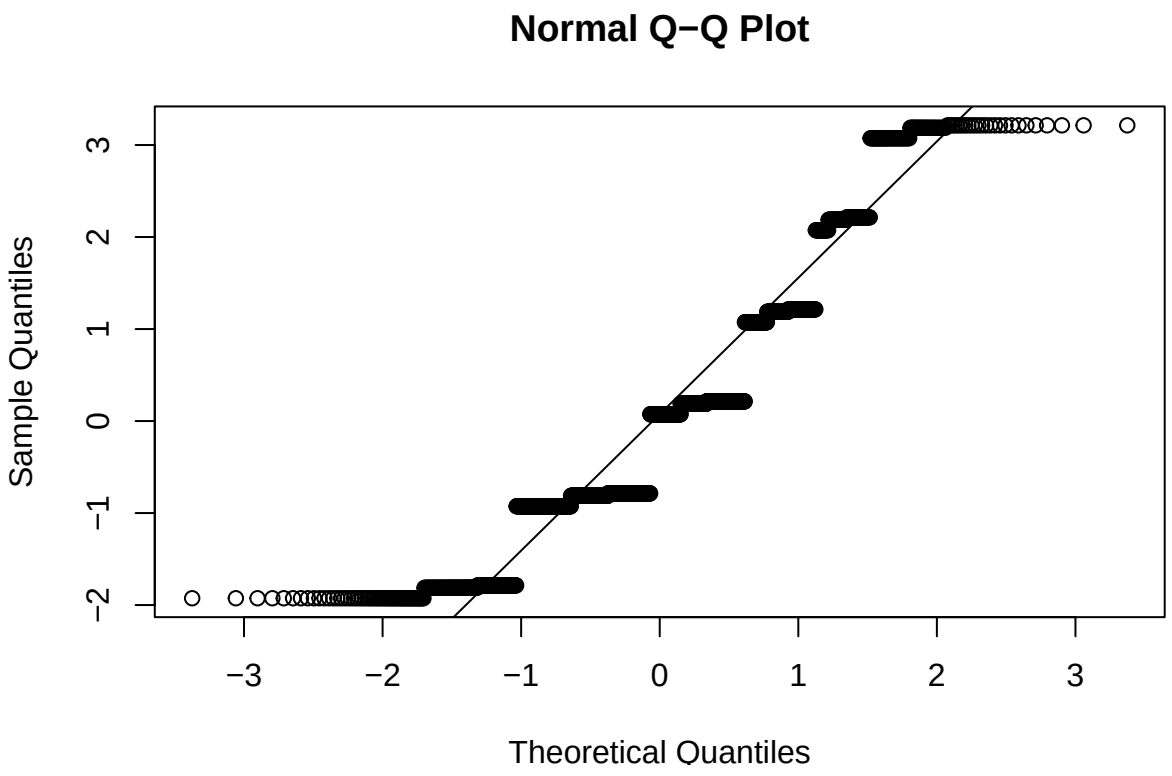

```
summary(lm(resilience$res.numeric ~ resilience$train))
```

```
##
## Call:
## lm(formula = resilience$res.numeric ~ resilience$train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.9259 -0.9259  0.0741  1.0741  3.2138
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 2.92593    0.06403  45.695   <2e-16 ***
## resilience$trainclinical   -0.13978    0.08907  -1.569    0.117
## resilience$trainresidency -0.11343    0.09383  -1.209    0.227
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.372 on 1347 degrees of freedom
## Multiple R-squared:  0.002013,   Adjusted R-squared:  0.0005308
## F-statistic: 1.358 on 2 and 1347 DF,  p-value: 0.2575
```

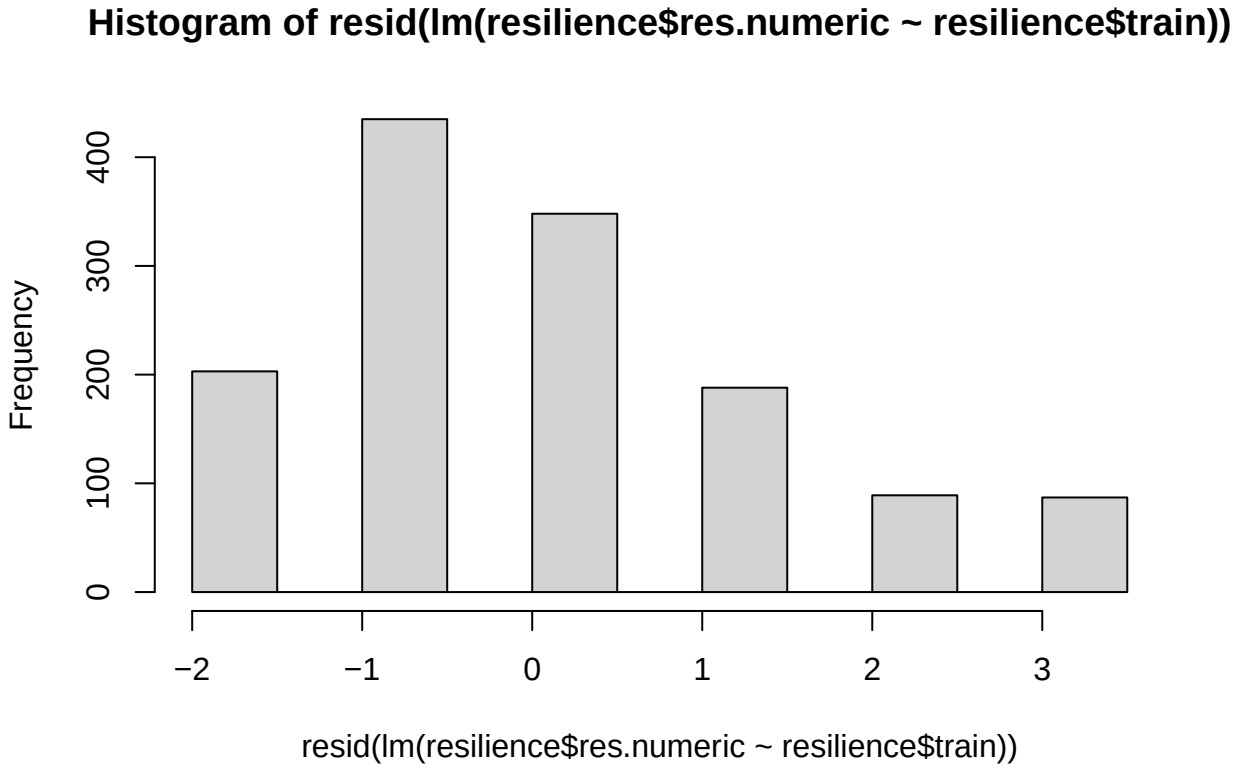
```
confint(lm(resilience$res.numeric ~ resilience$train), level = 0.95)
```

```
##                                2.5 %     97.5 %
## (Intercept)                2.8003130 3.05153885
## resilience$trainclinical  -0.3145003 0.03494992
## resilience$trainresidency -0.2975035 0.07065166
```

```
qqnorm(resid(lm(resilience$res.numeric ~ resilience$train)))
qqline(resid(lm(resilience$res.numeric ~ resilience$train)))
```




```
hist(resid(lm(resilience$res.numeric ~ resilience$train)))
```



Answer: There does not appear to be a relationship between resiliency and training. The f-statistic for the model is 1.358 with a p-value of 0.26, indicating no significant evidence that resilience scores are dependent on any of the training levels. The p-values for the relationship between clinical training level and residency training level are 0.117 and 0.227, respectively, indicating there is not significant evidence to reject the null

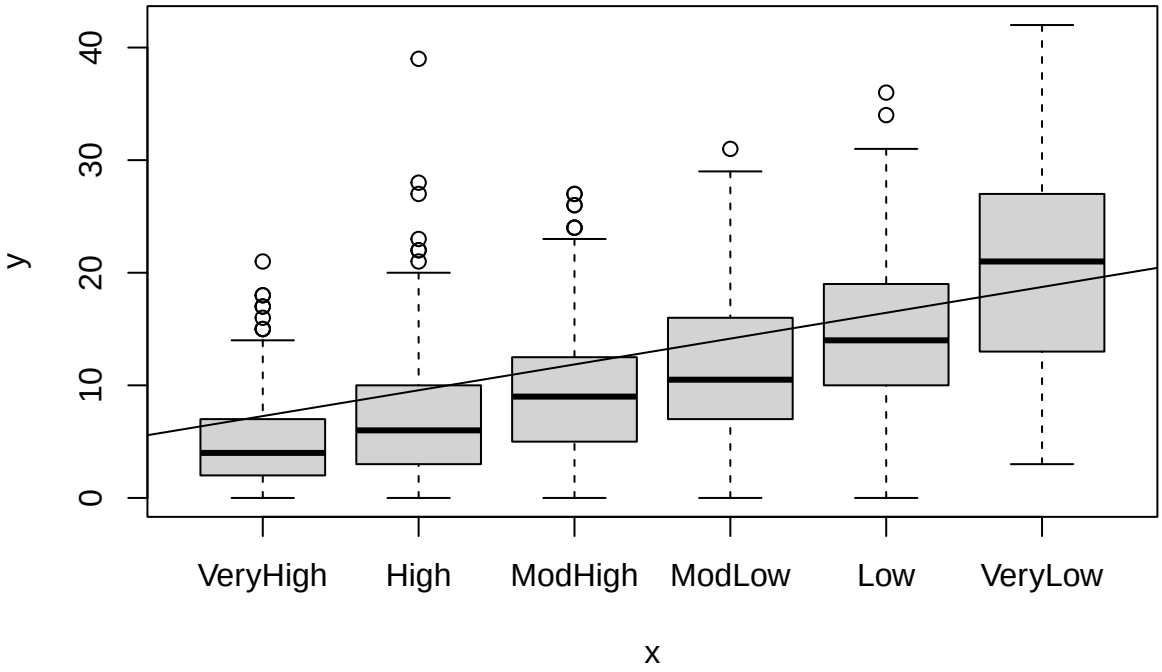
hypothesis of a population slope of 0. The 95% confidence intervals all include 0 within their margin of error, further confirming a lack of evidence to reject the null hypothesis.

Question D. Investigate the relationship between resilience and severity of depressive symptoms.
D. i. Create a plot illustrating the relationship between resilience and depressive symptoms. Describe what you see.

R code:

```
plot(resilience$res, resilience$bdi)
abline(lm(resilience$bdi ~ resilience$res))
```

```
## Warning in abline(lm(resilience$bdi ~ resilience$res)): only using the first
## two of 6 regression coefficients
```



Answer: A boxplot of the relationship between resiliency and BDI scores shows a clear linear relationship in which BDI scores decrease as resiliency increases.

D. ii. Conduct a formal analysis of the relationship between resilience and depressive symptoms. Summarize your findings. You may proceed with the analysis method you choose even if the assumptions do not seem to be reasonably satisfied; i.e., it is not necessary to check assumptions for this sub-question.

R code:

```
summary(lm(resilience$bdi ~ resilience$res))
```

```
##
## Call:
## lm(formula = resilience$bdi ~ resilience$res)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -17.621  -4.269  -0.686   3.314  31.731
```

```
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)              4.9754     0.4118  12.083  < 2e-16 ***
## resilience$resHigh       2.2936     0.4987   4.599 4.64e-06 ***
## resilience$resModHigh    4.3005     0.5181   8.300 2.51e-16 ***
## resilience$resModLow     6.7108     0.5938  11.301  < 2e-16 ***
## resilience$resLow        9.6538     0.7458  12.943  < 2e-16 ***
## resilience$resVeryLow   15.6453     0.7518  20.811  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.867 on 1344 degrees of freedom
## Multiple R-squared:  0.3052, Adjusted R-squared:  0.3027
## F-statistic: 118.1 on 5 and 1344 DF,  p-value: < 2.2e-16
```

```
confint(lm(resilience$bdi ~ resilience$res), level = 0.95)
```

```
##                            2.5 %    97.5 %
## (Intercept)            4.167590  5.783149
## resilience$resHigh     1.315326  3.271866
## resilience$resModHigh  3.284059  5.316926
## resilience$resModLow   5.545864  7.875738
## resilience$resLow      8.190692 11.116996
## resilience$resVeryLow 14.170523 17.120117
```

Answer: Linear regression analysis shows that there is a clear relationship between resiliency and BDI score. More specifically, BDI and resiliency are negatively correlated, where every increase in resiliency level corresponds with a significant decrease in BDI.

The p-value of the test for every resiliency category is significantly less than 0.01, indicating strong evidence to reject the null hypothesis that the population slope for any of the linear models is 0. The F-statistic is 118.5 with a p-value of less than 0.01, confirming that BDI is dependent on at least one of the categorical variables of resiliency level. None of the confidence intervals (95%) for any of the resiliency category slopes include 0, confirming the hypothesis test findings.

Question E. Investigate the association between resilience and quality of life as measured by the psychological health domain of the WHOQOL.
E. i. Without adjusting for any potential confounders, fit a model estimating the association between resilience and WHOQOL score in the psychological health domain. Describe the nature of the association.

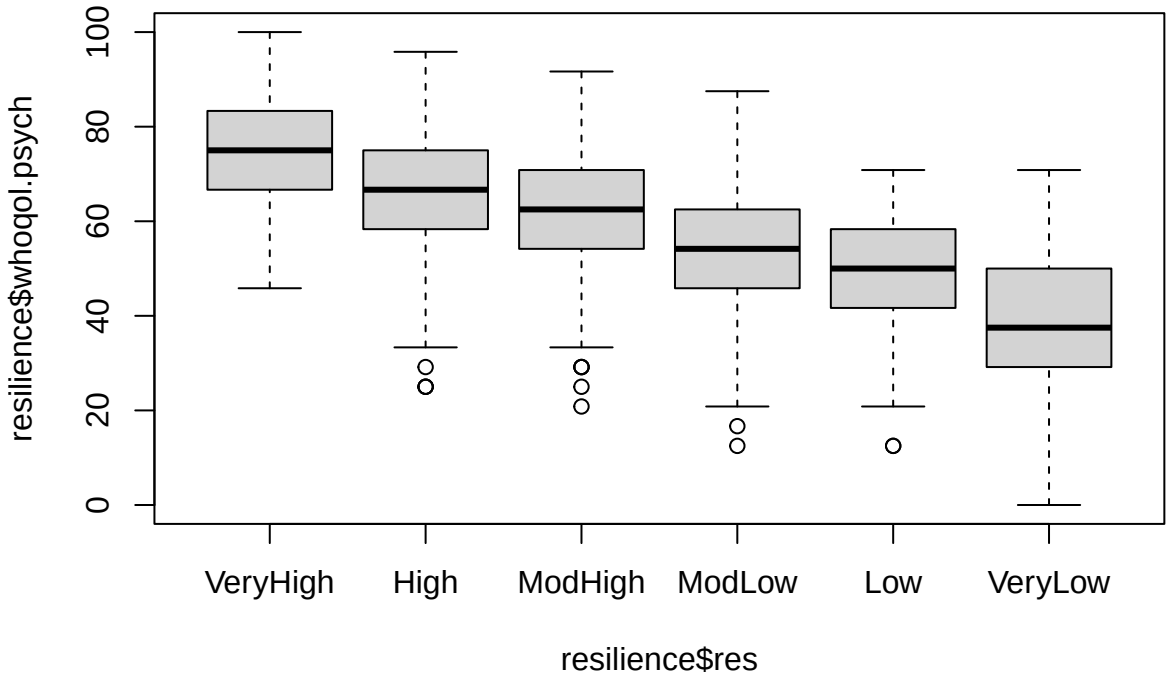
R code:

```
summary(lm(resilience$whoqol.psych ~ resilience$res))
```

```
##
## Call:
## lm(formula = resilience$whoqol.psych ~ resilience$res)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -42.031  -8.026   0.314   8.644  33.644
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)            75.0206     0.8718  86.053  < 2e-16 ***
## resilience$resHigh     -7.9899     1.0558  -7.568 7.02e-14 ***
## resilience$resModHigh -14.1131     1.0970 -12.865  < 2e-16 ***
## resilience$resModLow  -21.1642     1.2573 -16.834  < 2e-16 ***
## resilience$resLow     -26.5655     1.5791 -16.823  < 2e-16 ***
## resilience$resVeryLow -37.0897     1.5917 -23.302  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.42 on 1344 degrees of freedom
## Multiple R-squared:  0.376,  Adjusted R-squared:  0.3736
## F-statistic: 161.9 on 5 and 1344 DF,  p-value: < 2.2e-16
```

```
plot(resilience$whoqol.psych ~ resilience$res)
```




```
confint(lm(resilience$whoqol.psych ~ resilience$res), level = 0.95)
```

```
##                            2.5 %     97.5 %
## (Intercept)            73.31037  76.730811
## resilience$resHigh    -10.06108  -5.918722
## resilience$resModHigh -16.26510 -11.961140
## resilience$resModLow  -23.63060 -18.697820
## resilience$resLow     -29.66330 -23.467770
## resilience$resVeryLow -40.21209 -33.967251
```

Answer: There is significant statistical evidence that resiliency and WHOQOL psychological are associated. More specifically, they are positively associated, in that increases in resiliency levels are associated with increases in WHOQOL psychological score (higher WHOQOL psychological score indicates higher quality of life).

The F-statistic of the model analysis is 161.9 with a p-value of less than .01, indicating that WHOQOL psychological is dependent in at least one of the resiliency level analyses. The p-value for the hypothesis test among each of the resiliency levels is less than .001, indicating there is enough evidence to reject the null hypothesis of a slope parameter of 0 in any of the model equations.

E. ii. Is there evidence that resilience overall is a useful variable for predicting WHOQOL score in the psychological health domain? Explain your answer.

Answer: Yes, there is evidence that resilience overall may be a useful variable for predicting WHOQOL. The statistical significance of the hypothesis test findings are strong at every resiliency level. None of the confidence intervals calculated for the resiliency categories includes 0, confirming the hypothesis test findings. There is a clear visual trend observed when plotting the data in boxplots. However, it is worth analyzing whether other variables are playing a role in this relationship. Socio-economic variables such as income (not recorded in this sample) may influence medical students' resiliency and WHOQOL score, for example.

E. iii. Report and interpret the model R2 for the model fit in part i.

Answer: The multiple R2 for the model in part i. is 0.376, indicating that 37.6% of the variation in WHOQOL psychological scores are explained by resiliency levels. The adjusted R-squared is 0.374. These scores suggest that it may be worth analyzing additional independent variables (in new models) to find variables that explain a larger percent of the variation in WHOQOL psychological scores.