Lab 3

Ben Chu

February 6, 2018

Loading data and packages.

```
load("C:/Users/Branly Mclanbry/Downloads/lab3.RData")
lab3 <- lab3 %>% janitor::clean_names()
```

Q1

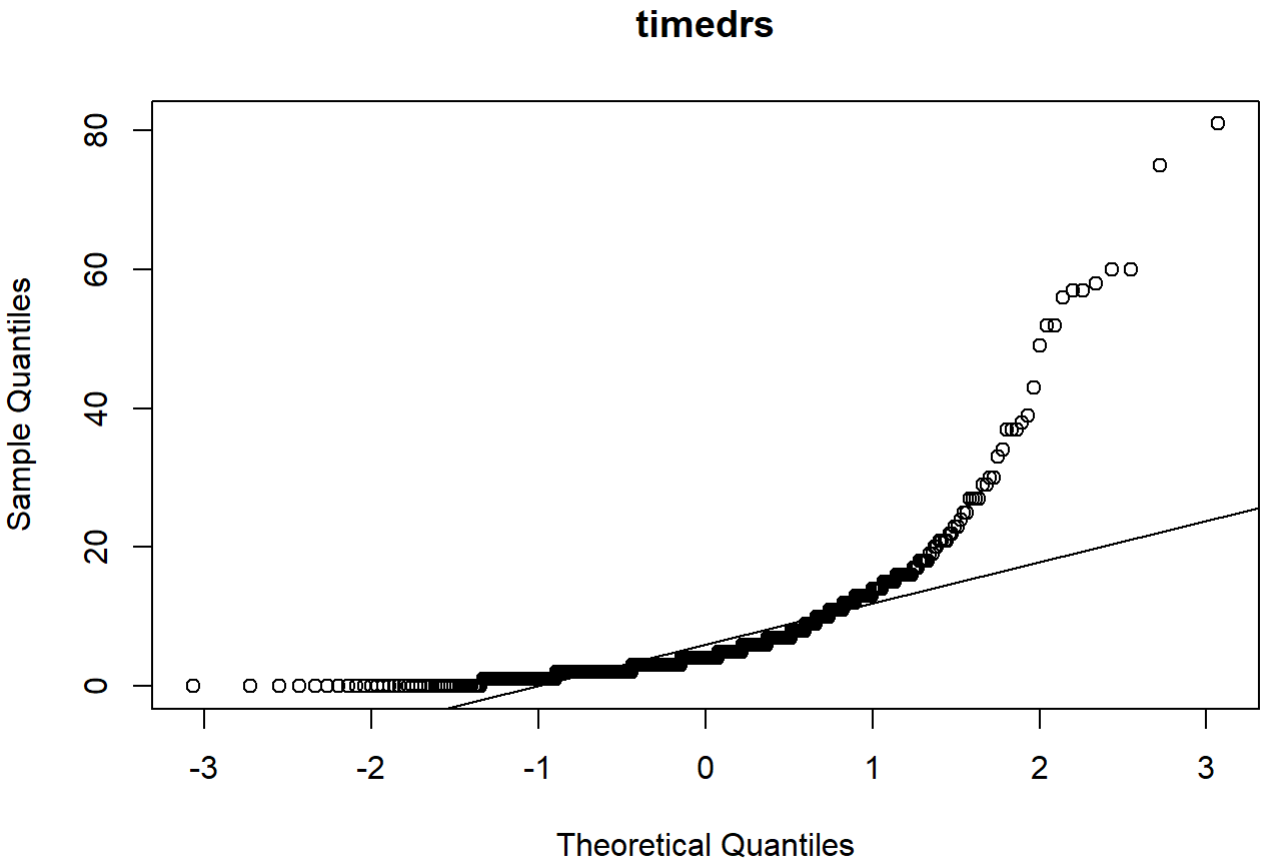
It appears that all variables are positively skewed. This is apparent in the 99% confidence intervals because 0 is no within the range. Similarly, kurtosis was an issue for all variables excelt mental health. ###Descriptive statistics

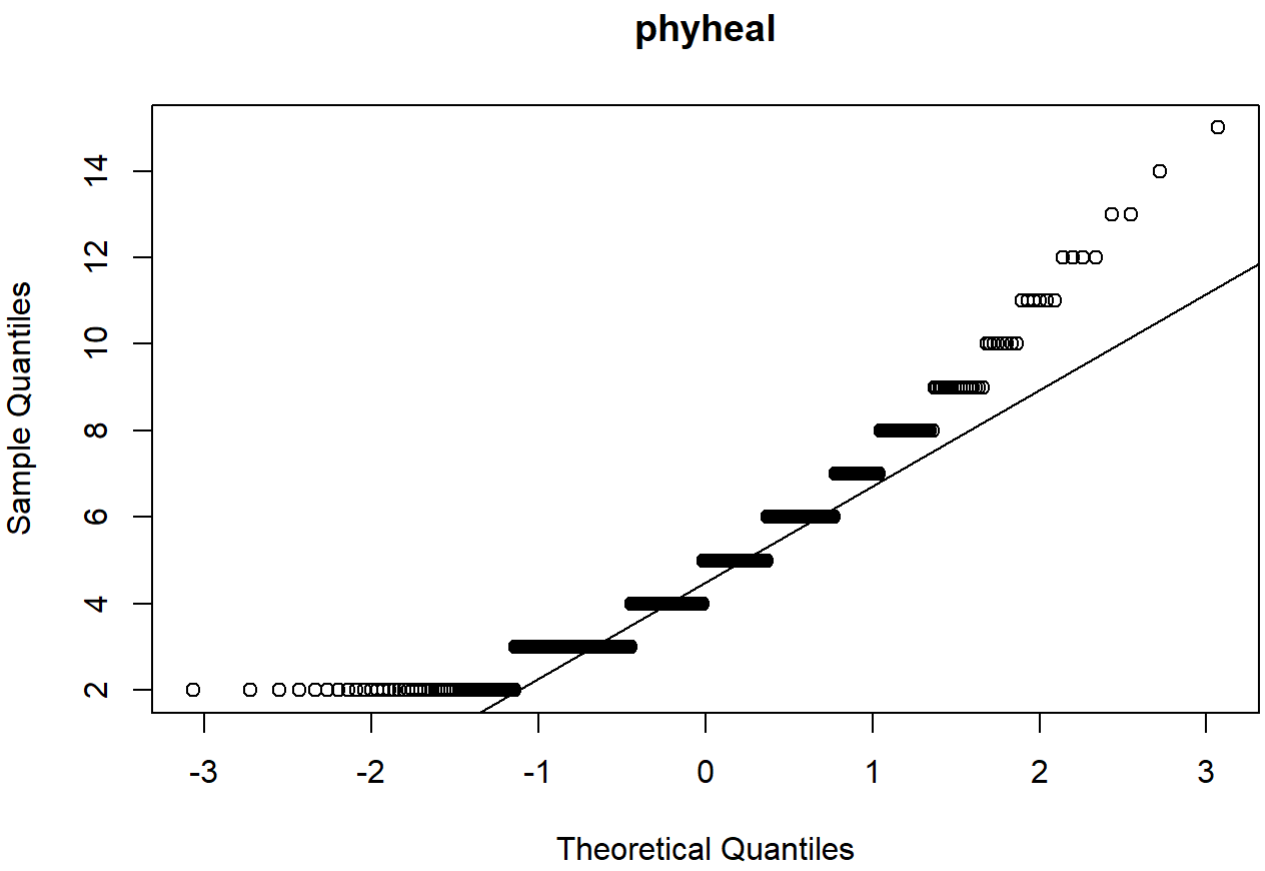
```
##         vars   n   mean     sd median trimmed    mad min max range skew
## subjno     1 465 317.38 194.16    314  313.26 256.49   1 758   757 0.14
## timedrs    2 465   7.90  10.95      4    5.61   4.45   0  81    81 3.23
## phyheal    3 465   4.97   2.39      5    4.72   2.97   2  15    13 1.02
## menheal    4 465   6.12   4.19      6    5.81   4.45   0  18    18 0.60
## stress     5 465 204.22 135.79    178  191.74 133.43   0 920   920 1.04
##         kurtosis   se
## subjno     -0.99 9.00
## timedrs    12.88 0.51
## phyheal     1.08 0.11
## menheal    -0.31 0.19
## stress      1.75 6.30
```

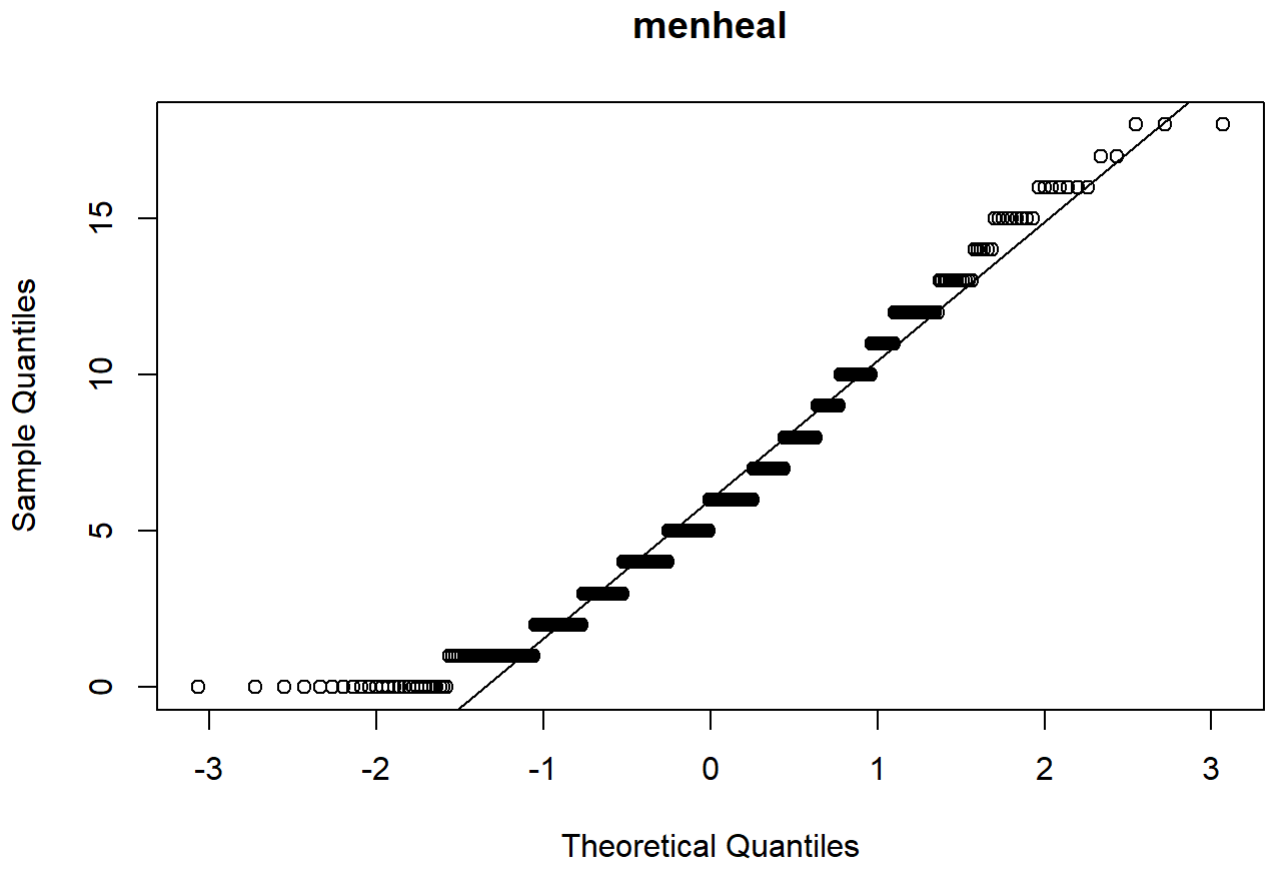
Functions for plots

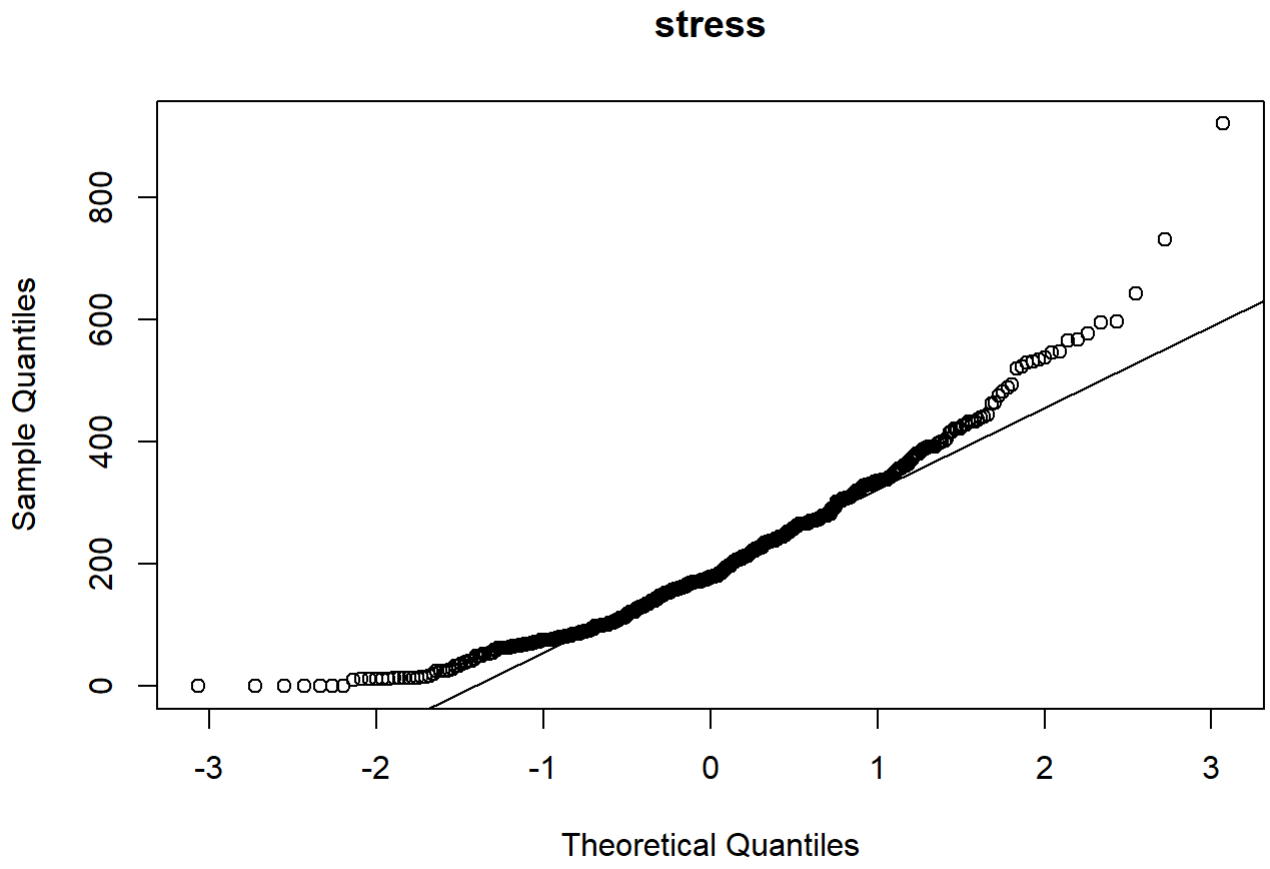
QQ Plots

```
walk2(p_list,names,pphehe)
```


timedrs
Sample Quantiles
0
20
40
60
80
-3
-2
-1
0
1
2
3
Theoretical Quantiles

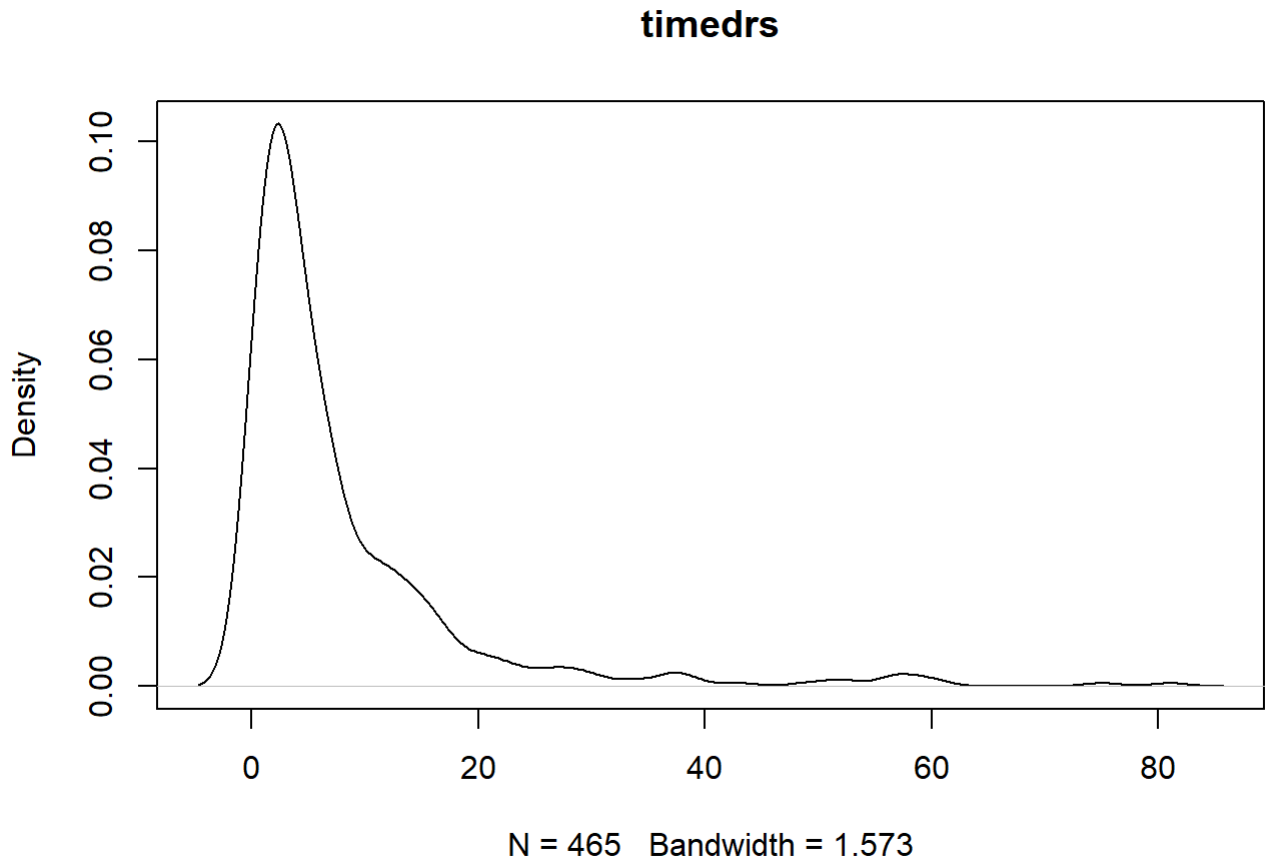

phyheal
Sample Quantiles
2
4
6
8
10
12
14
-3
-2
-1
0
1
2
3
Theoretical Quantiles

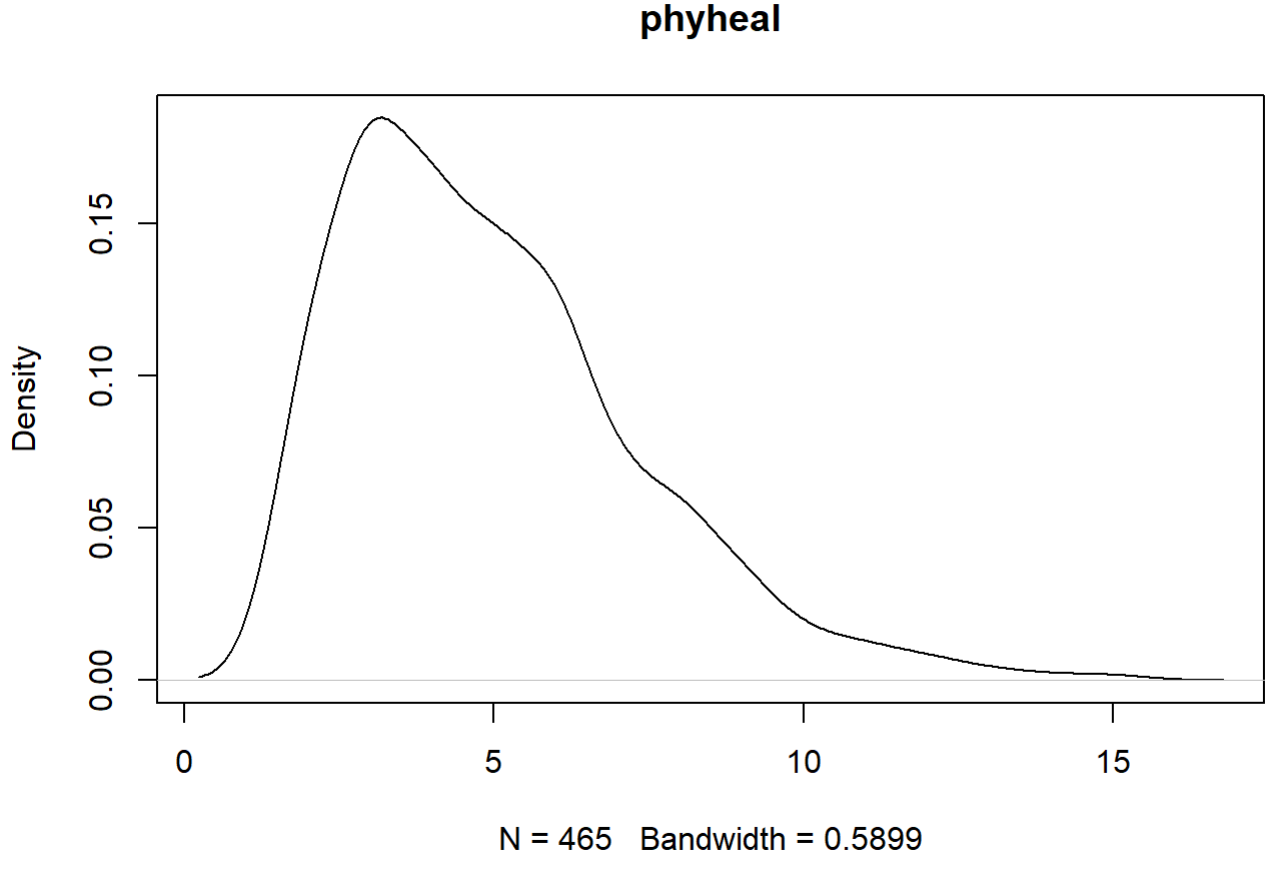

menheal
Sample Quantiles
0
5
10
15
-3
-2
-1
0
1
2
3
Theoretical Quantiles

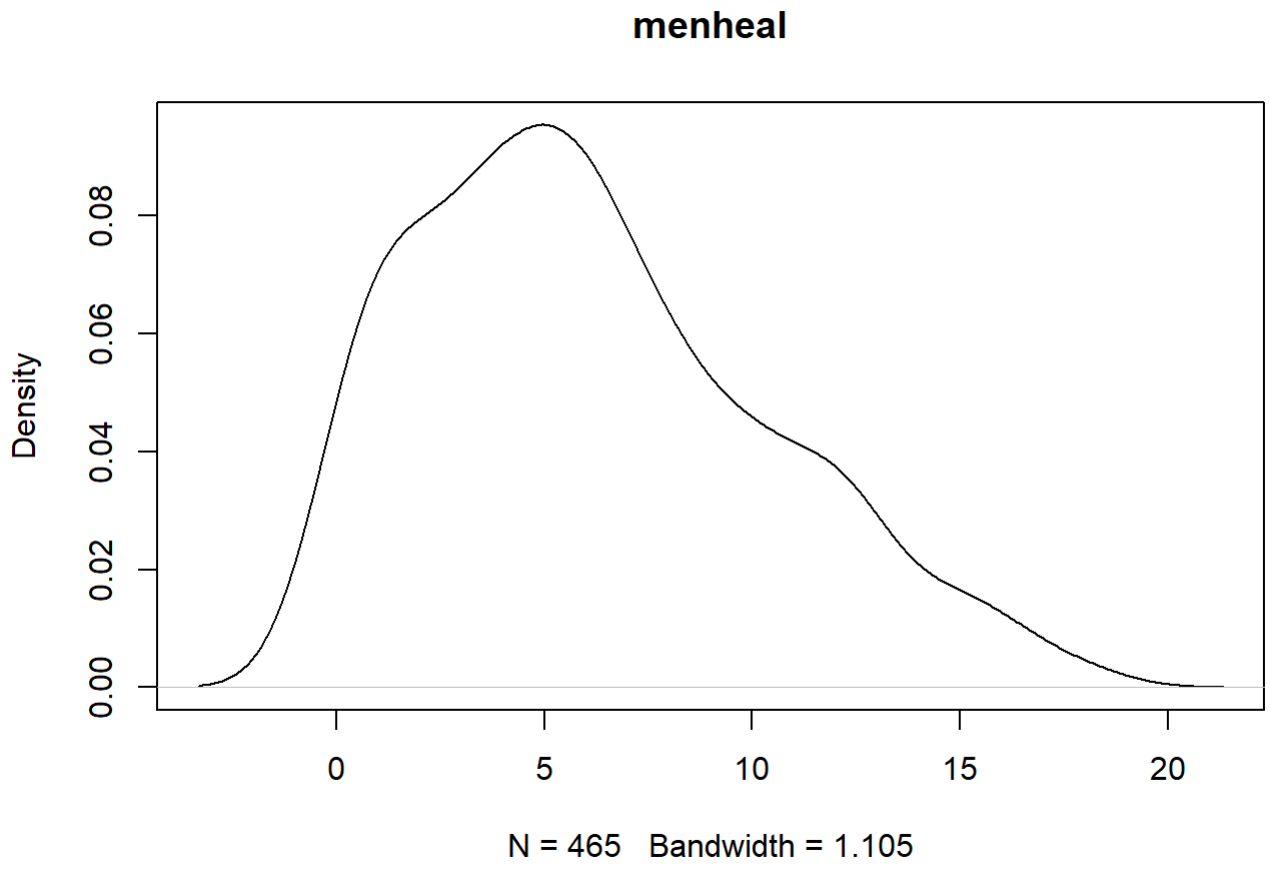

stress
Sample Quantiles
0
200
400
600
800
-3
-2
-1
0
1
2
3
Theoretical Quantiles

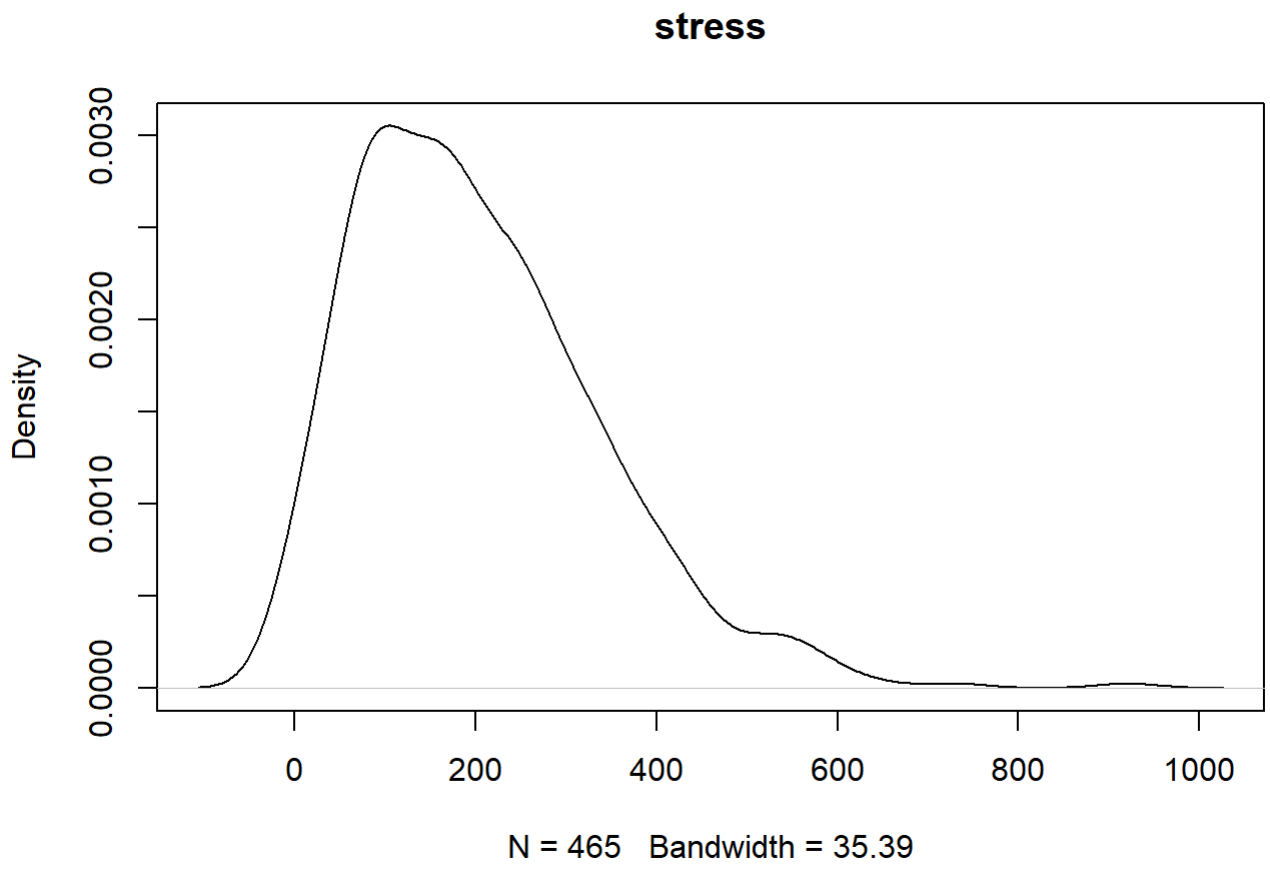
Density Plots

```
walk2(p_list,names,denss)
```


timedrs
Density
0.00
0.02
0.04
0.06
0.08
0.10
0
20
40
60
80
N = 465 Bandwidth = 1.573


phyheal
Density
0.00
0.05
0.10
0.15
0
5
10
15
N = 465 Bandwidth = 0.5899


menheal
Density
0.00
0.02
0.04
0.06
0.08
0
5
10
15
20
N = 465 Bandwidth = 1.105


stress
Density
0.0000
0.0010
0.0020
0.0030
0
200
400
600
800
1000
N = 465 Bandwidth = 35.39

Function for Skew and Kurtosis

```
skurt <-function(x,var) {
  skew.1 <- round(DescTools::Skew(x, method = 2, conf.level = .99),2)
  print(var)
  kurt.1 <- round(DescTools::Kurt(x, method = 2, conf.level = .99),2)
  print(list(skew.1,kurt.1))
}
```

Skew and Kurt for all variables

```
walk2(p_list,names,skurt)
```

```
## [1] "timedrs"
## [[1]]
##   skew lwr.ci upr.ci
##   3.25   2.68   4.29
##
## [[2]]
##   kurt lwr.ci upr.ci
##  13.10   7.73  22.79
##
## [1] "phyheal"
## [[1]]
##   skew lwr.ci upr.ci
##   1.03   0.76   1.38
##
## [[2]]
##   kurt lwr.ci upr.ci
##   1.12   0.07   2.71
##
## [1] "menheal"
## [[1]]
##   skew lwr.ci upr.ci
##   0.60   0.41   0.83
##
## [[2]]
##   kurt lwr.ci upr.ci
##  -0.29  -0.64   0.29
##
## [1] "stress"
## [[1]]
##   skew lwr.ci upr.ci
##   1.04   0.69   1.60
##
## [[2]]
##   kurt lwr.ci upr.ci
##   1.80   0.08   5.15
```

Function for transofmrations

```
transform <- function(x,var){
  print(var)
  print("squareroot")
  squareroot <- (x+1)^.5
  print(round(DescTools::Skew(squareroot,na.rm=TRUE, method=2,conf.level=.99),2))
  print("log")
  log <- log10(x+1)
  print(round(DescTools::Skew(log,na.rm=TRUE, method=2,conf.level=.99),2))
  print("inverse")
  inverse <- 1/(x+1)
  print(round(DescTools::Skew(inverse,na.rm=TRUE, method=2,conf.level=.99),2))
    print(var)
  print("squareroot")
  squareroot <- (x+1)^.5
  print(round(DescTools::Kurt(squareroot,na.rm=TRUE, method=2,conf.level=.99),2))
  print("log")
  log <- log10(x+1)
  print(round(DescTools::Kurt(log,na.rm=TRUE, method=2,conf.level=.99),2))
  print("inverse")
  inverse <- 1/(x+1)
  print(round(DescTools::Kurt(inverse,na.rm=TRUE, method=2,conf.level=.99),2))
}
```

Q2

Transformations

```
walk2(p_list,names,transform)
```

```
## [1] "timedrs"
## [1] "squareroot"
##   skew lwr.ci upr.ci
##   1.63   1.26   2.05
## [1] "log"
##   skew lwr.ci upr.ci
##   0.23   0.04   0.44
## [1] "inverse"
##   skew lwr.ci upr.ci
##   1.75   1.49   2.04
## [1] "timedrs"
## [1] "squareroot"
##   kurt lwr.ci upr.ci
##   3.41   1.86   5.90
## [1] "log"
##   kurt lwr.ci upr.ci
##  -0.18  -0.52   0.27
## [1] "inverse"
##   kurt lwr.ci upr.ci
##   2.35   1.03   4.11
## [1] "phyheal"
## [1] "squareroot"
##   skew lwr.ci upr.ci
##   0.57   0.37   0.81
## [1] "log"
##   skew lwr.ci upr.ci
##   0.16  -0.03   0.36
## [1] "inverse"
##   skew lwr.ci upr.ci
##   0.53   0.37   0.71
## [1] "phyheal"
## [1] "squareroot"
##   kurt lwr.ci upr.ci
##  -0.08  -0.55   0.79
## [1] "log"
##   kurt lwr.ci upr.ci
##  -0.64  -0.88  -0.28
## [1] "inverse"
##   kurt lwr.ci upr.ci
##  -0.61  -0.91  -0.27
## [1] "menheal"
## [1] "squareroot"
##   skew lwr.ci upr.ci
##  -0.02  -0.20   0.16
## [1] "log"
##   skew lwr.ci upr.ci
##  -0.75  -0.96  -0.56
## [1] "inverse"
##   skew lwr.ci upr.ci
##   2.32   1.96   2.72
## [1] "menheal"
## [1] "squareroot"
##   kurt lwr.ci upr.ci
```

```
##  -0.67  -0.89  -0.42
## [1] "log"
##   kurt lwr.ci upr.ci
##   0.02  -0.42   0.51
## [1] "inverse"
##   kurt lwr.ci upr.ci
##   5.08   2.71   7.61
## [1] "stress"
## [1] "squareroot"
##   skew lwr.ci upr.ci
##  -0.04  -0.27   0.31
## [1] "log"
##   skew lwr.ci upr.ci
##  -2.22  -2.84  -1.66
## [1] "inverse"
##   skew lwr.ci upr.ci
##   7.85   4.15  14.18
## [1] "stress"
## [1] "squareroot"
##   kurt lwr.ci upr.ci
##   0.00  -0.38   0.81
## [1] "log"
##   kurt lwr.ci upr.ci
##   7.87   4.85  11.07
## [1] "inverse"
##   kurt lwr.ci upr.ci
##  60.62  15.71 214.67
```

That's alot of numbers. At any rate, here are the best ones. `time at doctors` with a `log. Physical health` with `log. Mental Health` with `square root` and `mental health` with `square root` here's some code that adds it into the data file

```
lab3 <- lab3 %>%
  mutate(timedrs_log = log10(timedrs+1),
         phyheal_log = log(phyheal+1),
         menheal_sqrt = sqrt(menheal +1),
         stress_sqrt = sqrt(stress +1))
```

```
## Warning: package 'bindrcpp' was built under R version 3.3.3
```

Q3

cause i'm lazy

```
lazy.list <- list(lab3$timedrs,lab3$phyheal,lab3$menheal,lab3$stress,lab3$timedrs_log,lab3$phyhe
al_log,lab3$menheal_sqrt,lab3$stress_sqrt)
lazy.names <- names(lab3[2:9])
zz<-walk2(lazy.list,lazy.names,skurt)
```

```
## [1] "timedrs"
## [[1]]
##   skew lwr.ci upr.ci
##   3.25   2.62   4.10
##
## [[2]]
##   kurt lwr.ci upr.ci
##  13.10   7.80  24.05
##
## [1] "phyheal"
## [[1]]
##   skew lwr.ci upr.ci
##   1.03   0.73   1.37
##
## [[2]]
##   kurt lwr.ci upr.ci
##   1.12   0.08   3.12
##
## [1] "menheal"
## [[1]]
##   skew lwr.ci upr.ci
##   0.60   0.42   0.81
##
## [[2]]
##   kurt lwr.ci upr.ci
##  -0.29  -0.69   0.25
##
## [1] "stress"
## [[1]]
##   skew lwr.ci upr.ci
##   1.04   0.69   1.75
##
## [[2]]
##   kurt lwr.ci upr.ci
##   1.80   0.11   5.57
##
## [1] "timedrs_log"
## [[1]]
##   skew lwr.ci upr.ci
##   0.23   0.01   0.45
##
## [[2]]
##   kurt lwr.ci upr.ci
##  -0.18  -0.52   0.22
##
## [1] "phyheal_log"
## [[1]]
##   skew lwr.ci upr.ci
##   0.16  -0.02   0.39
##
## [[2]]
##   kurt lwr.ci upr.ci
##  -0.64  -0.87  -0.27
```

```
##
## [1] "menheal_sqrt"
## [[1]]
##   skew lwr.ci upr.ci
##  -0.02  -0.20   0.13
##
## [[2]]
##   kurt lwr.ci upr.ci
##  -0.67  -0.89  -0.40
##
## [1] "stress_sqrt"
## [[1]]
##   skew lwr.ci upr.ci
##  -0.04  -0.29   0.31
##
## [[2]]
##   kurt lwr.ci upr.ci
##   0.00  -0.39   0.76
```

Table.1 Skewness, kurtosis and confidence intervals

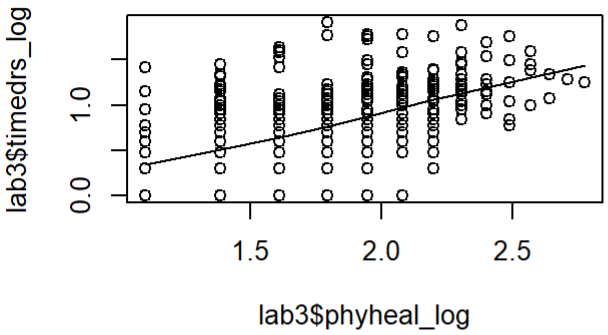
Variable(transformation)	Skewness[confidence interval]	Kurtosis[confidence interval]
Timedrs	3.25[2.64,4.03]	13.10[8.16,21.32]
Timedrs(log)*	0.23[0.00,.44]	-0.18[-0.46,0.29]
phyheal	1.03[0.76,1.43]	1.12[0.10,2.69]
phyheal_log*	0.16[-0.02,0.35]	-0.64[-0.89,-0.31]
menheal	0.60[0.41,0.84]	-0.29[-0.69,0.35]
menheal_sqrt*	-0.02[-0.19,0.14]	-0.67[-0.89,-0.40]
stress	1.04[0.67,1.69]	1.80[0.08,5.52]
stress_sqrt*	-0.04[-0.27,0.30]	0.00[-0.41,0.89]

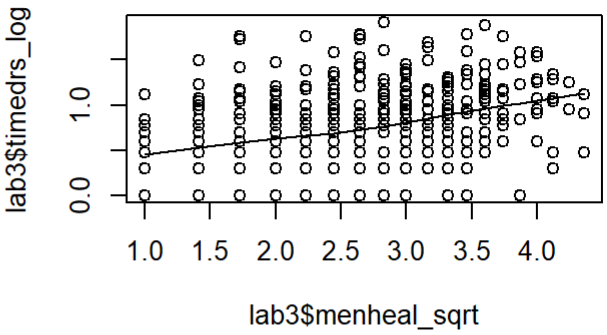
*fixed problems with skew.

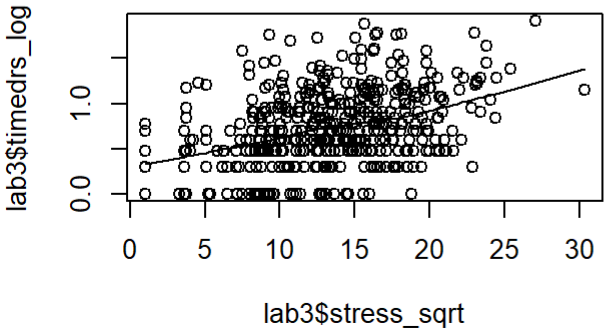
Q4

plots

```
par(mfrow = c(2,2))
plot(lab3$phyheal_log, lab3$timedrs_log)
lines(lowess(lab3$phyheal_log, lab3$timedrs_log))
plot(lab3$menheal_sqrt, lab3$timedrs_log)
lines(lowess(lab3$menheal_sqrt, lab3$timedrs_log))
plot(lab3$stress_sqrt, lab3$timedrs_log)
lines(lowess(lab3$stress_sqrt, lab3$timedrs_log))
```







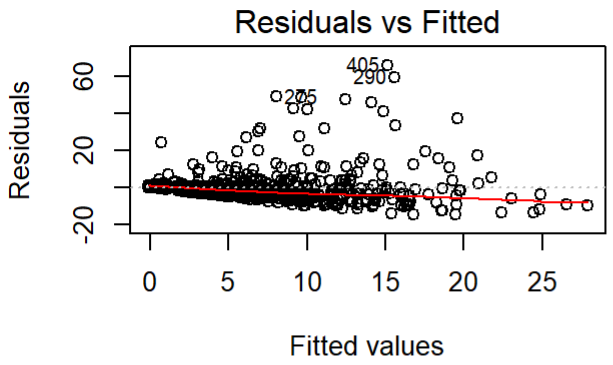
Q5

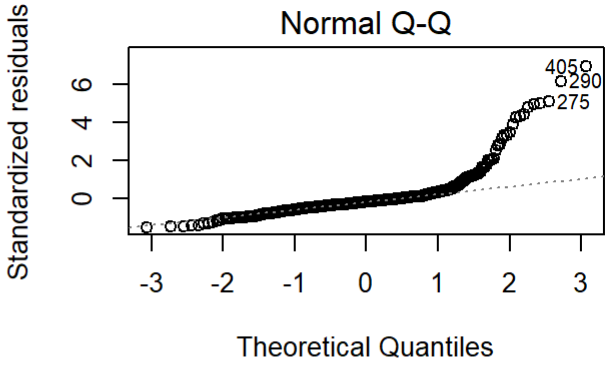
untransformed

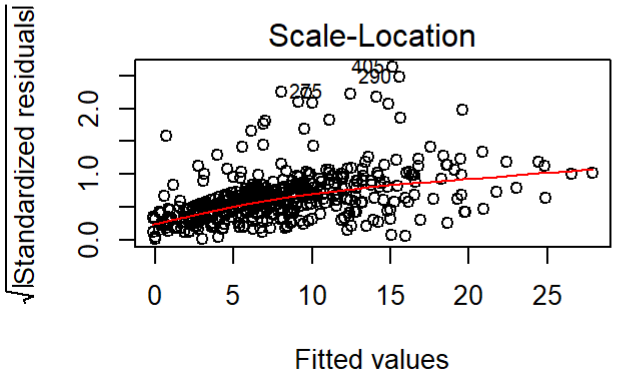
```
heal.mod <- lm(timedrs ~ phyheal + menheal + stress, data = lab3)
summary(heal.mod)
```

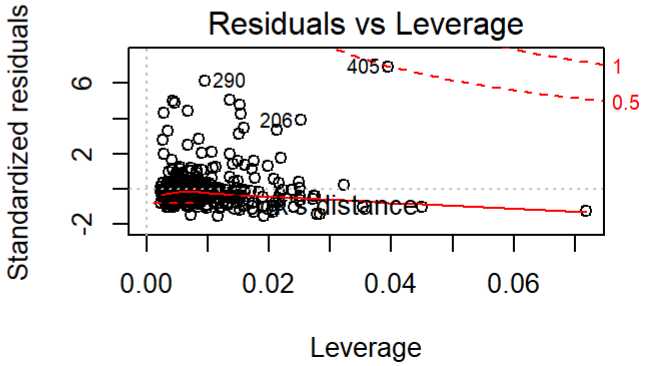
```
##
## Call:
## lm(formula = timedrs ~ phyheal + menheal + stress, data = lab3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -14.792  -4.353  -1.815   0.902  65.886
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3.704848   1.124195  -3.296 0.001058 **
## phyheal      1.786948   0.221074   8.083  5.6e-15 ***
## menheal     -0.009666   0.129029  -0.075 0.940318
## stress       0.013615   0.003612   3.769 0.000185 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.708 on 461 degrees of freedom
## Multiple R-squared:  0.2188, Adjusted R-squared:  0.2137
## F-statistic: 43.03 on 3 and 461 DF,  p-value: < 2.2e-16
```

```
par(mfrow = c(2,2))
plot(heal.mod)
```









```
vif(heal.mod)
```

```
##  phyheal  menheal   stress
## 1.372358 1.441328 1.184410
```

```
1/vif(heal.mod)
```

```
##   phyheal   menheal    stress
## 0.7286726 0.6938048 0.8443025
```

```
lmtest::bptest(heal.mod, varformula = ~ fitted.values(heal.mod), studentize = FALSE)
```

```
##
##  Breusch-Pagan test
##
## data:  heal.mod
## BP = 148.83, df = 1, p-value < 2.2e-16
```

Q6

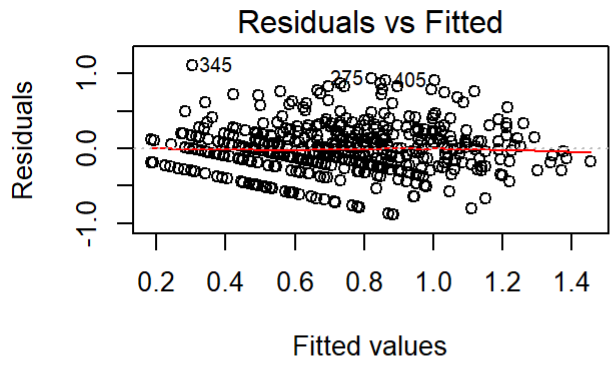
transformed

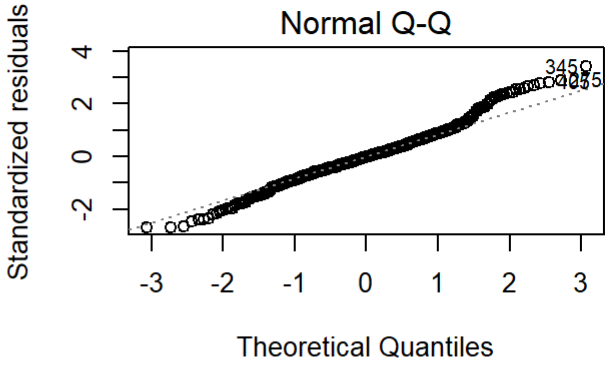
Transformed variables are better than the untransformed. Furthermore, the residuals and plots are more evenly distributed. The Breusch-Pagan is not significant which indicates that data is homoscedastic.

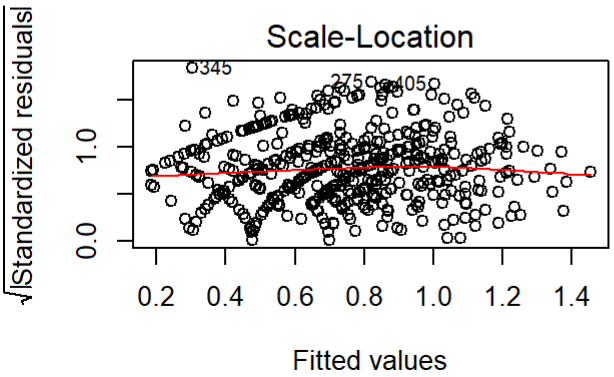
```
heal.mod.2 <- lm(timedrs_log ~ phyheal_log  + menheal_sqrt  + stress_sqrt, data = lab3)
summary(heal.mod.2)
```

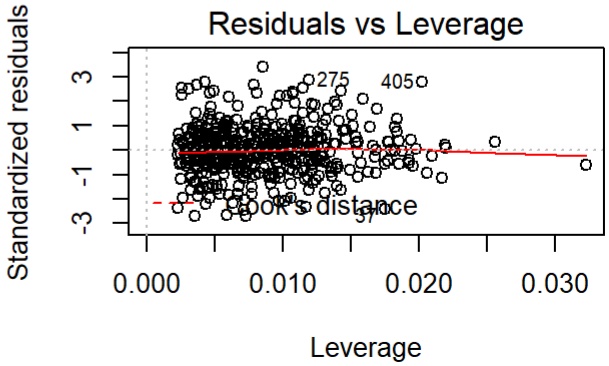
```
##
## Call:
## lm(formula = timedrs_log ~ phyheal_log + menheal_sqrt + stress_sqrt,
##     data = lab3)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.88556 -0.18896 -0.00823  0.18160  1.11316
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.457612   0.073422  -6.233 1.04e-09 ***
## phyheal_log   0.558314   0.046672  11.962  < 2e-16 ***
## menheal_sqrt  0.012376   0.022639   0.547    0.585
## stress_sqrt   0.015700   0.003397   4.621 4.96e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3283 on 461 degrees of freedom
## Multiple R-squared:  0.3791, Adjusted R-squared:  0.3751
## F-statistic: 93.84 on 3 and 461 DF,  p-value: < 2.2e-16
```

```
par(mfrow = c(2,2))
plot(heal.mod.2)
```









```
lmtest::bptest(heal.mod.2, varformula = ~ fitted.values(heal.mod.2), studentize = FALSE)
```

```
##
##  Breusch-Pagan test
##
## data:  heal.mod.2
## BP = 0.88699, df = 1, p-value = 0.3463
```

Q7

Multicolinearity is also not a problem because the variance inflation factor and 1/VIF are low.

```
vif(heal.mod.2)
```

```
##  phyheal_log menheal_sqrt  stress_sqrt
##     1.384361     1.460800     1.201470
```

```
1/vif(heal.mod.2)
```

```
##  phyheal_log menheal_sqrt  stress_sqrt
##    0.7223551    0.6845564    0.8323140
```

Q8

Multivariate outliers are not a problem because *p* was not <.001.

```
n <- 465
hat <- hatvalues(heal.mod.2)
mahun<-((n-1)*(hat))-1
tail(sort(mahun),10)
```

```
##       358        45       405       226       237       125       280
##  8.093537  8.196096  8.408870  8.570152  8.744034  9.072863  9.192450
##       446       159       403
##  9.215315 10.883877 13.990406
```

```
1-pchisq(13.99, df = 3)
```

```
## [1] 0.002918796
```