Statistical models : Homework 2

Priyanshi Shah and Sourabh Prakash

2023-01-27

Question 1

a. Perform some simulations to check that the least squares coefficients are indeed normally distributed under the standard assumptions.

Solution: We first installed the package 'car' for this solution and imported using the following code.

```
library(ggplot2)
library(car)
```

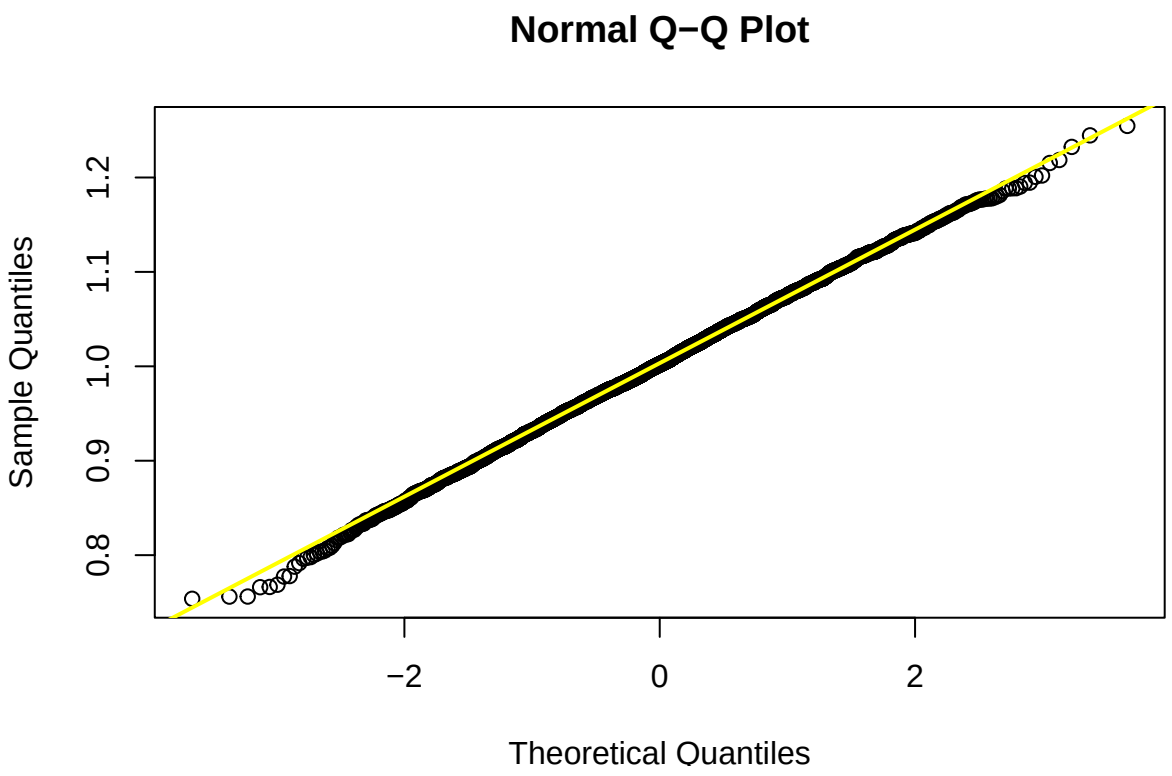
```
## Loading required package: carData
```

```
require(car)
```

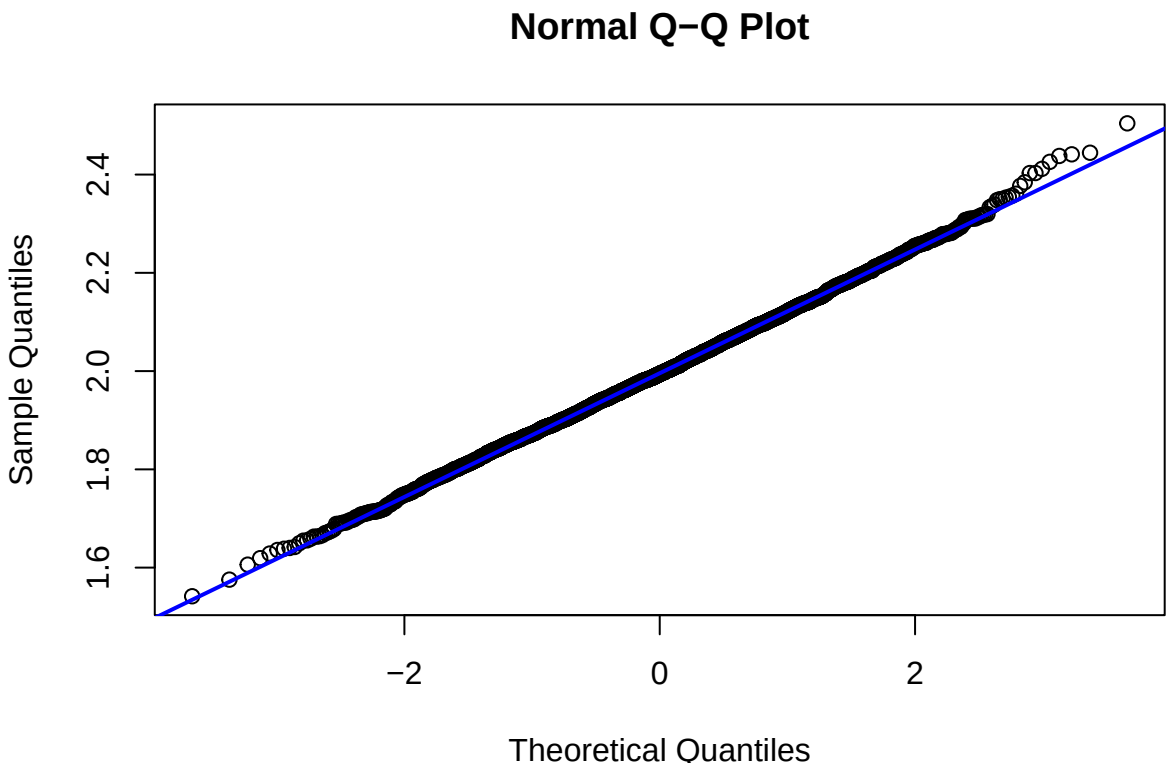
```
#making a function sample_size which returns slop and intercept list
sample_size<-function(n){
# x with a uniform distribution of [-1,1]
x= runif(n,-1,1)
e=rnorm(n,0,0.5)
#Computing y
y=1+2*x+e
linear_model=lm(y~x)
#print(summary(model_fit))
values<- coef(linear_model)
# getting the slope and intercept values from the linear model
Intercept <-values[1]
Slope <-values[2]
#putting them into variables
  value_list<-list("int"=Intercept,"slope"=Slope)
  return(value_list)
}
iter<-list(50,100,200,500)
final_list<-list()
final_list_int<-list()
final_list_slope<-list()
for(i in 1:1000){
for (n in iter){
      model_n=sample_size(50)
    final_list_int<-append(final_list_int,model_n$int)
```

```
    final_list_slope<-append(final_list_slope,model_n$slope)
}
  }
```

```
# Plot Normal Q_Q Plot of intercepts
qqnorm(unlist(final_list_int))
qqline(unlist(final_list_int),col="yellow",lwd=2)
```




```
#Plot Normal Q_Q Plot of slopes
qqnorm(unlist(final_list_slope))
qqline(unlist(final_list_slope),col="blue",lwd=2)
```

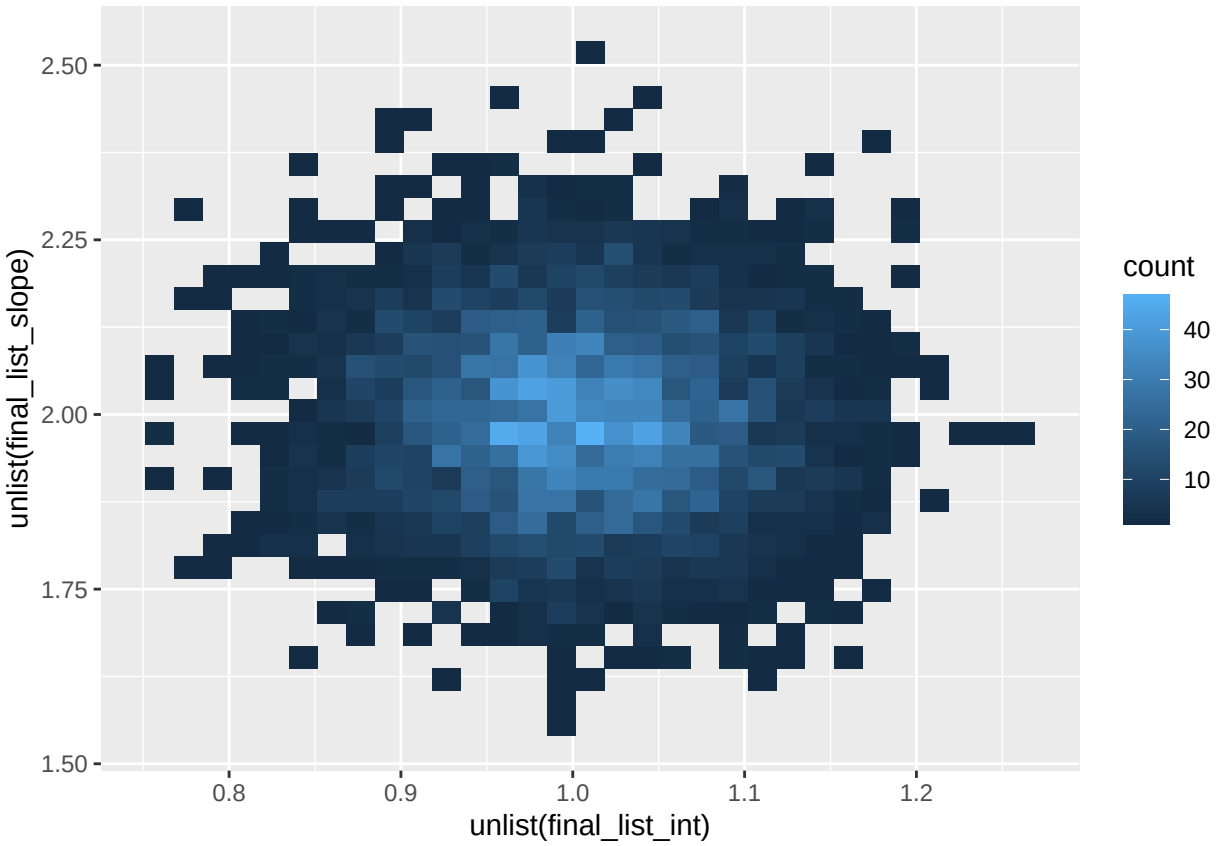


Joint Normal Plots

Now plotting joint normal plot s for slope and intercepts

```
qplot(unlist(final_list_int),unlist(final_list_slope),geom='bin2d')
```

```
## Warning: `qplot()` was deprecated in ggplot2 3.4.0.
```



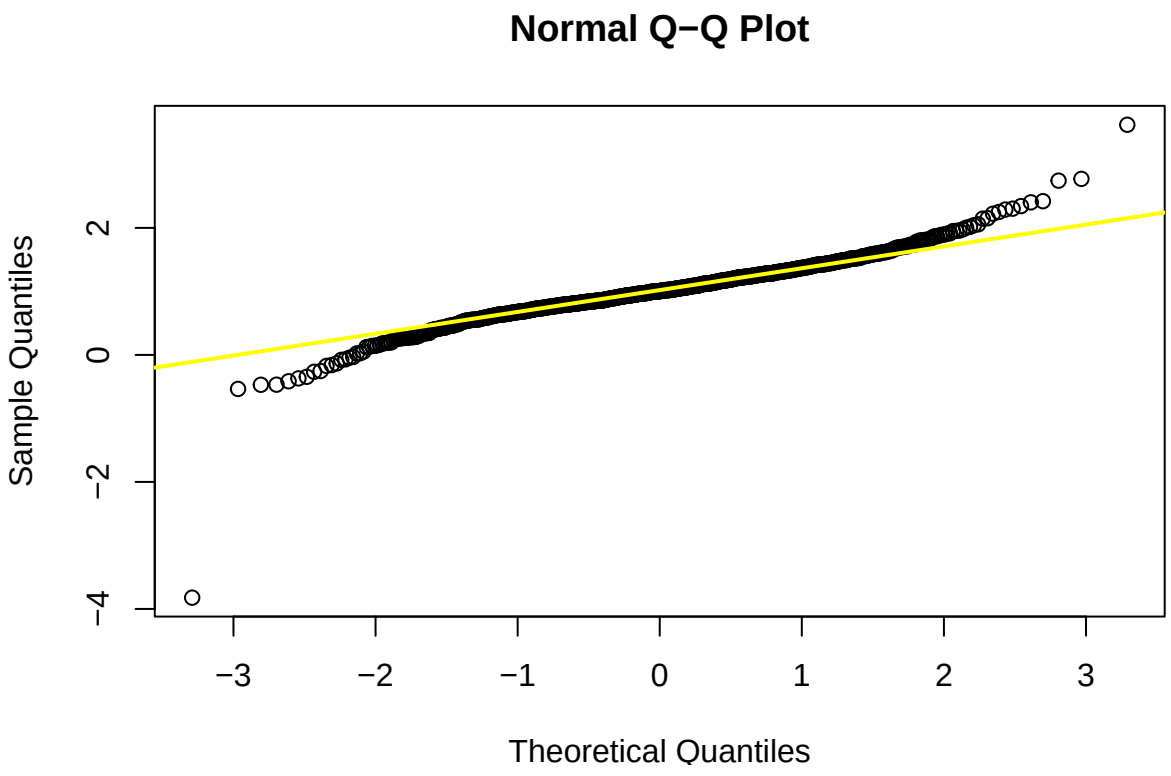
Inference: From the marginal plots we can clearly see almost all the points fitting almost linearly in the plot. The joint distribution plot clearly shows the dependence

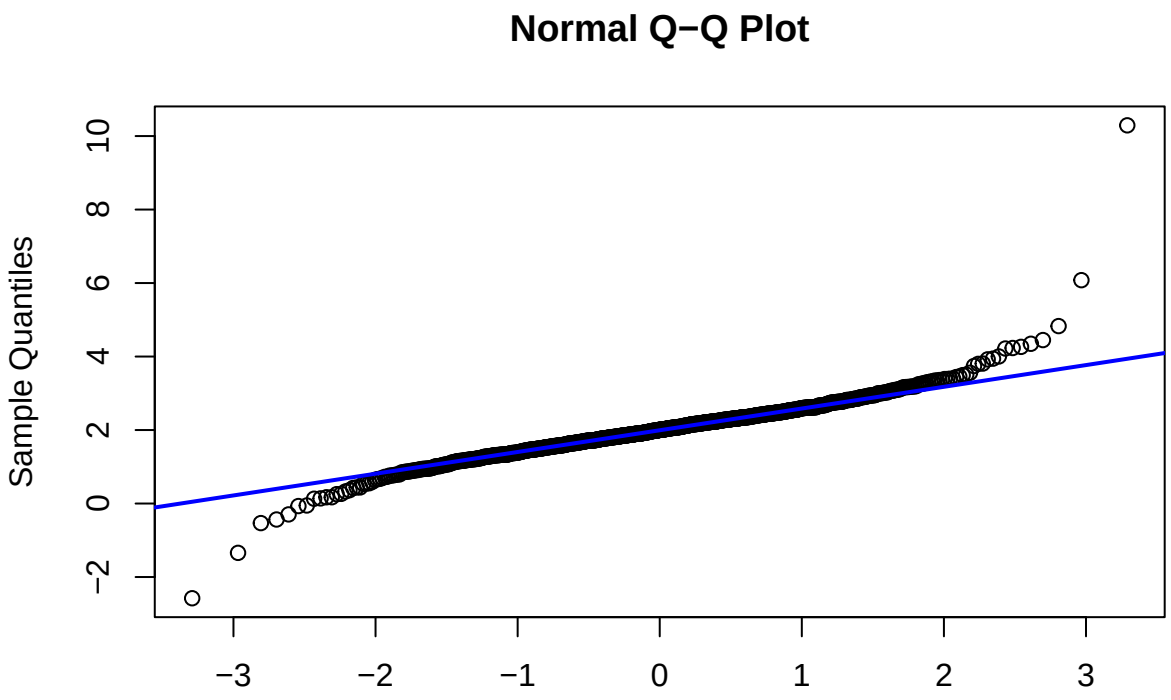
Question 1(b)

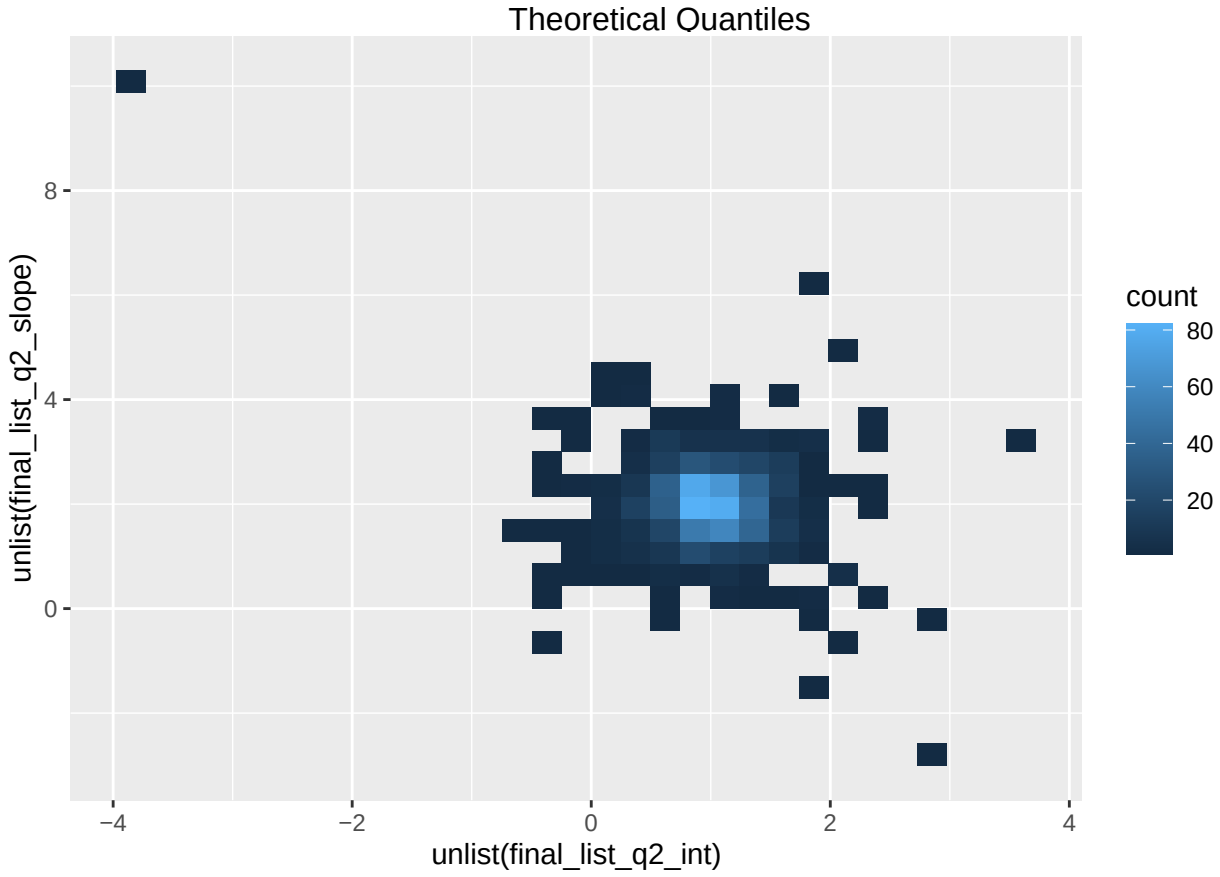
```
N<-list(50,100,200,500)
K<-list(2,5,10,20,50)
final_list_q2_int<-list()
final_list_q2_slope<-list()
n_sample_question2<-function(n,e){ x_= runif(n,-1,1)
y_=2*x_+1+e
  model_fit=lm(y_~x_)
  temp<- coef(model_fit)
  Intercept <-temp[1]
  Slope <-temp[2]
  list_temp<-list("int"=Intercept,"slope"=Slope)
  return(list_temp)
}
```

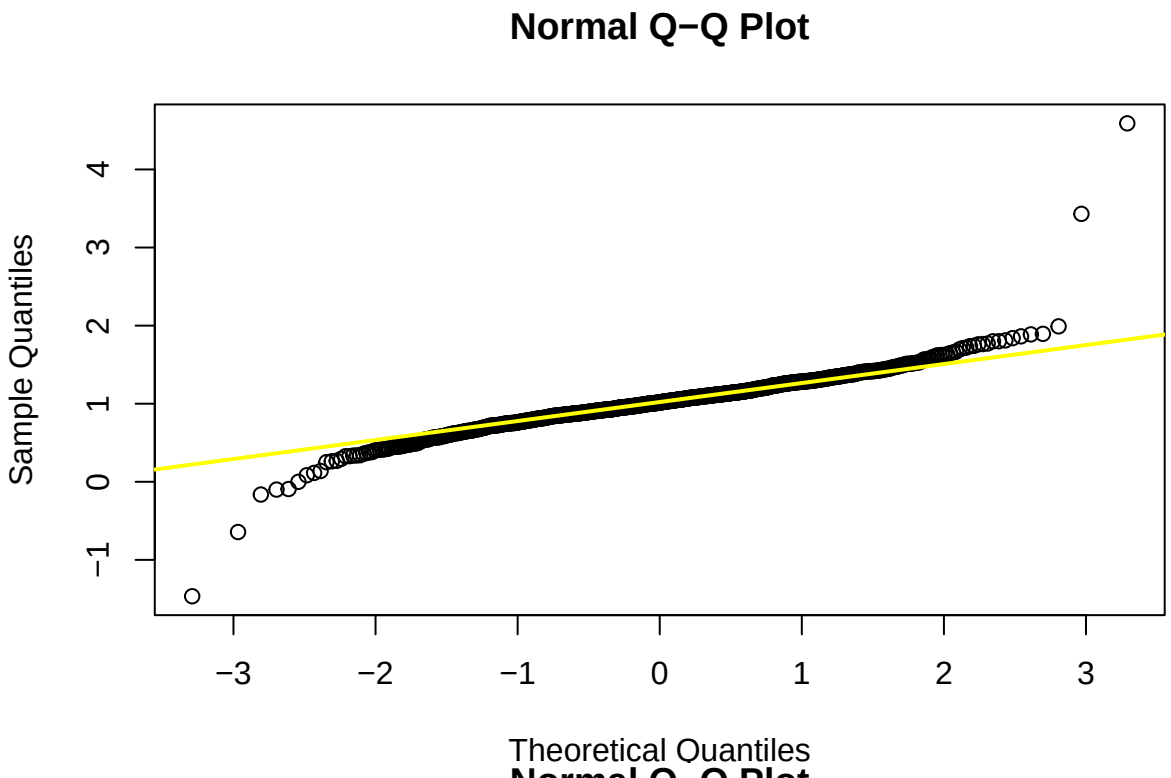
```
# Calculating for 20 different setting and
for (p in K){
for(k in N){
final_list_q2_int<-list()
final_list_q2_slope<-list()
for(j in 1:1000){
```

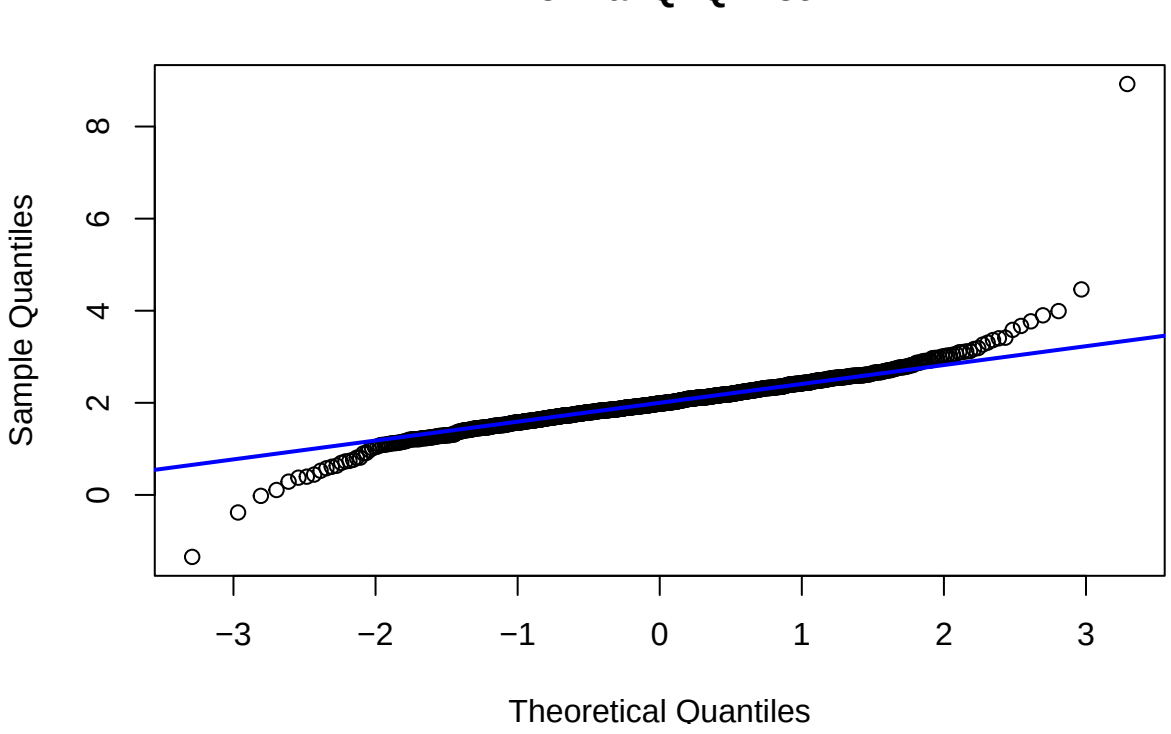
```
    e_dash=rt(k,p)
    list_question2=n_sample_question2(k,e_dash)
    final_list_q2_int<-append(final_list_q2_int,list_question2$int)
    final_list_q2_slope<-append(final_list_q2_slope,list_question2$slope)
}
#Marginal
qqnorm(unlist(final_list_q2_int))
qqline(unlist(final_list_q2_int), col="yellow", lwd=2)
qqnorm(unlist(final_list_q2_slope))
qqline(unlist(final_list_q2_slope), col="blue", lwd=2)

#Joint
print(qplot(unlist(final_list_q2_int),unlist(final_list_q2_slope), geom='bin2d'))

  }}
```

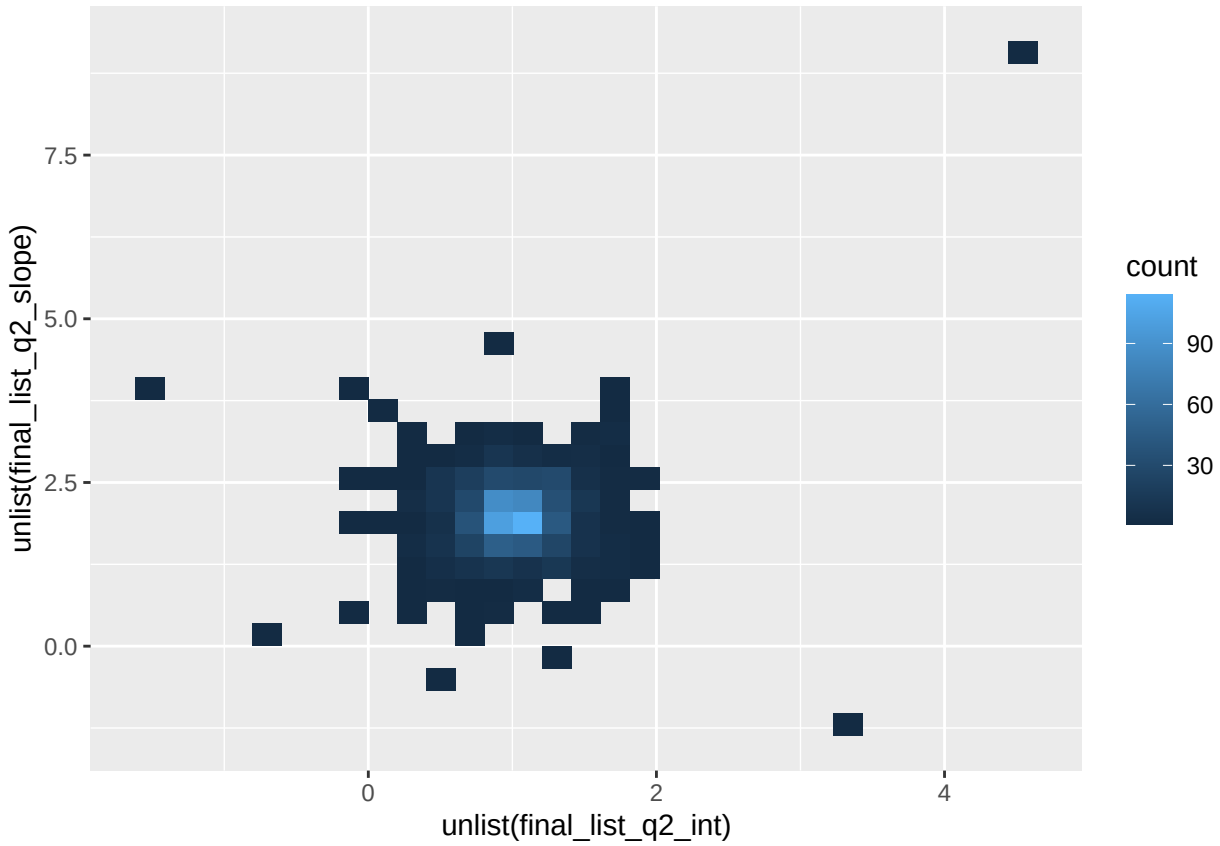


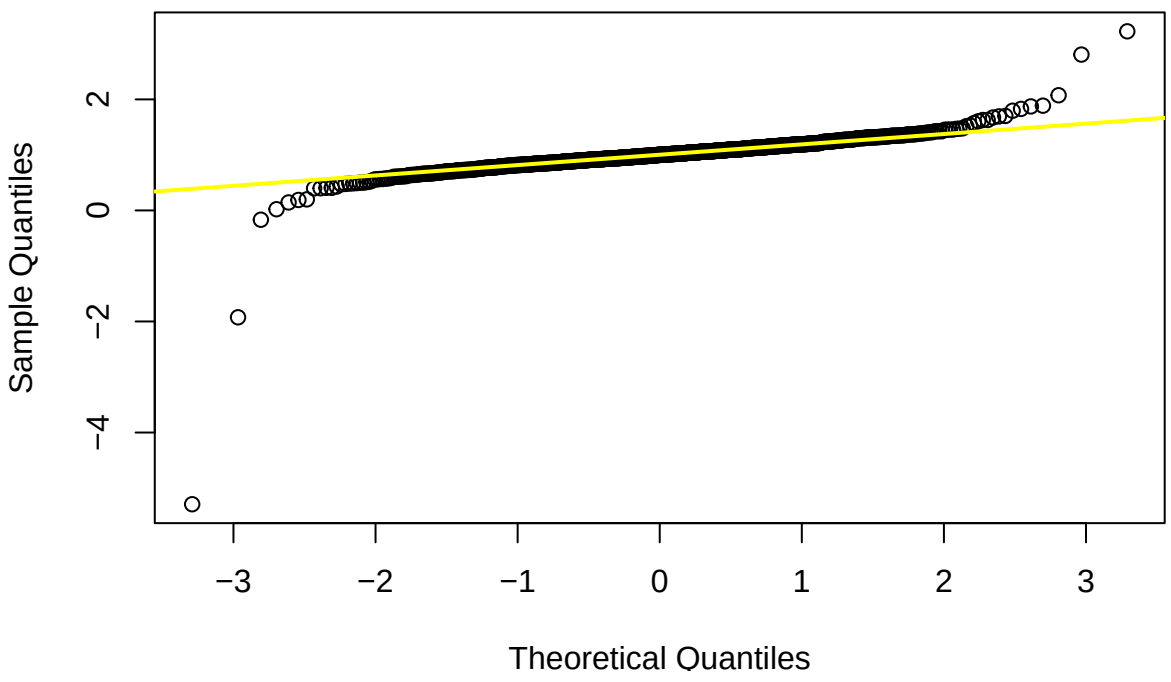

Normal Q−Q Plot
Sample Quantiles
Theoretical Quantiles
−2
0
2
4
6
8
10
−3
−2
−1
0
1
2
3

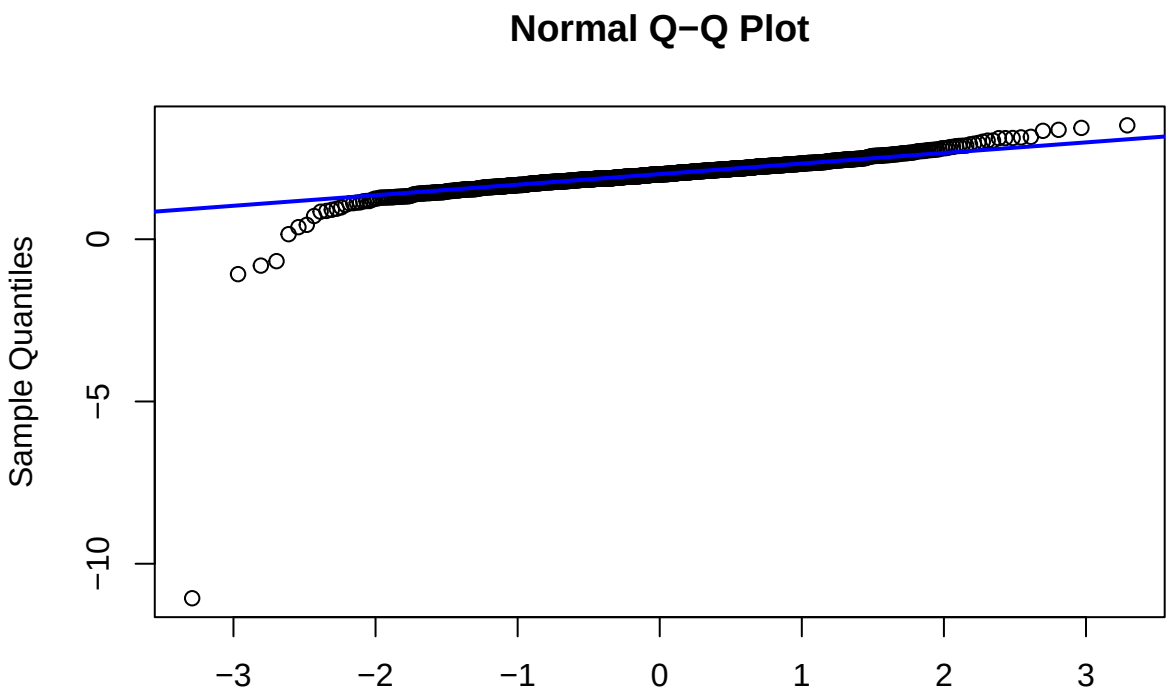

unlist(final_list_q2_slope)
unlist(final_list_q2_int)
count
80
60
40
20
8
4
0
−4
−2
0
2
4

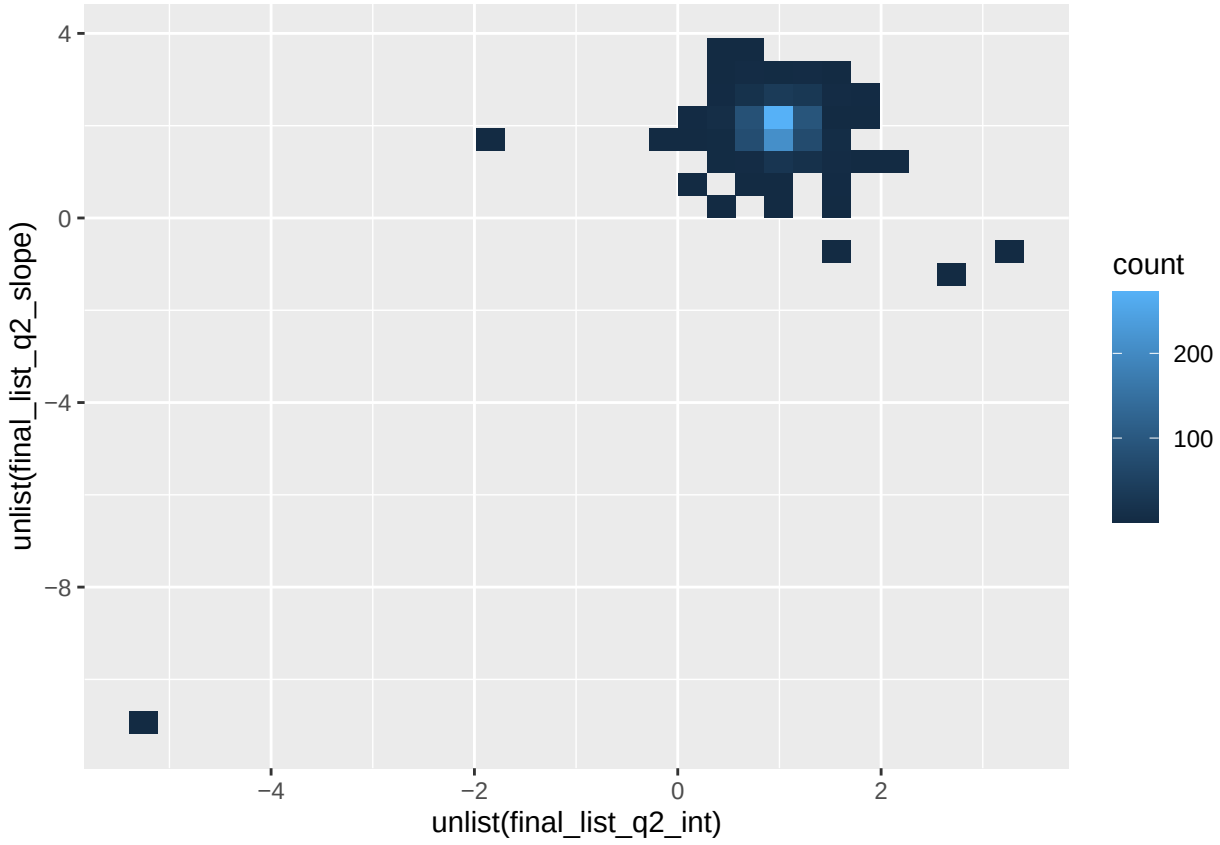

Normal Q−Q Plot
Sample Quantiles
Theoretical Quantiles
−3
−2
−1
0
1
2
3
−1
0
1
2
3
4

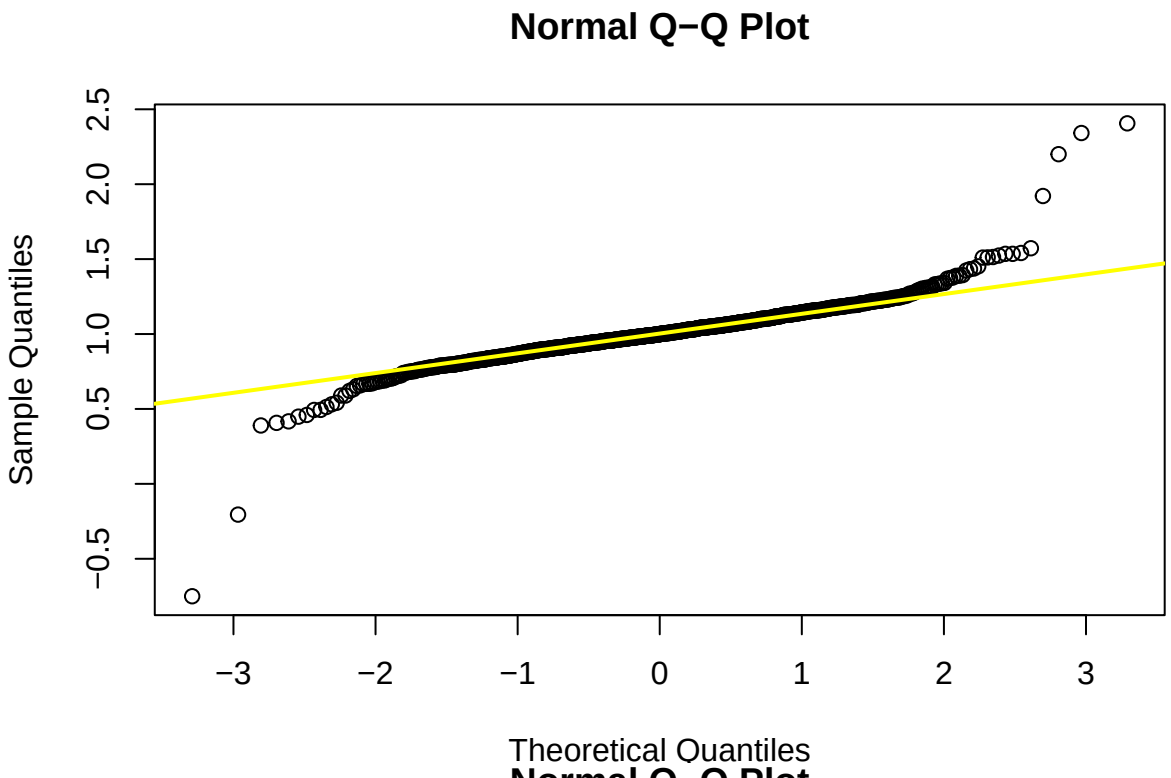

Normal Q−Q Plot
Sample Quantiles
Theoretical Quantiles
−3
−2
−1
0
1
2
3
0
2
4
6
8

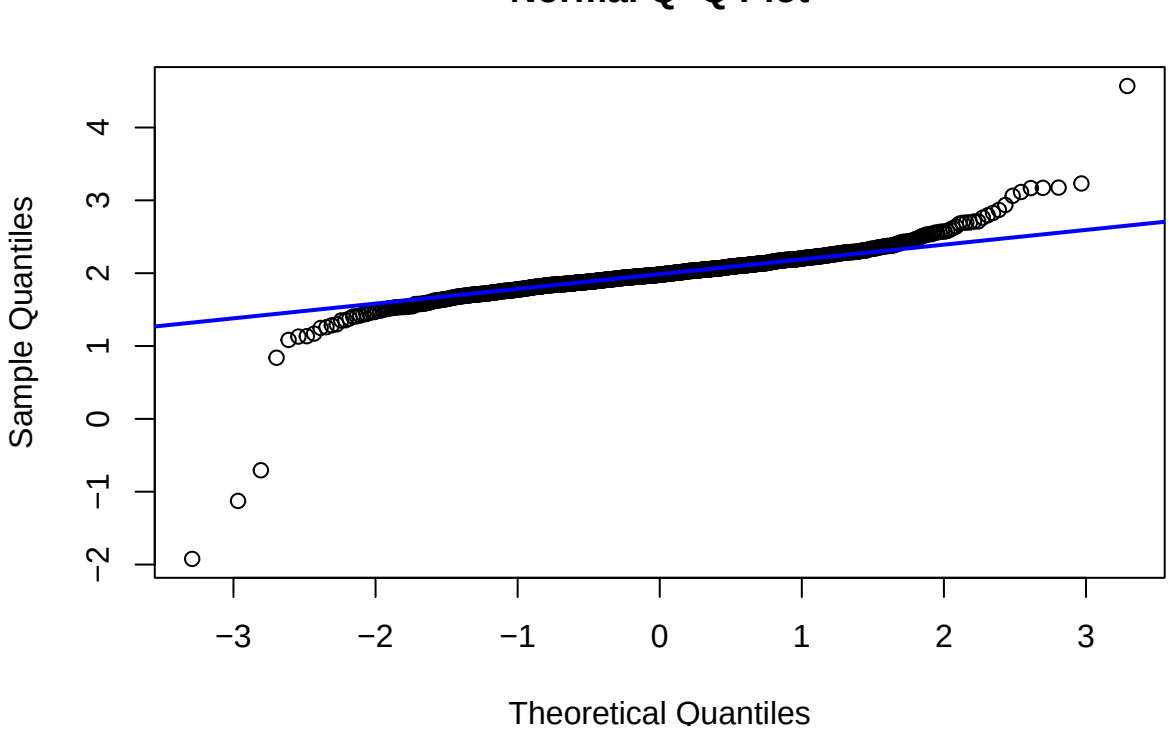

7.5
5.0
2.5
0.0
unlist(final_list_q2_slope)
0
2
4
unlist(final_list_q2_int)
count
90
60
30

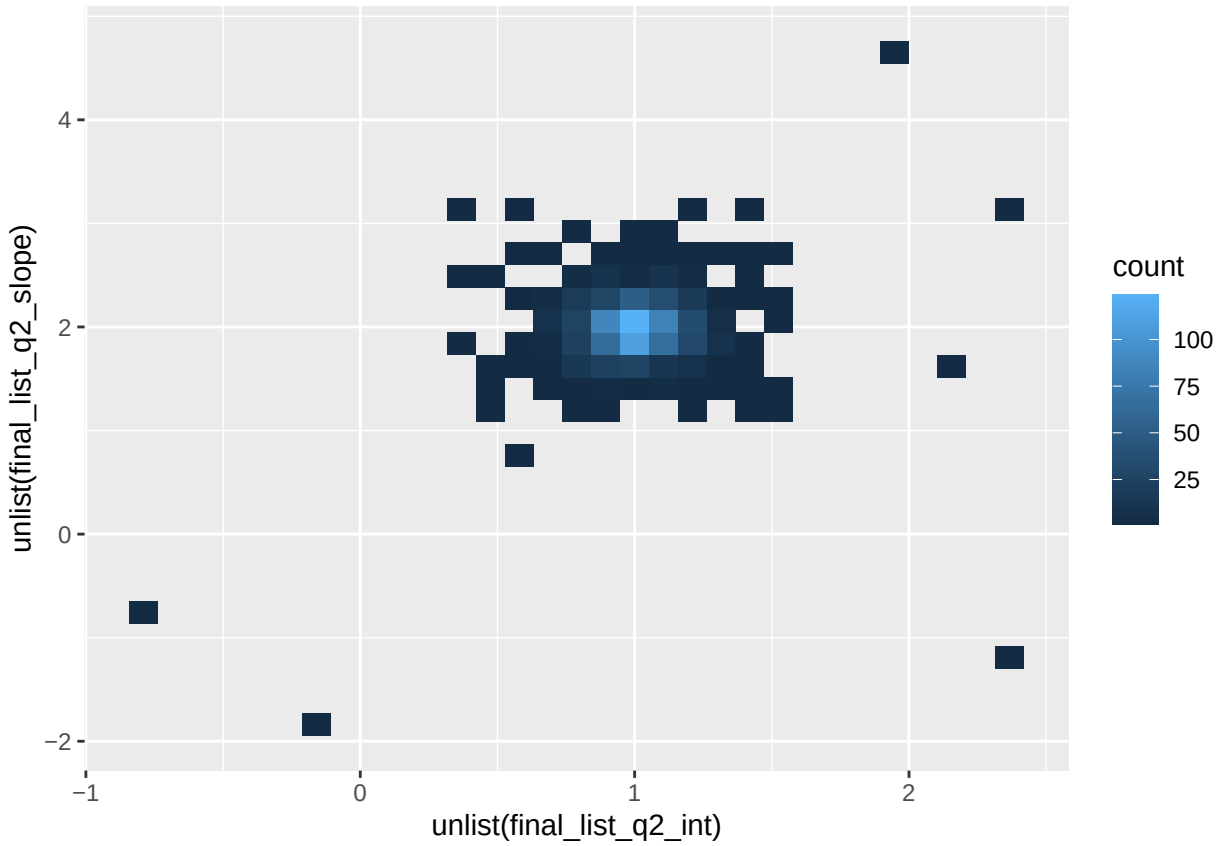

Normal Q−Q Plot
Sample Quantiles
2
0
−2
−4
−3
−2
−1
0
1
2
3
Theoretical Quantiles


Normal Q−Q Plot
Sample Quantiles
0
−5
−10
−3
−2
−1
0
1
2
3
Theoretical Quantiles

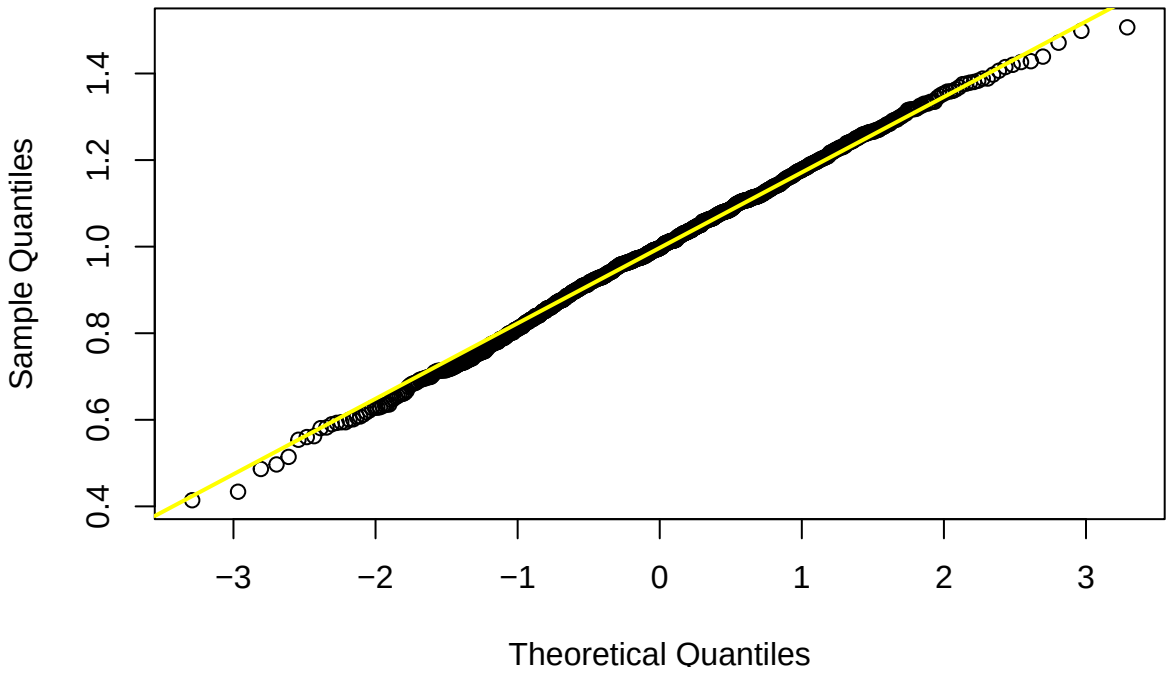

4
0
−4
−8
−4
−2
0
2
unlist(final_list_q2_slope)
unlist(final_list_q2_int)
count
200
100

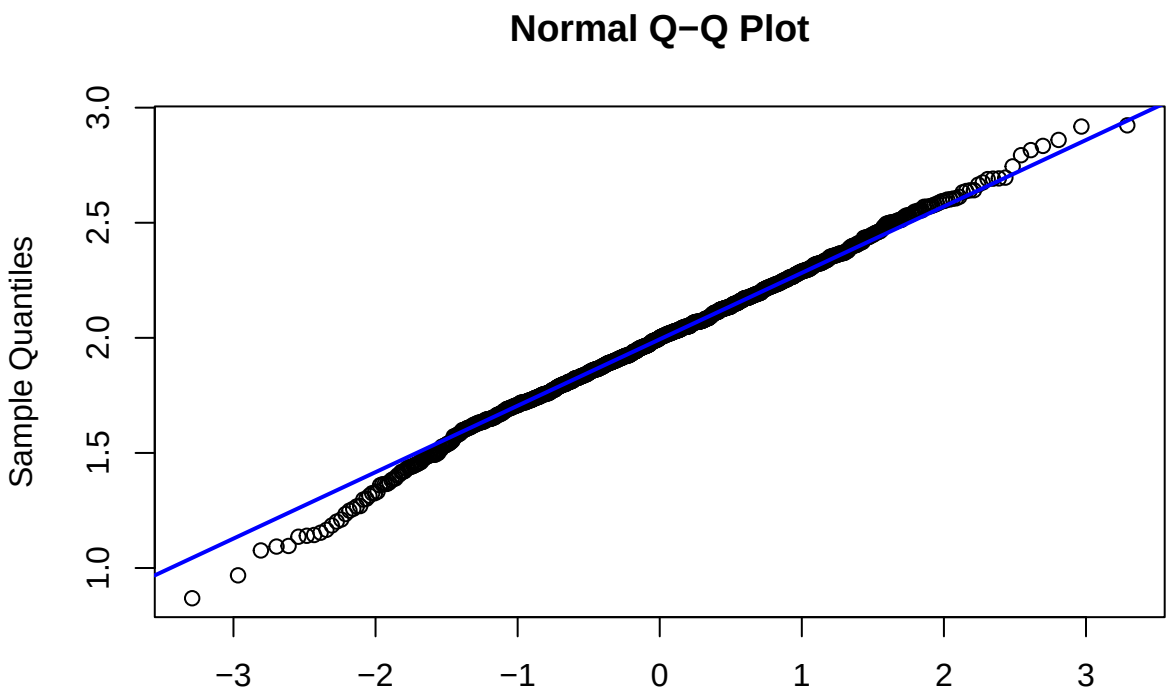

Normal Q−Q Plot
Sample Quantiles
Theoretical Quantiles
−3
−2
−1
0
1
2
3
−0.5
0.5
1.0
1.5
2.0
2.5

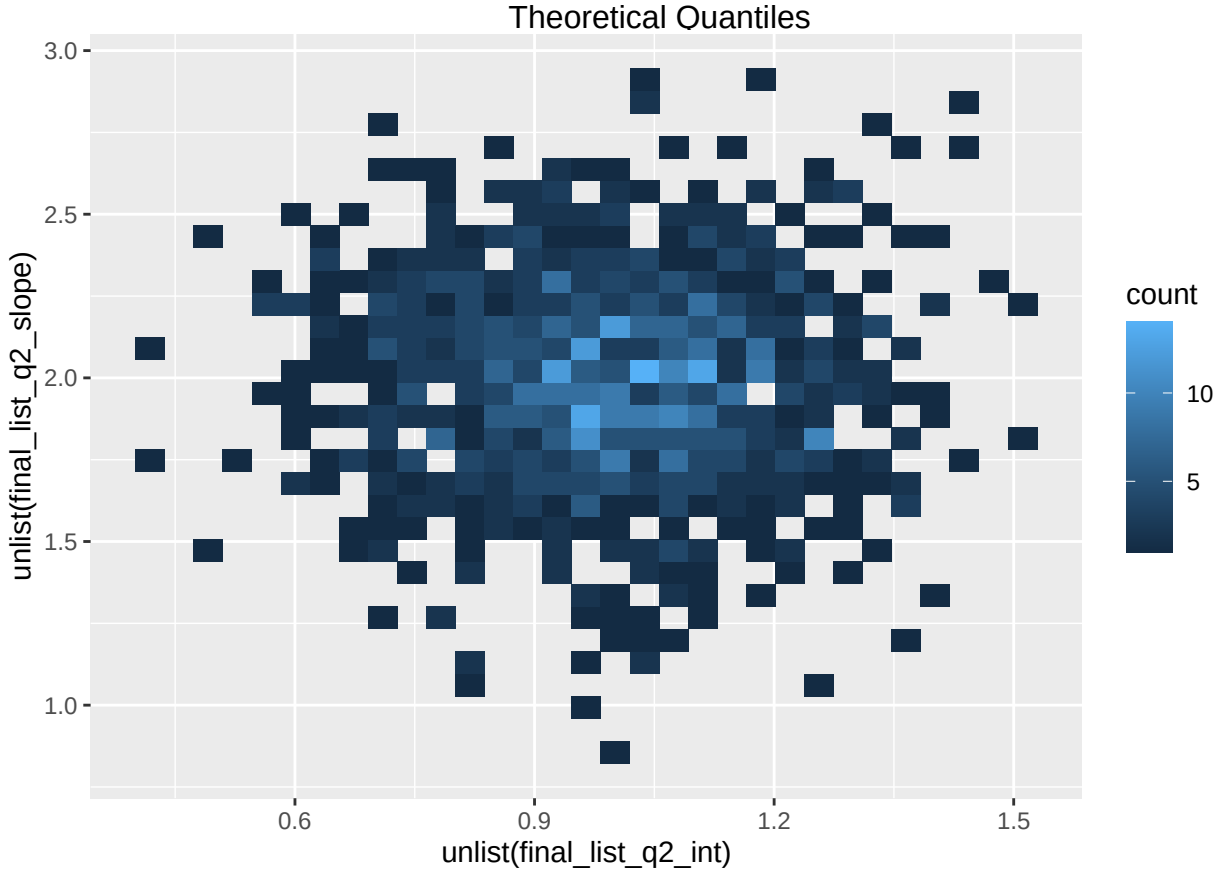

Normal Q−Q Plot
Sample Quantiles
Theoretical Quantiles
−3
−2
−1
0
1
2
3
−2
−1
0
1
2
3
4

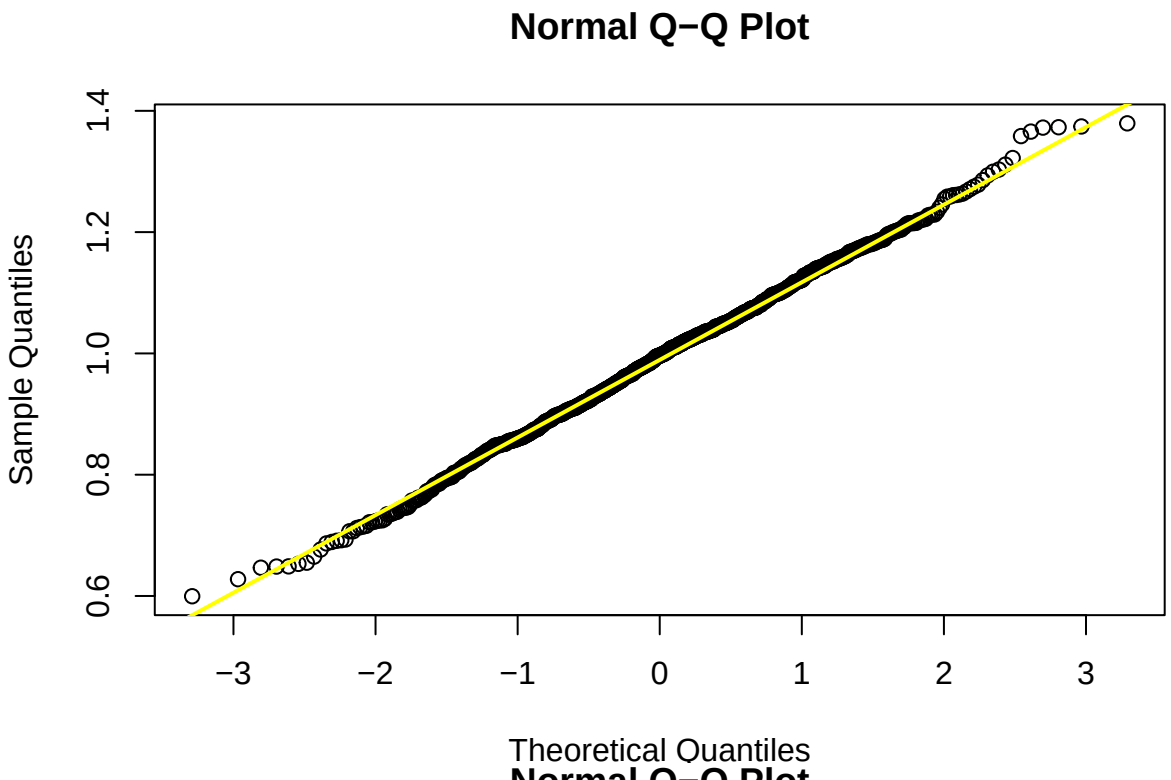


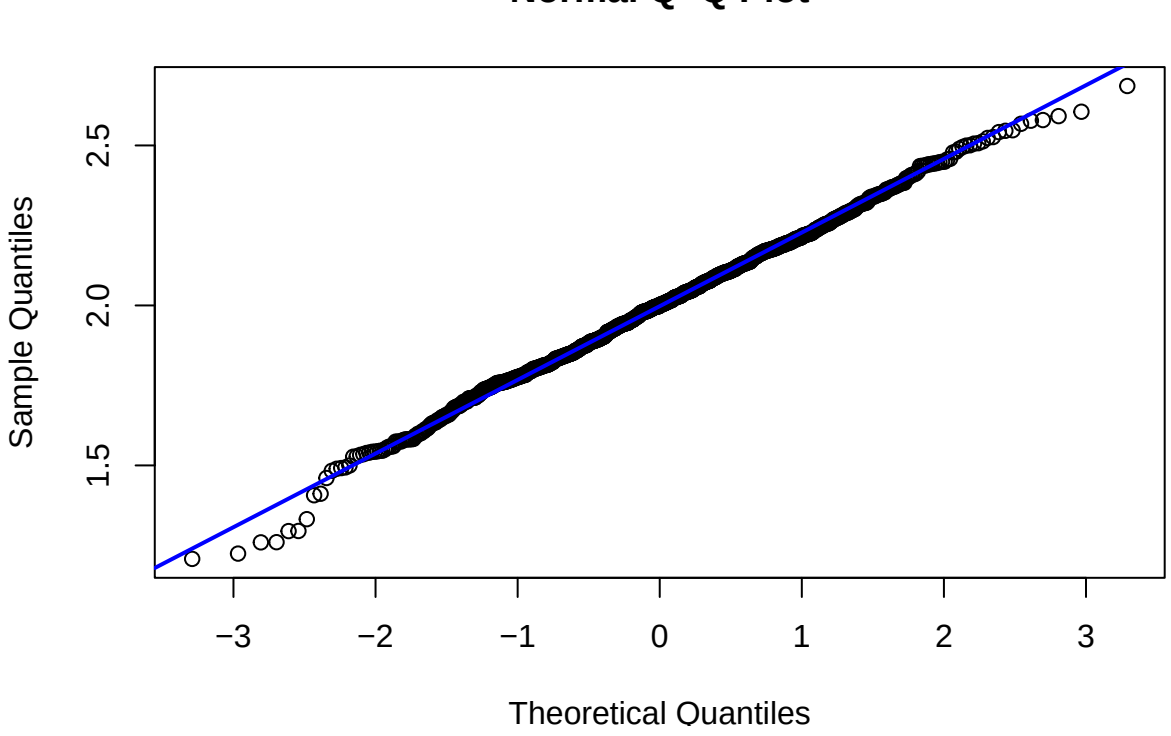
Normal Q−Q Plot

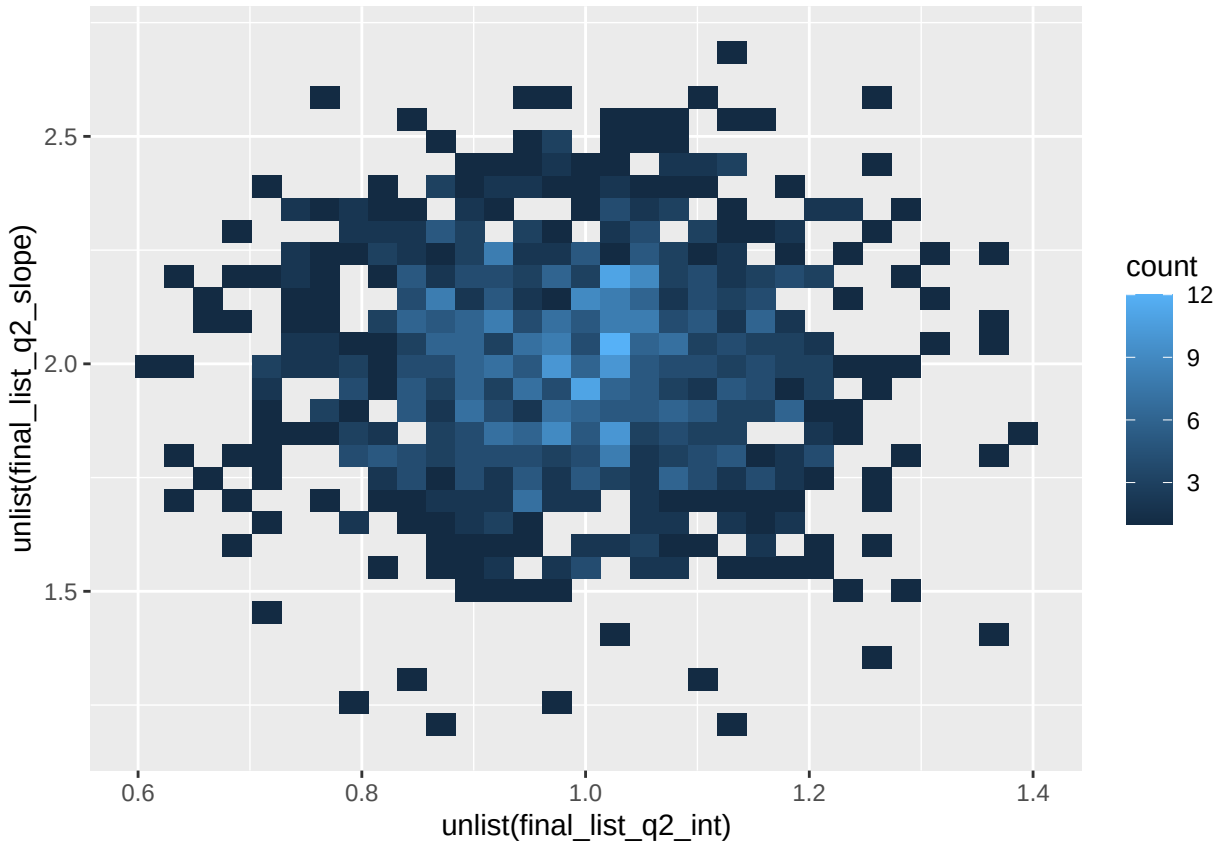


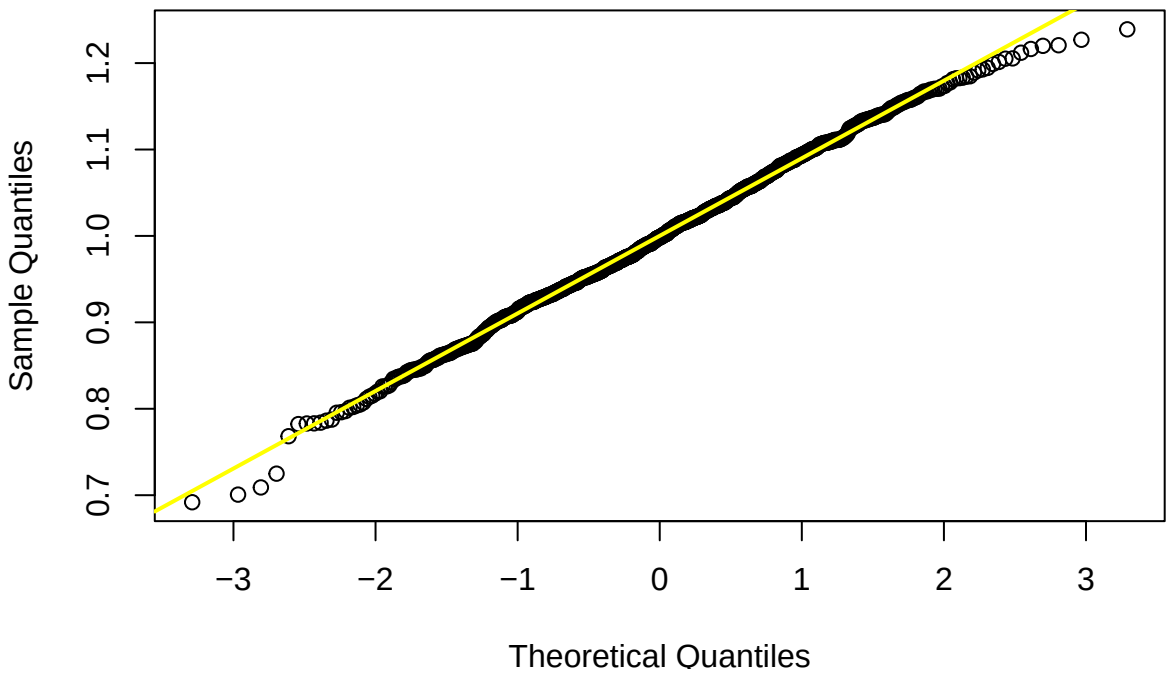

Normal Q−Q Plot
Sample Quantiles
Theoretical Quantiles
3.0
2.5
2.0
1.5
1.0
−3
−2
−1
0
1
2
3

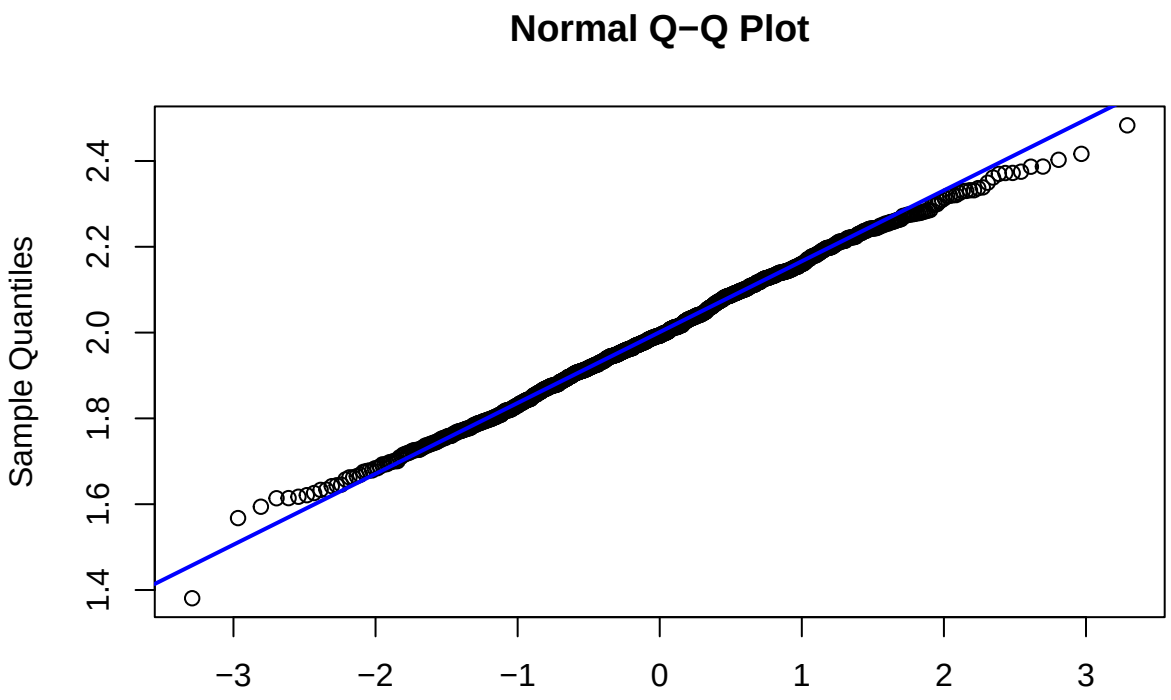

unlist(final_list_q2_slope)
unlist(final_list_q2_int)
count
10
5
3.0
2.5
2.0
1.5
1.0
0.6
0.9
1.2
1.5

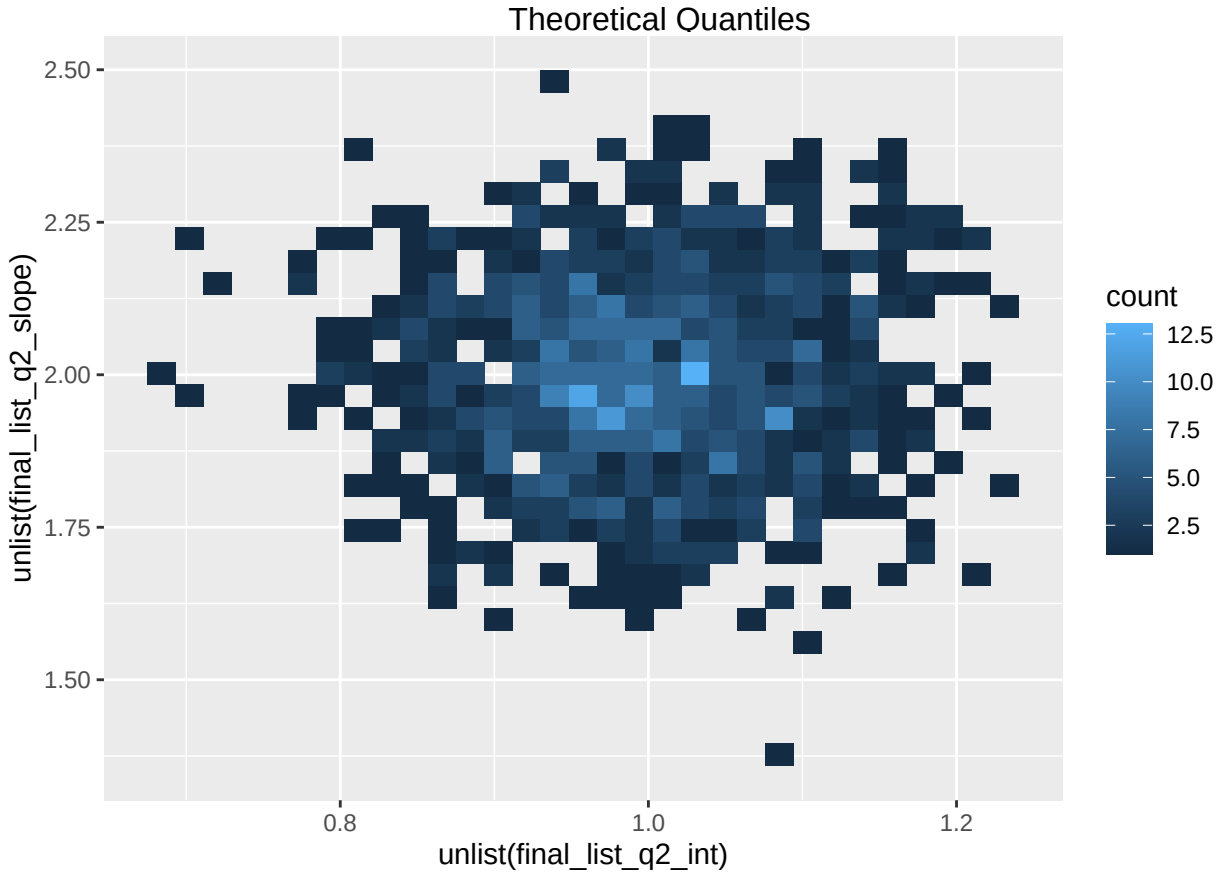

Normal Q−Q Plot
Sample Quantiles
0.6
0.8
1.0
1.2
1.4
−3
−2
−1
0
1
2
3
Theoretical Quantiles

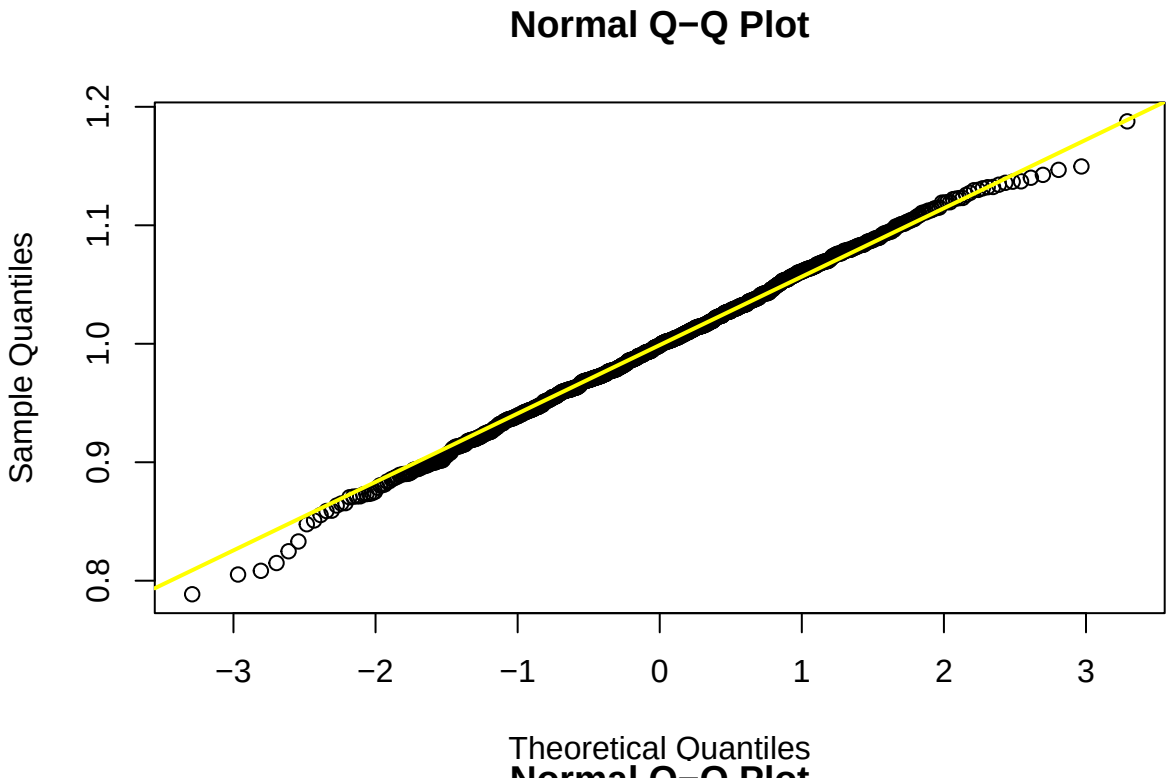

Normal Q−Q Plot
Sample Quantiles
1.5
2.0
2.5
−3
−2
−1
0
1
2
3
Theoretical Quantiles

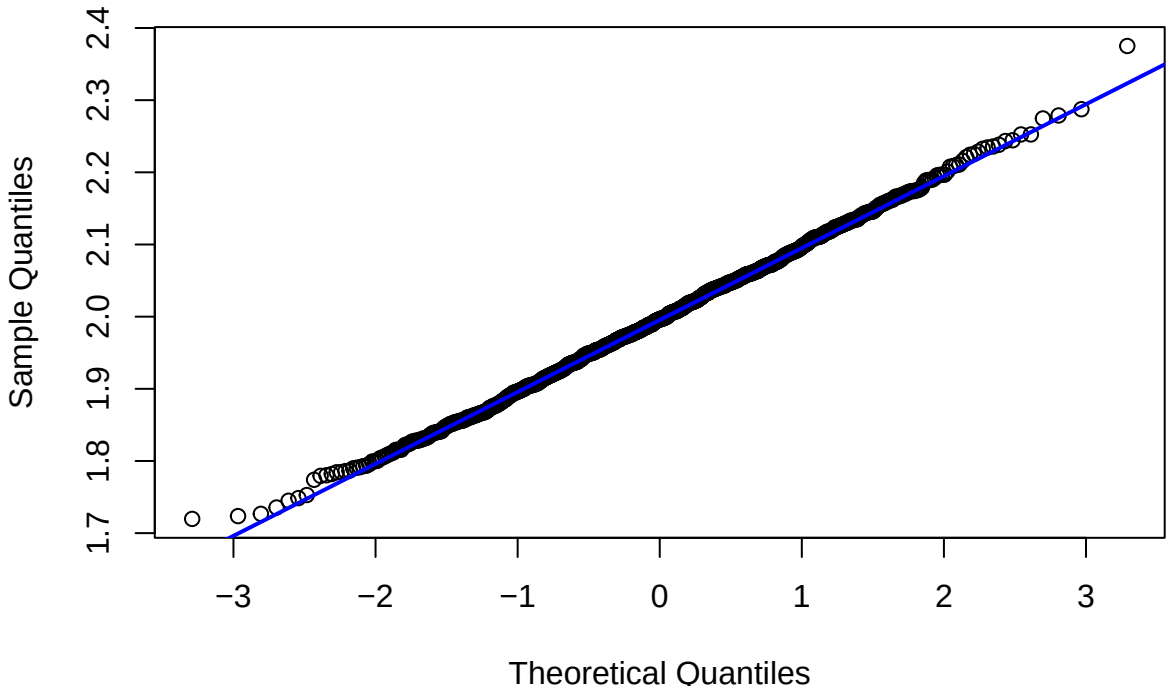

unlist(final_list_q2_slope)
2.5
2.0
1.5
0.6
0.8
1.0
1.2
1.4
unlist(final_list_q2_int)
count
12
9
6
3

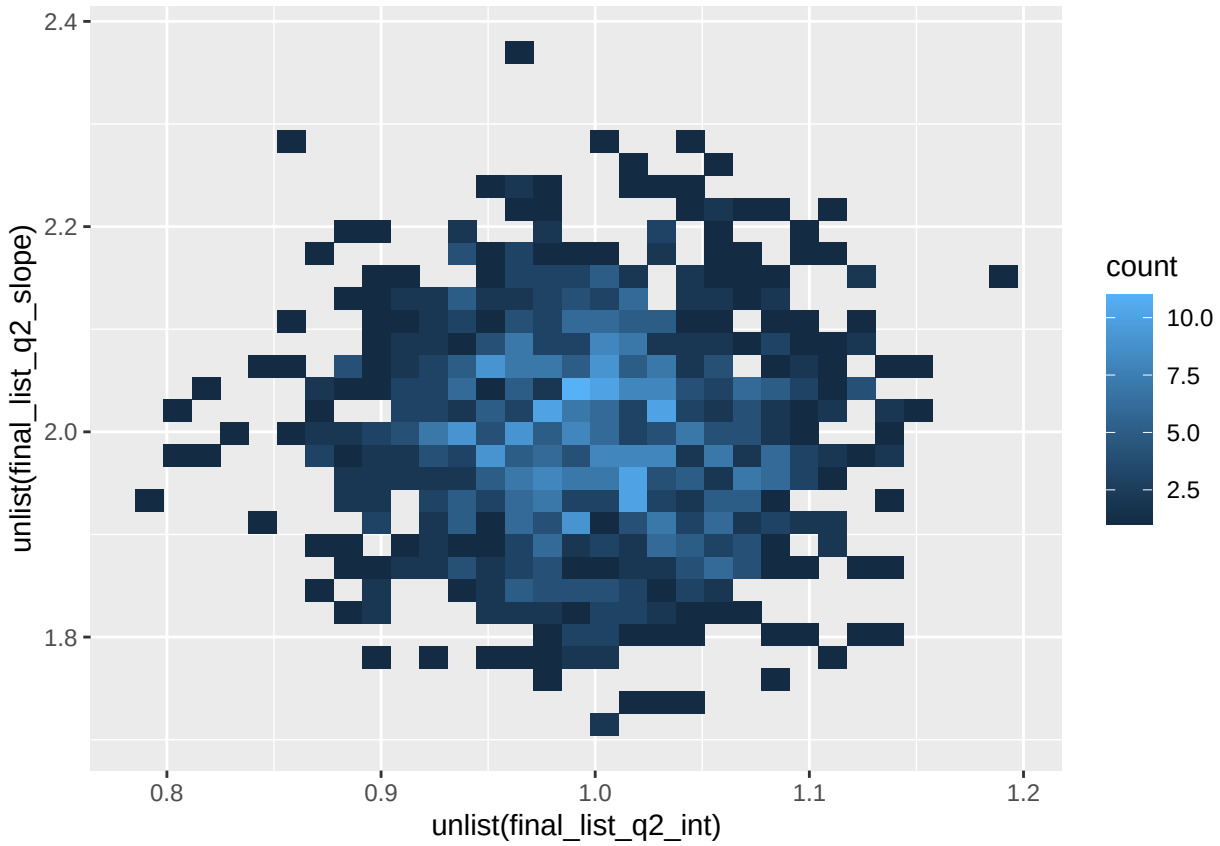

Normal Q−Q Plot
Sample Quantiles
0.7
0.8
0.9
1.0
1.1
1.2
−3
−2
−1
0
1
2
3
Theoretical Quantiles


Normal Q−Q Plot
Sample Quantiles
Theoretical Quantiles
1.4
1.6
1.8
2.0
2.2
2.4
−3
−2
−1
0
1
2
3

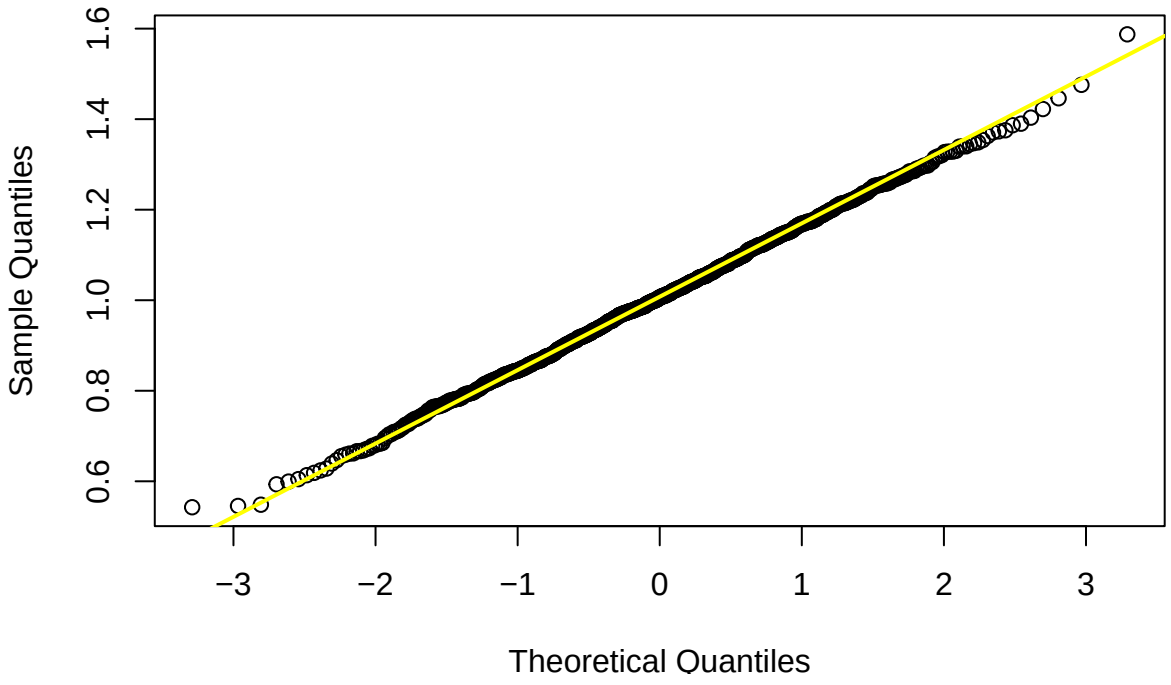

unlist(final_list_q2_slope)
unlist(final_list_q2_int)
count
2.50
2.25
2.00
1.75
1.50
0.8
1.0
1.2
12.5
10.0
7.5
5.0
2.5

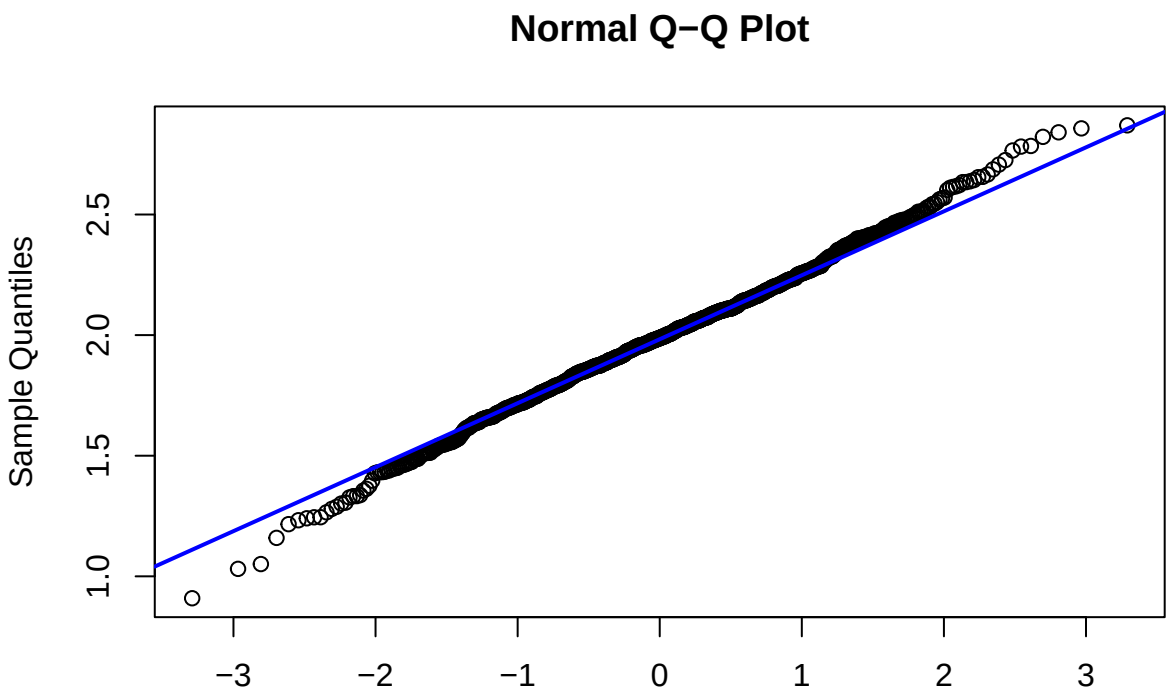

Normal Q−Q Plot
Sample Quantiles
Theoretical Quantiles
0.8
0.9
1.0
1.1
1.2
−3
−2
−1
0
1
2
3

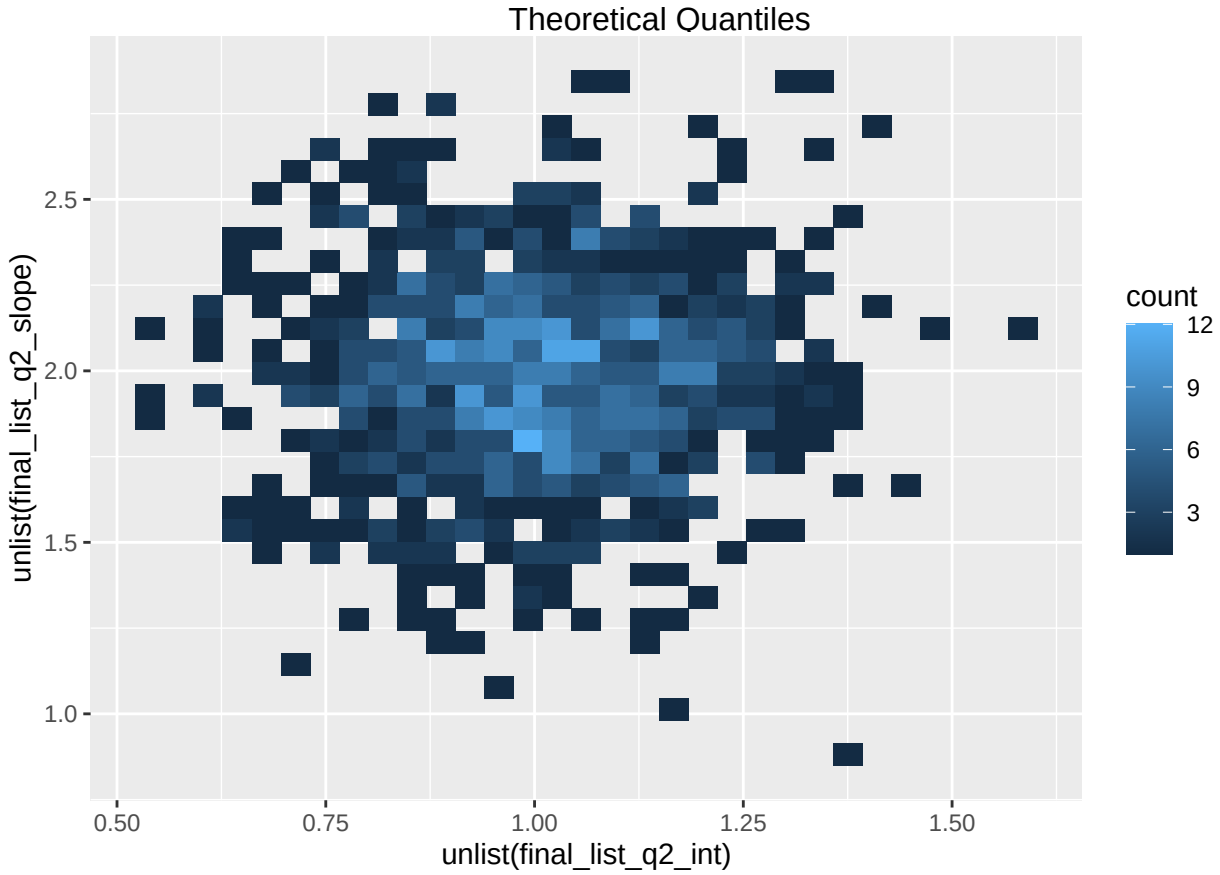

Normal Q−Q Plot
Sample Quantiles
Theoretical Quantiles
1.7
1.8
1.9
2.0
2.1
2.2
2.3
2.4
−3
−2
−1
0
1
2
3

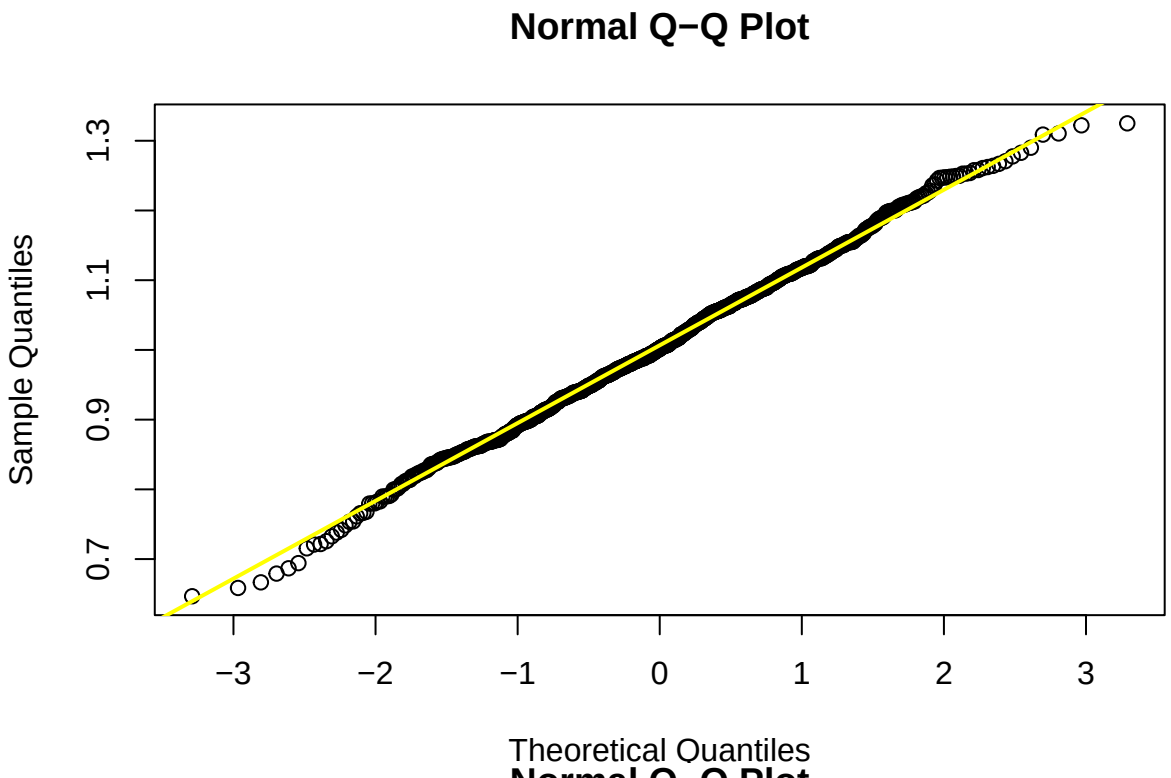


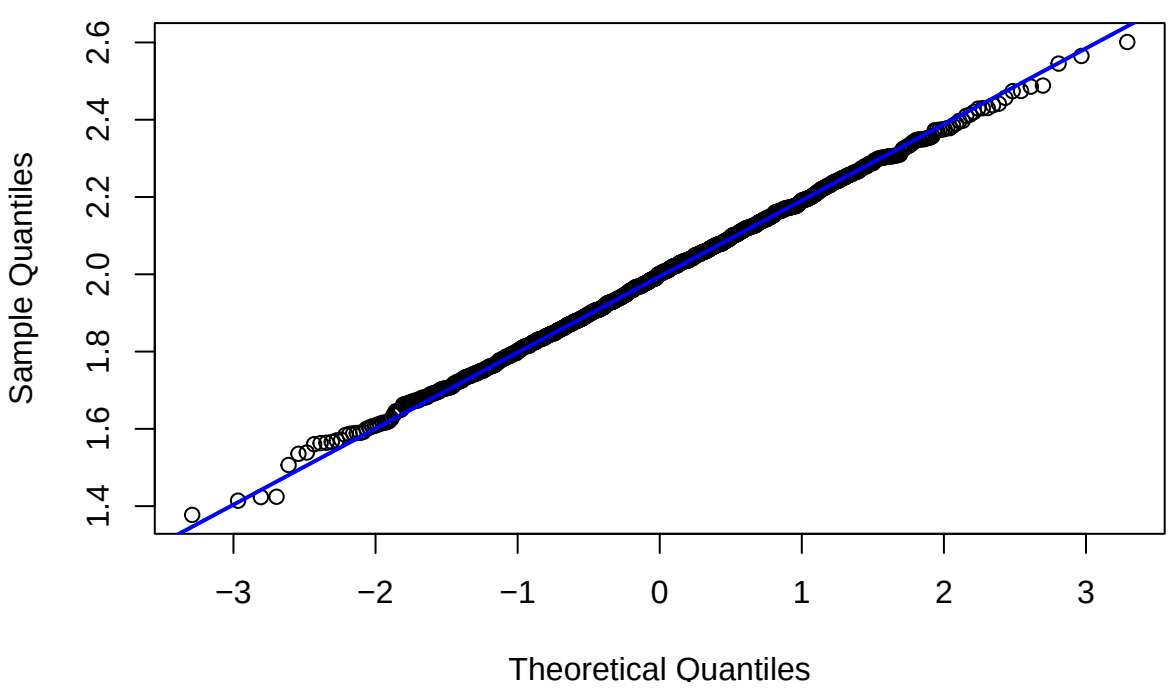
Normal Q−Q Plot

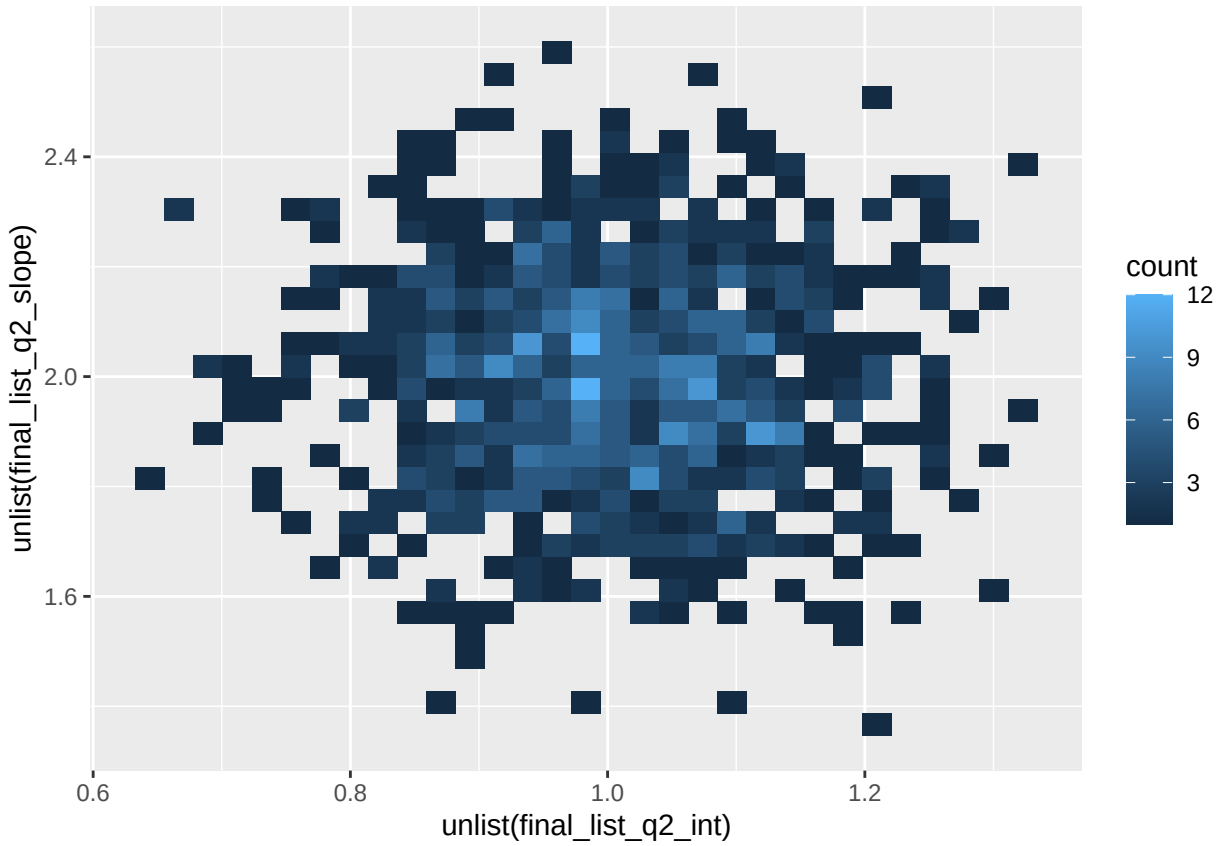


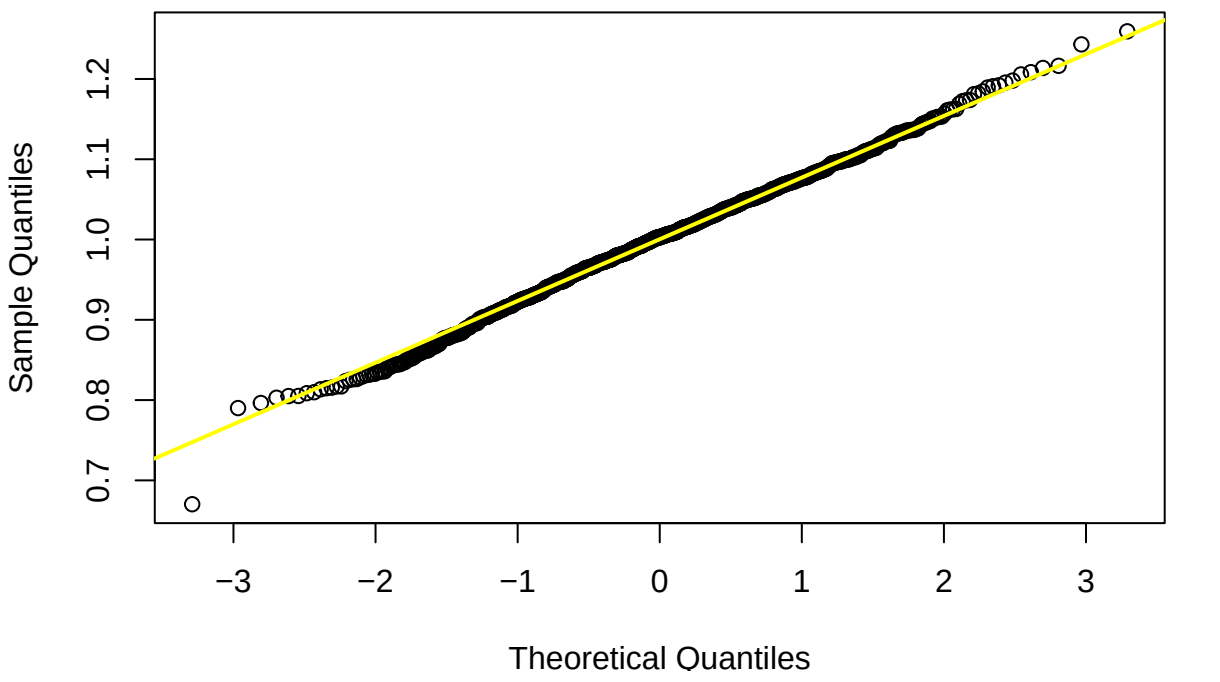

Normal Q−Q Plot
Sample Quantiles
Theoretical Quantiles
1.0
1.5
2.0
2.5
−3
−2
−1
0
1
2
3

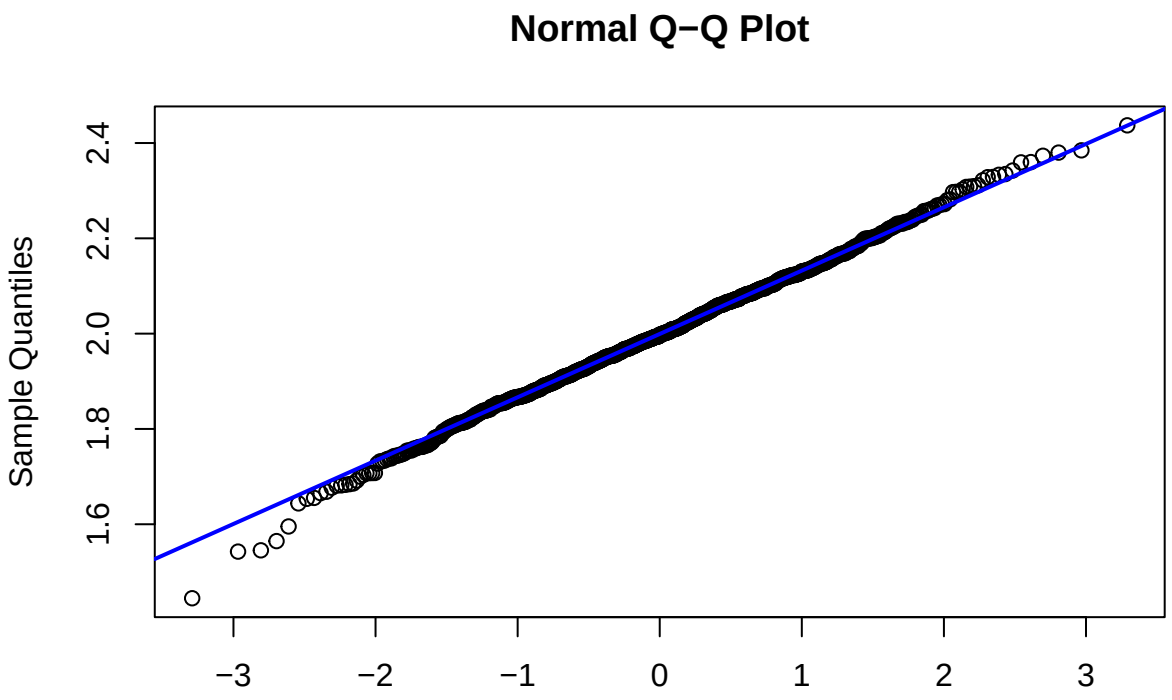

unlist(final_list_q2_slope)
unlist(final_list_q2_int)
count
12
9
6
3
0.50
0.75
1.00
1.25
1.50
1.0
1.5
2.0
2.5

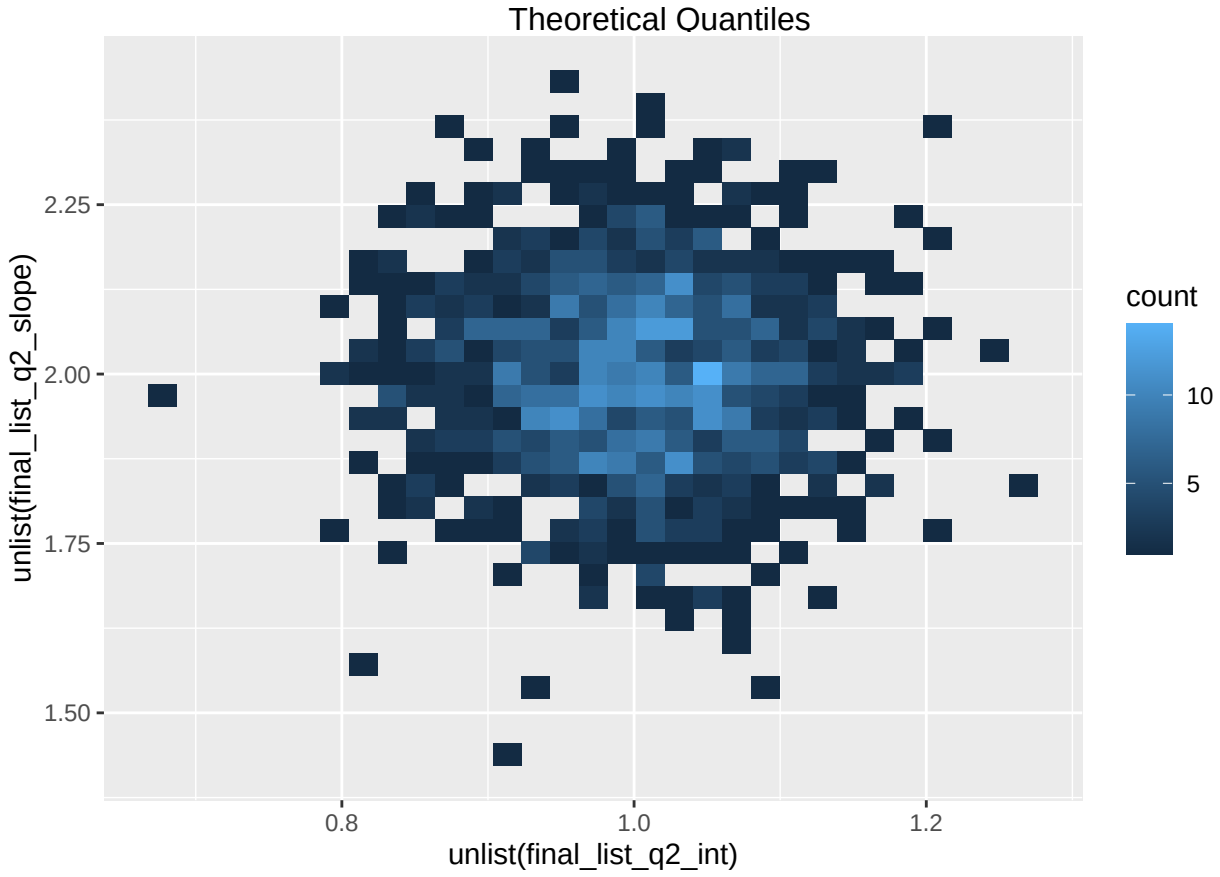

Normal Q−Q Plot
Sample Quantiles
0.7
0.9
1.1
1.3
−3
−2
−1
0
1
2
3
Theoretical Quantiles

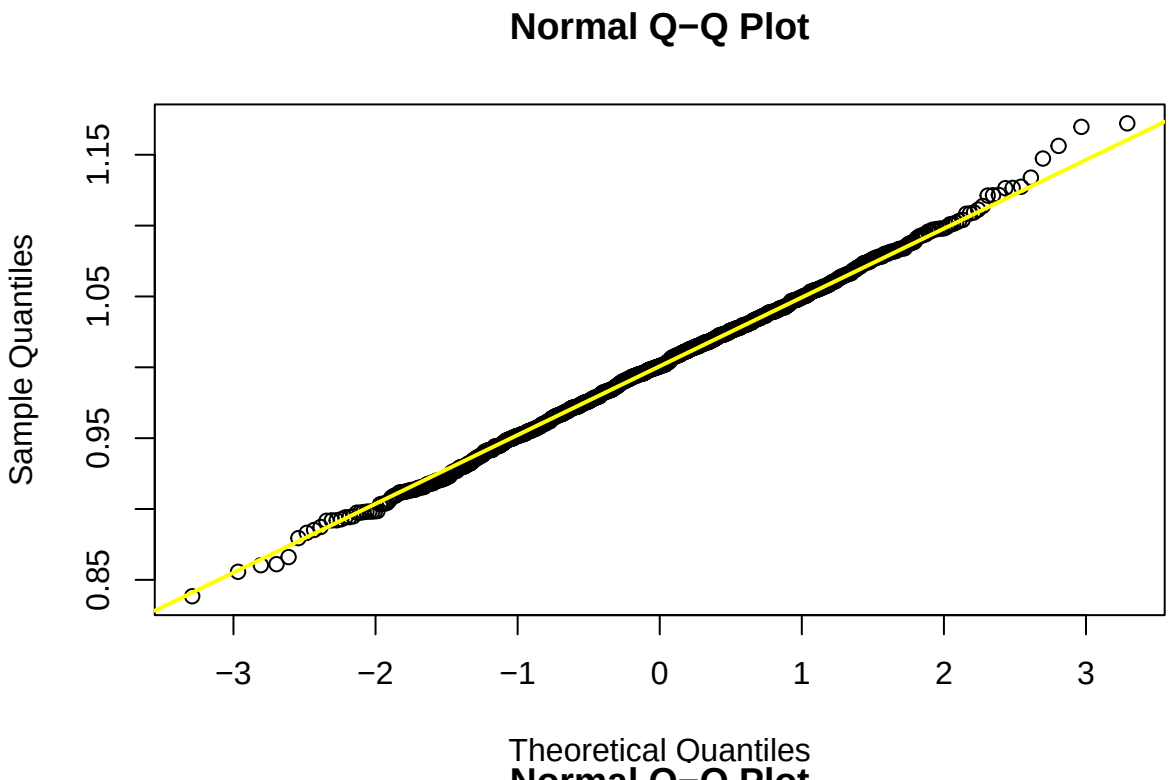

Normal Q−Q Plot
Sample Quantiles
1.4
1.6
1.8
2.0
2.2
2.4
2.6
−3
−2
−1
0
1
2
3
Theoretical Quantiles

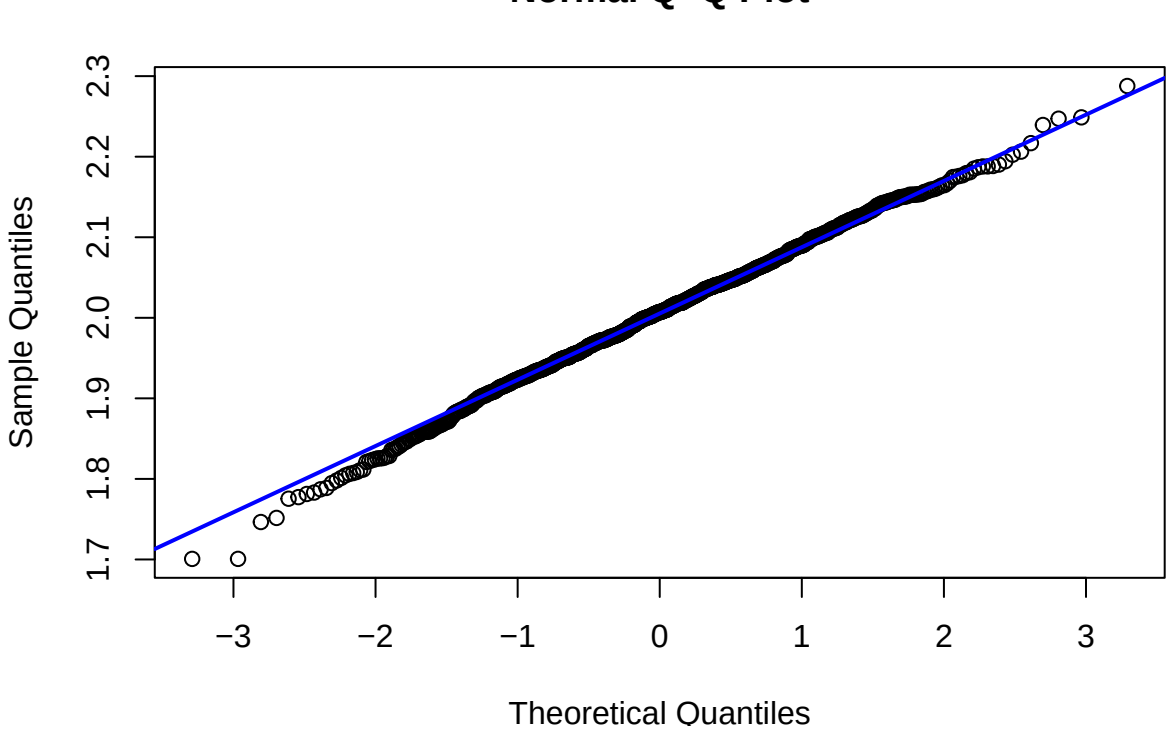

unlist(final_list_q2_slope)
2.4
2.0
1.6
0.6
0.8
1.0
1.2
unlist(final_list_q2_int)
count
12
9
6
3

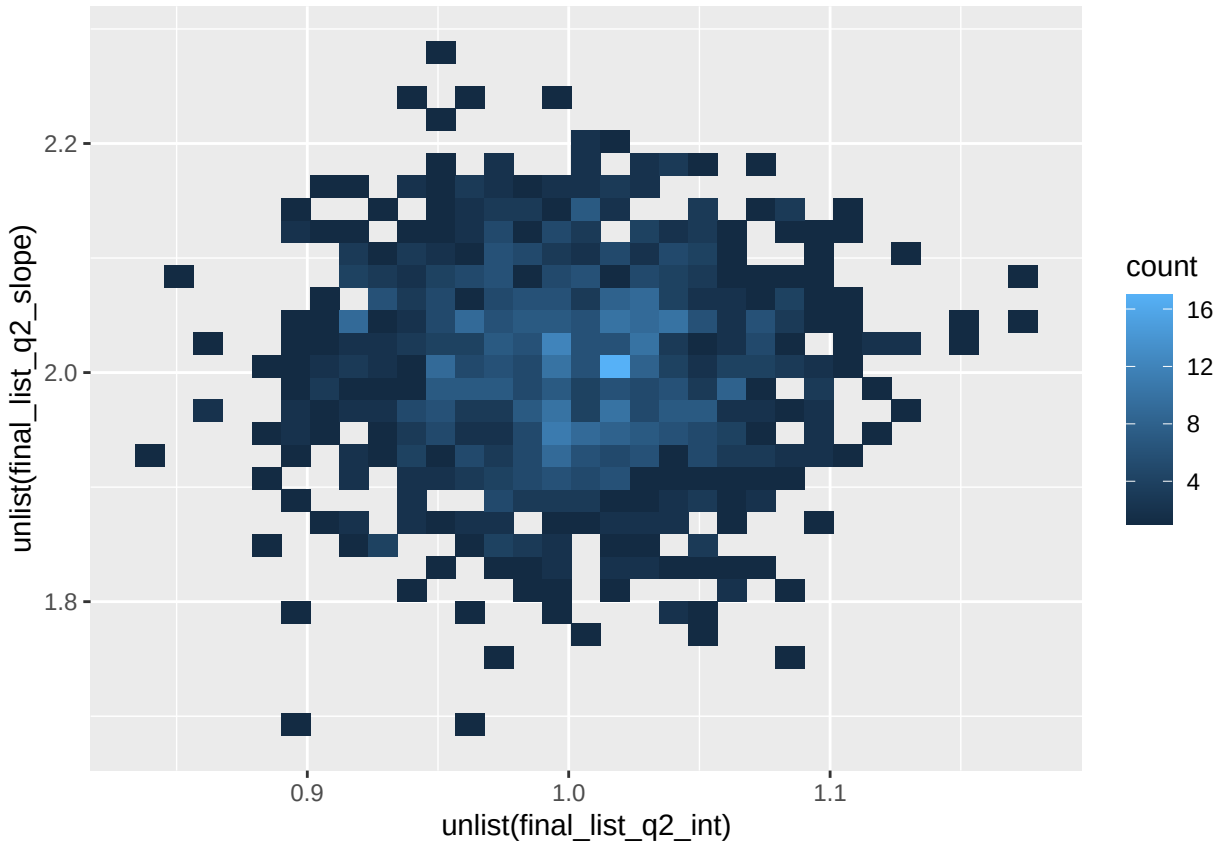

Normal Q−Q Plot
Sample Quantiles
0.7
0.8
0.9
1.0
1.1
1.2
−3
−2
−1
0
1
2
3
Theoretical Quantiles

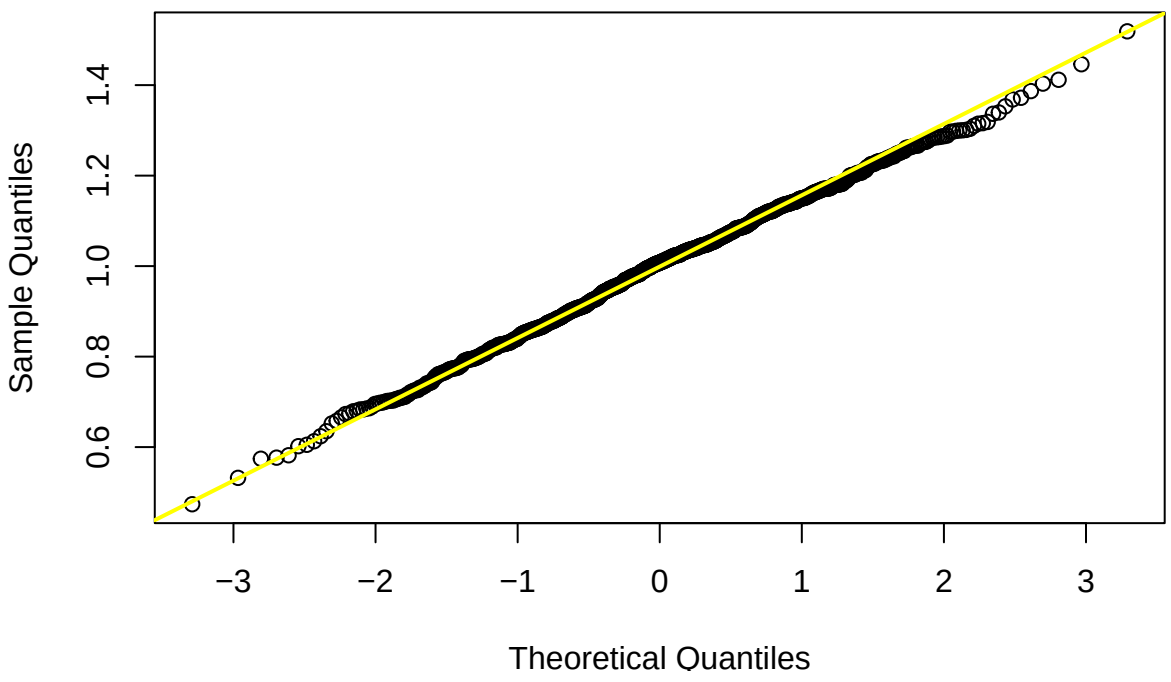

Normal Q−Q Plot
Sample Quantiles
Theoretical Quantiles
1.6
1.8
2.0
2.2
2.4
−3
−2
−1
0
1
2
3

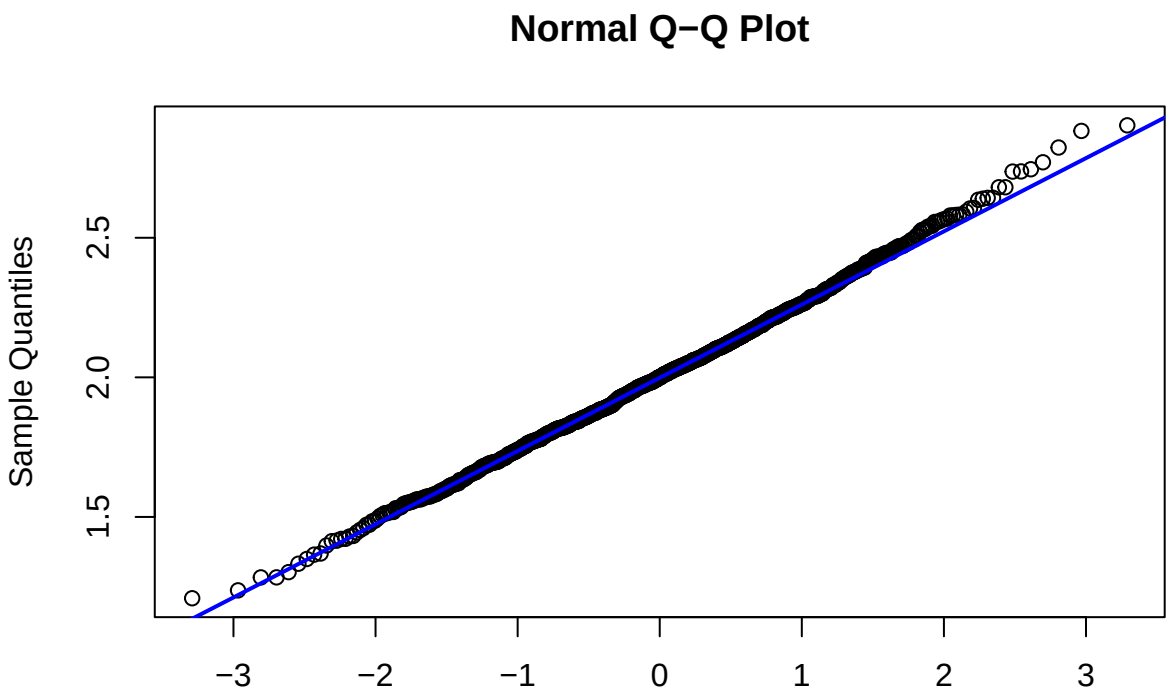

unlist(final_list_q2_slope)
unlist(final_list_q2_int)
count
2.25
2.00
1.75
1.50
0.8
1.0
1.2
10
5

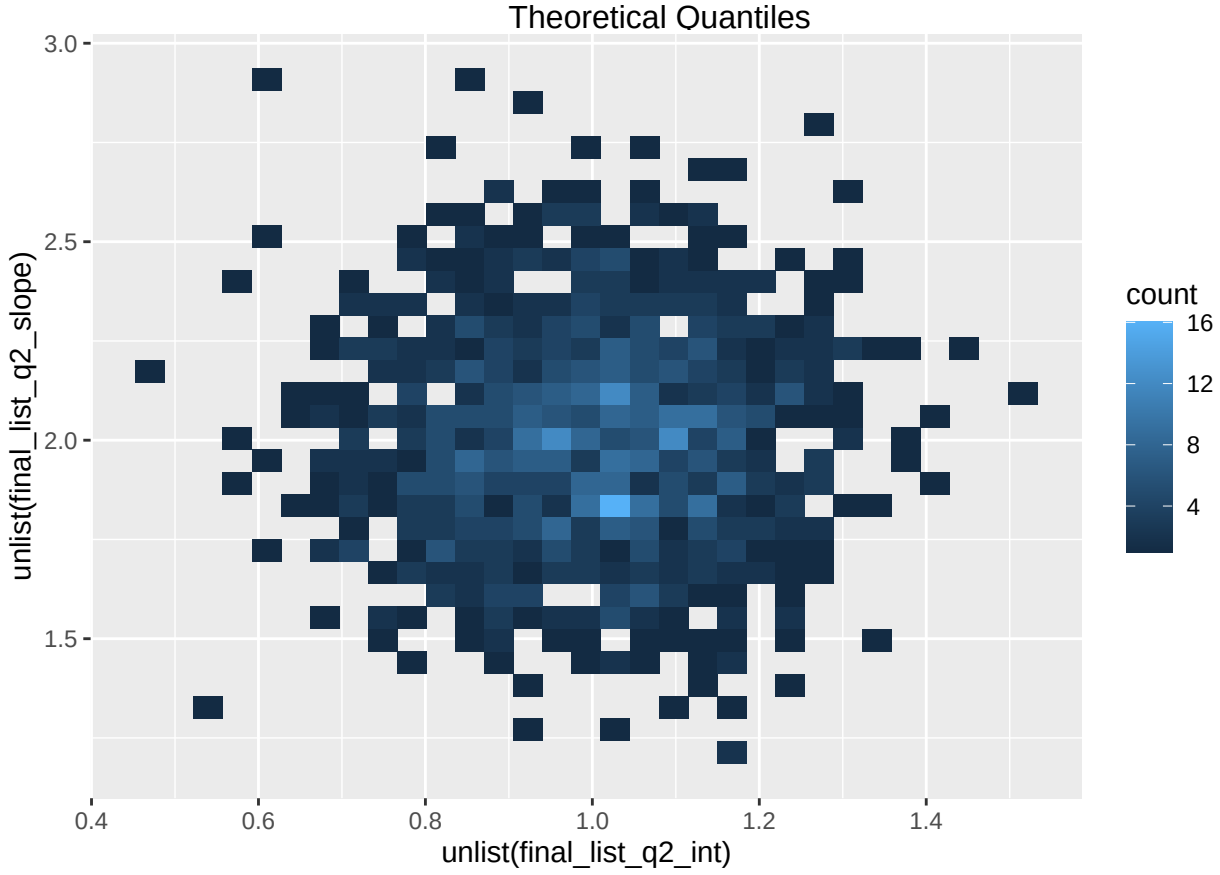

Normal Q−Q Plot
Sample Quantiles
0.85
0.95
1.05
1.15
−3
−2
−1
0
1
2
3
Theoretical Quantiles

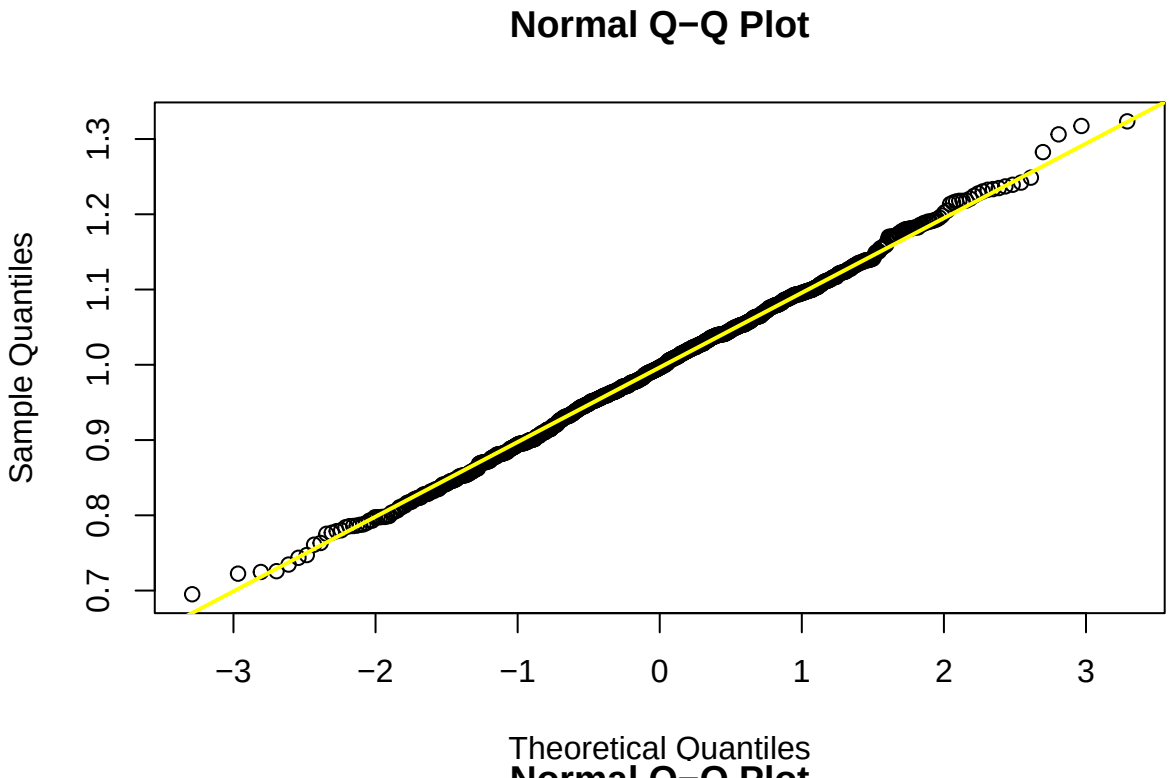

Normal Q−Q Plot
Sample Quantiles
1.7
1.8
1.9
2.0
2.1
2.2
2.3
−3
−2
−1
0
1
2
3
Theoretical Quantiles

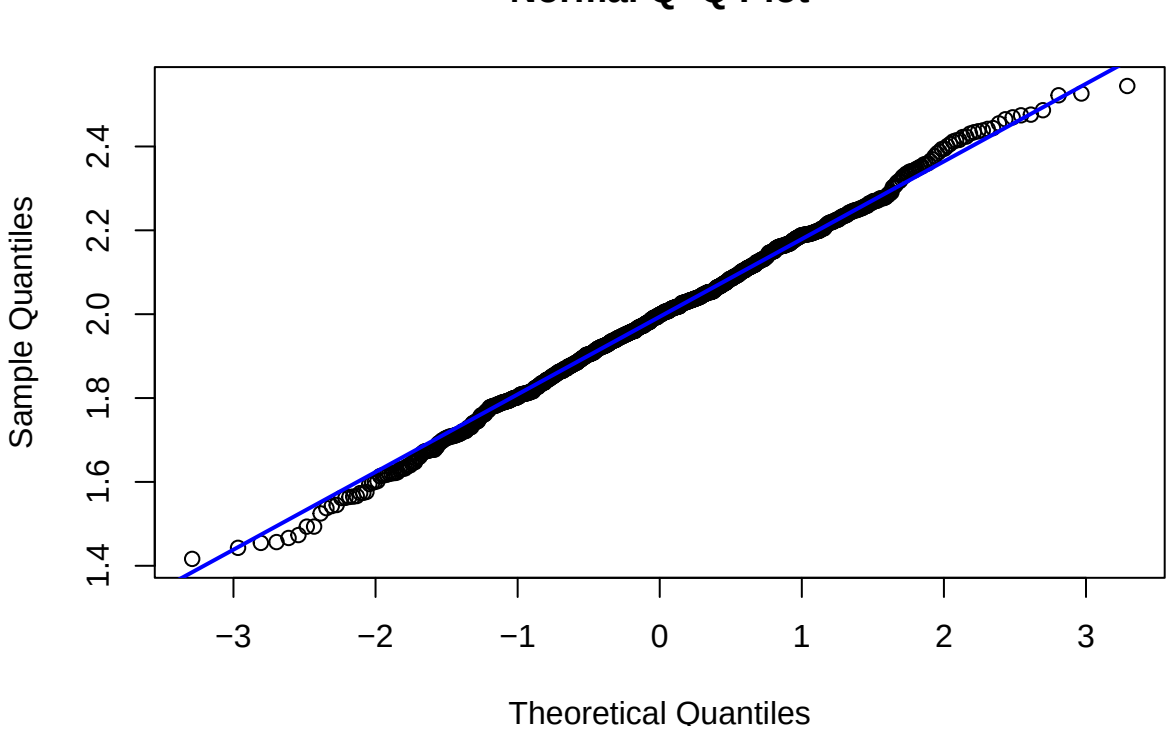

2.2
2.0
1.8
unlist(final_list_q2_slope)
0.9
1.0
1.1
unlist(final_list_q2_int)
count
16
12
8
4

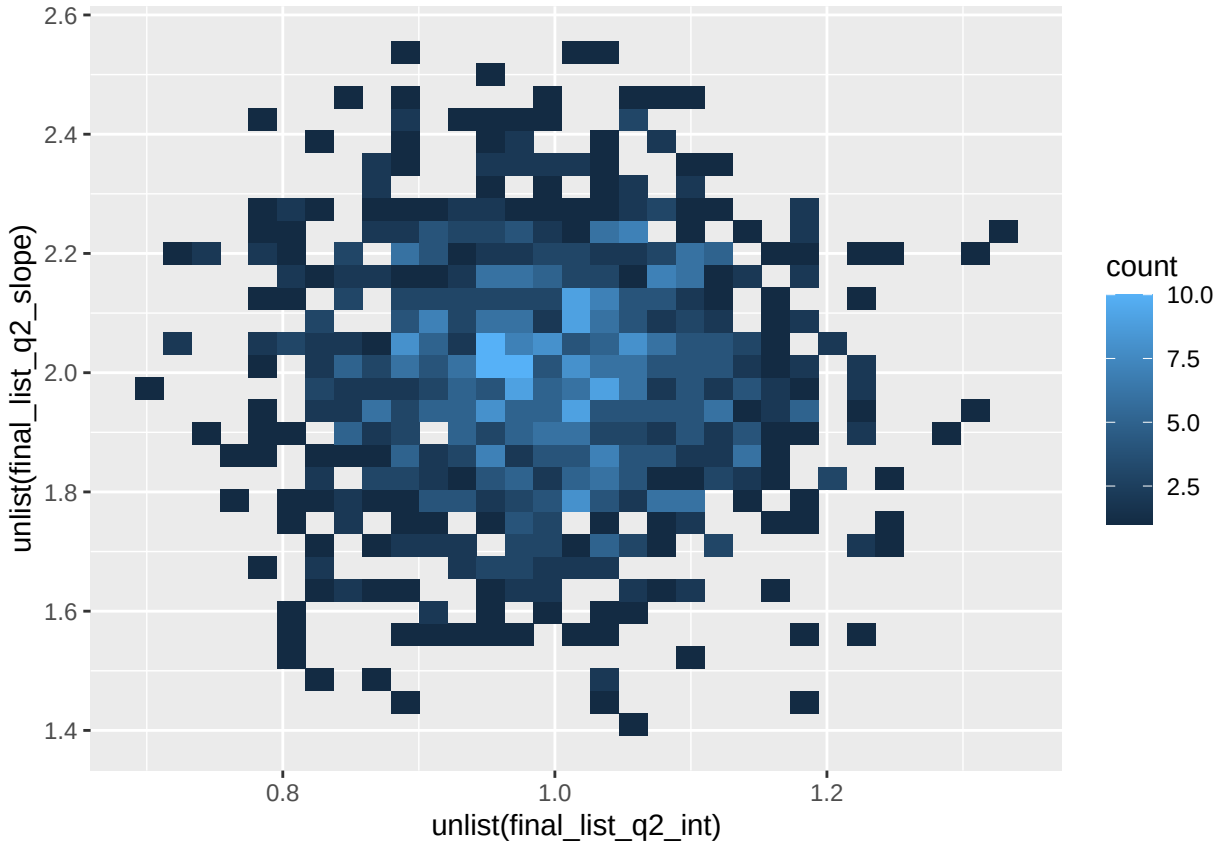

Normal Q−Q Plot
Sample Quantiles
0.6
0.8
1.0
1.2
1.4
−3
−2
−1
0
1
2
3
Theoretical Quantiles

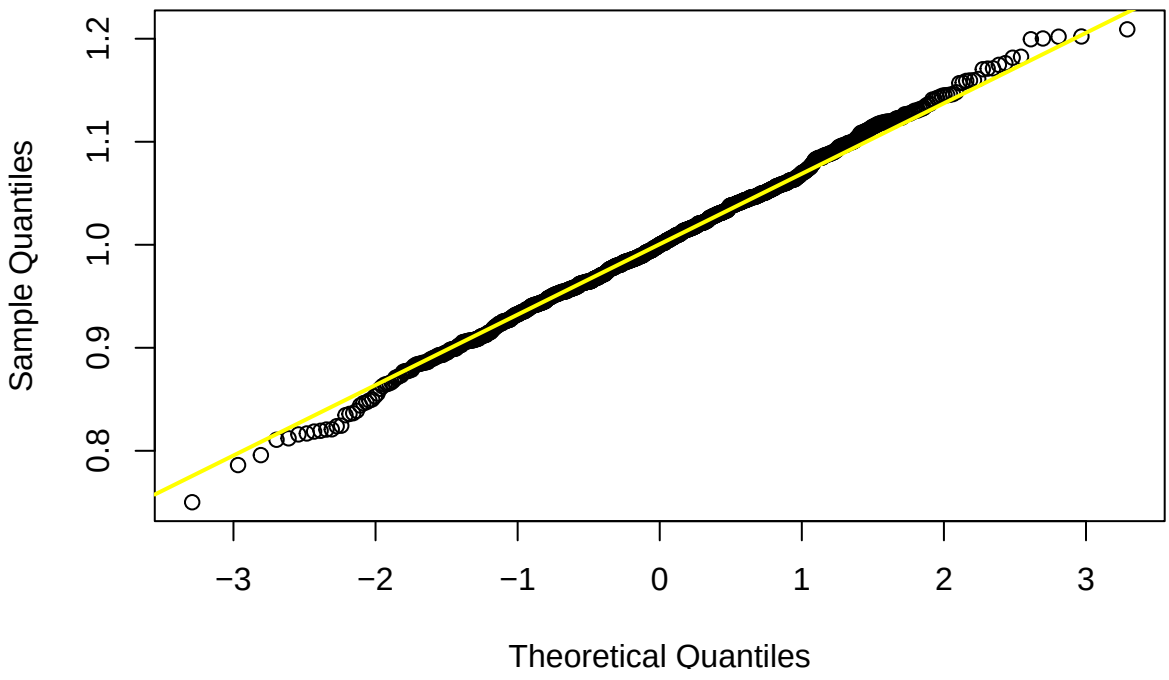

Normal Q−Q Plot
Sample Quantiles
Theoretical Quantiles
1.5
2.0
2.5
−3
−2
−1
0
1
2
3

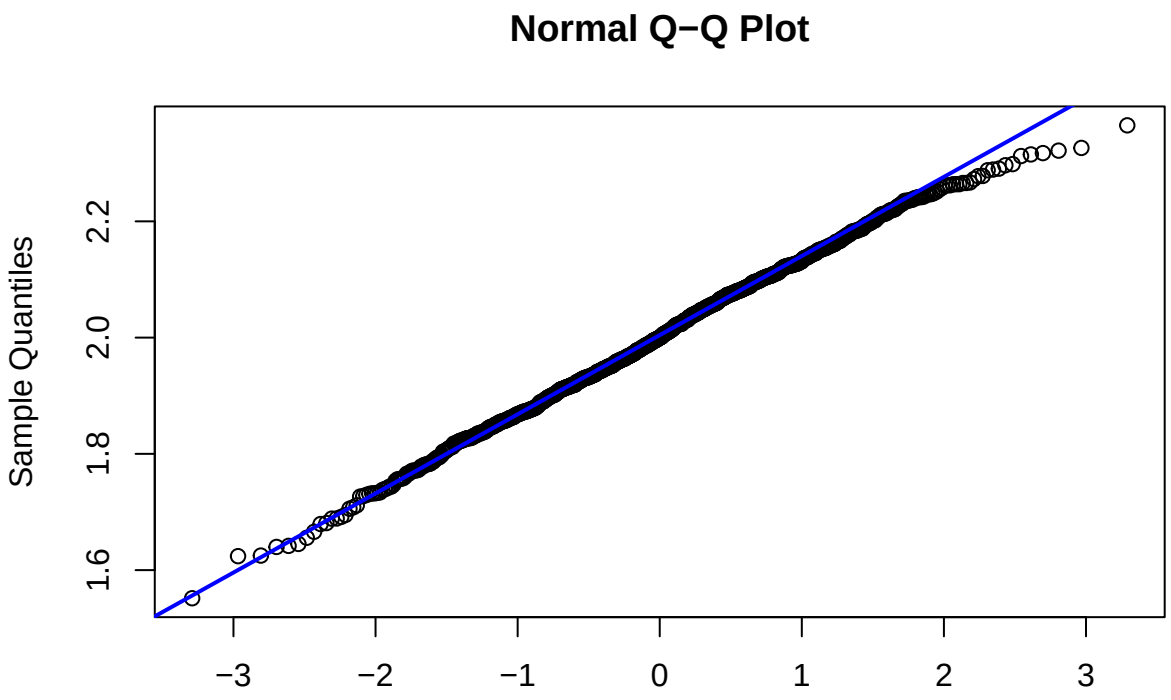

unlist(final_list_q2_slope)
unlist(final_list_q2_int)
count
16
12
8
4
3.0
2.5
2.0
1.5
0.4
0.6
0.8
1.0
1.2
1.4

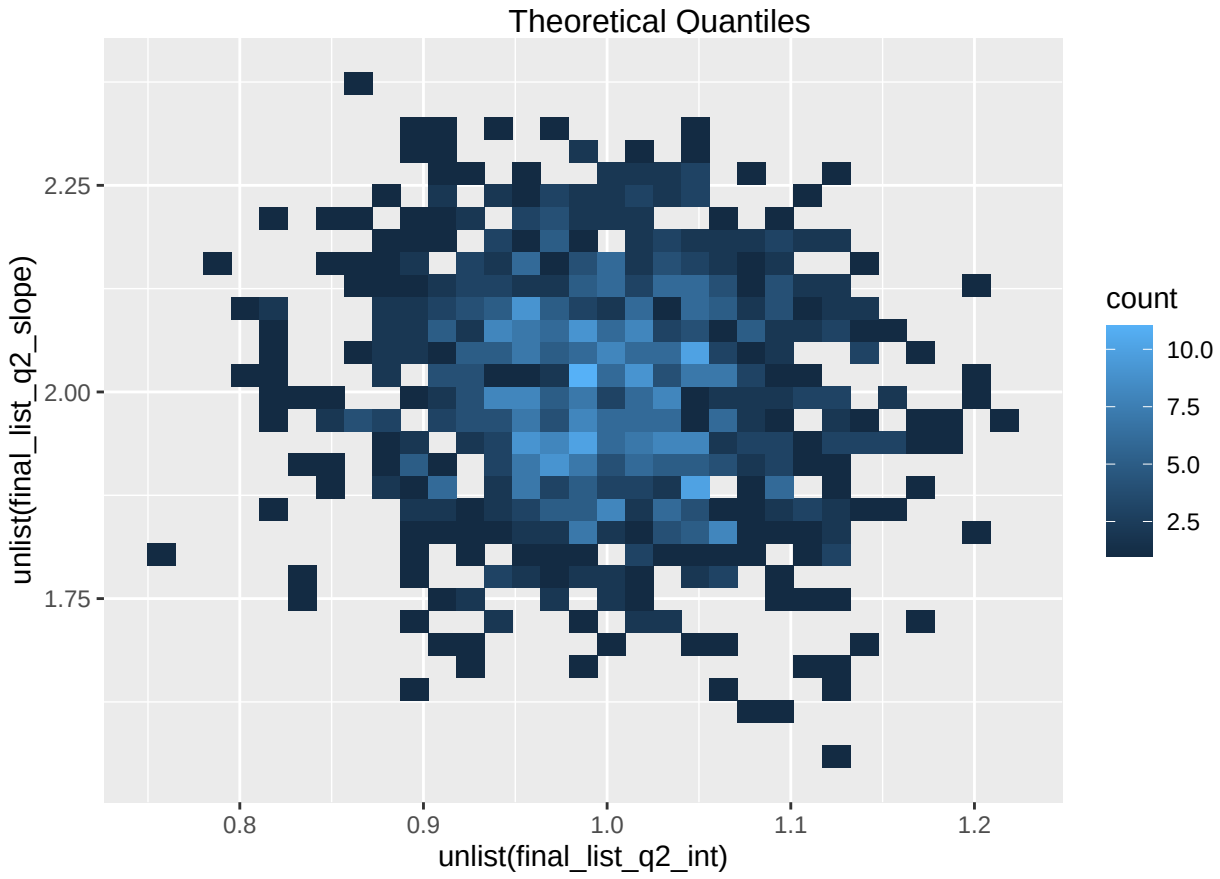

Normal Q−Q Plot
Sample Quantiles
0.7
0.8
0.9
1.0
1.1
1.2
1.3
−3
−2
−1
0
1
2
3
Theoretical Quantiles

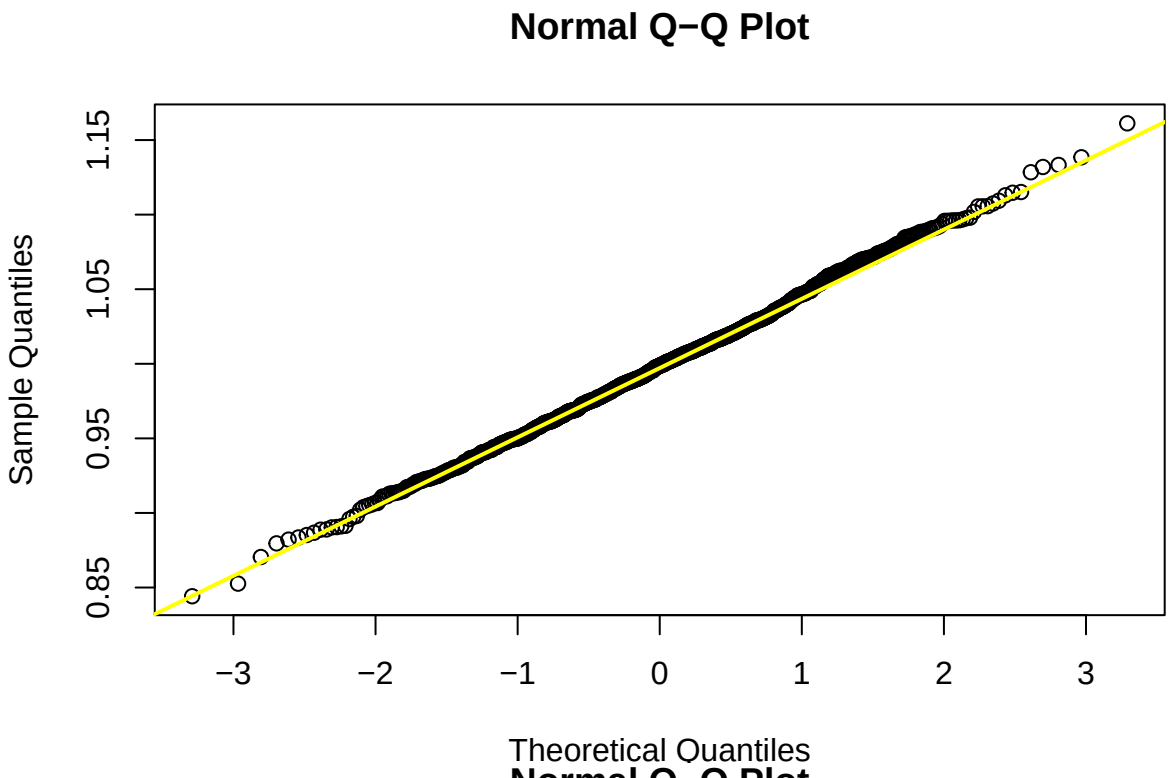

Normal Q−Q Plot
Sample Quantiles
1.4
1.6
1.8
2.0
2.2
2.4
−3
−2
−1
0
1
2
3
Theoretical Quantiles

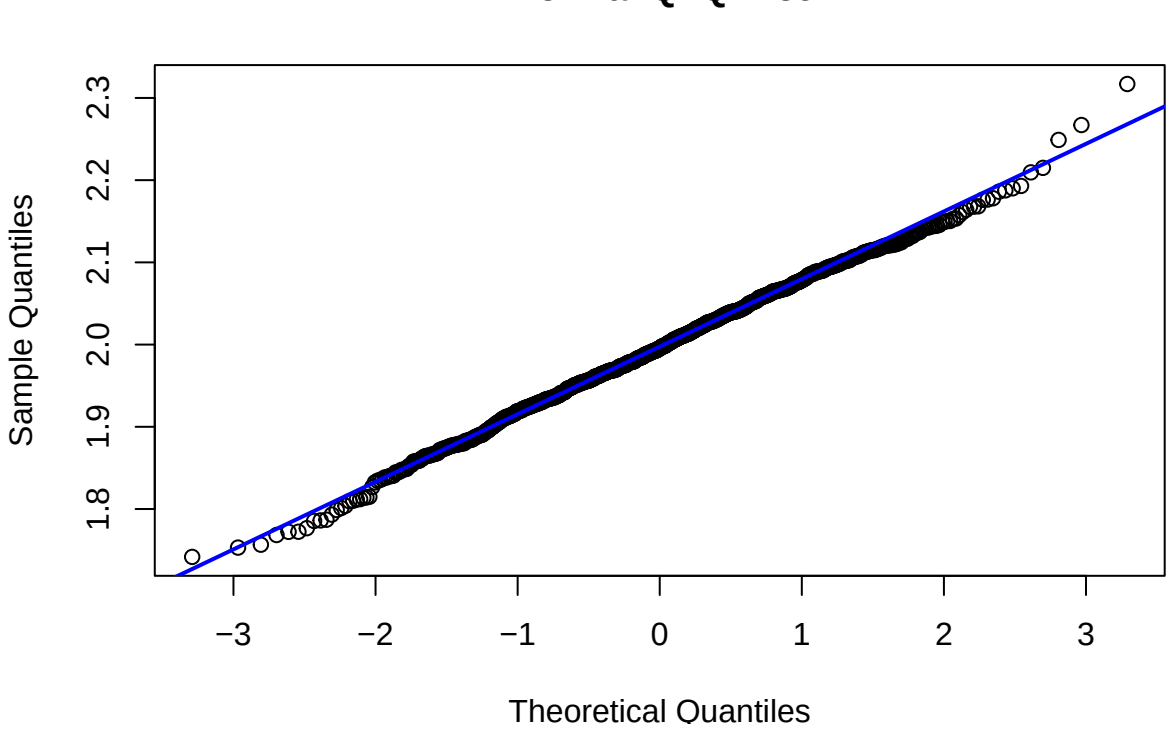

2.6
2.4
2.2
2.0
1.8
1.6
1.4
unlist(final_list_q2_slope)
0.8
1.0
1.2
unlist(final_list_q2_int)
count
10.0
7.5
5.0
2.5

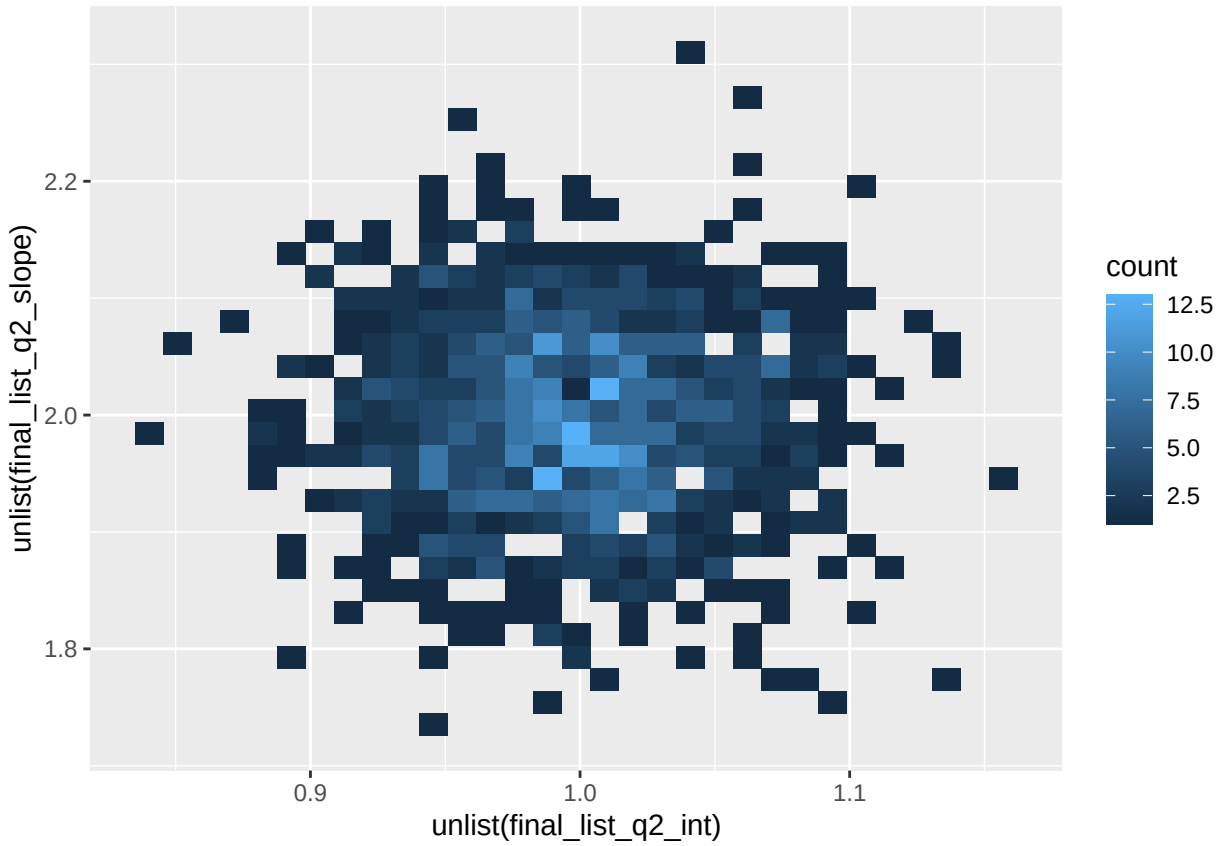

Normal Q−Q Plot
Sample Quantiles
0.8
0.9
1.0
1.1
1.2
−3
−2
−1
0
1
2
3
Theoretical Quantiles

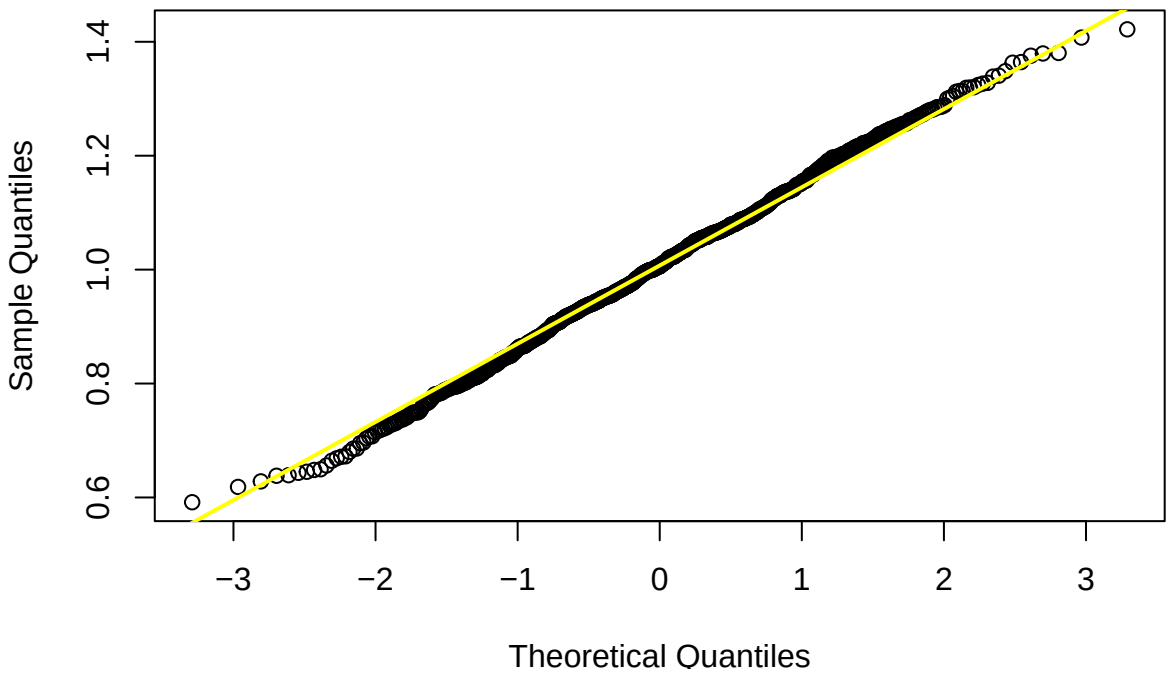

Normal Q−Q Plot
Sample Quantiles
Theoretical Quantiles
1.6
1.8
2.0
2.2
−3
−2
−1
0
1
2
3

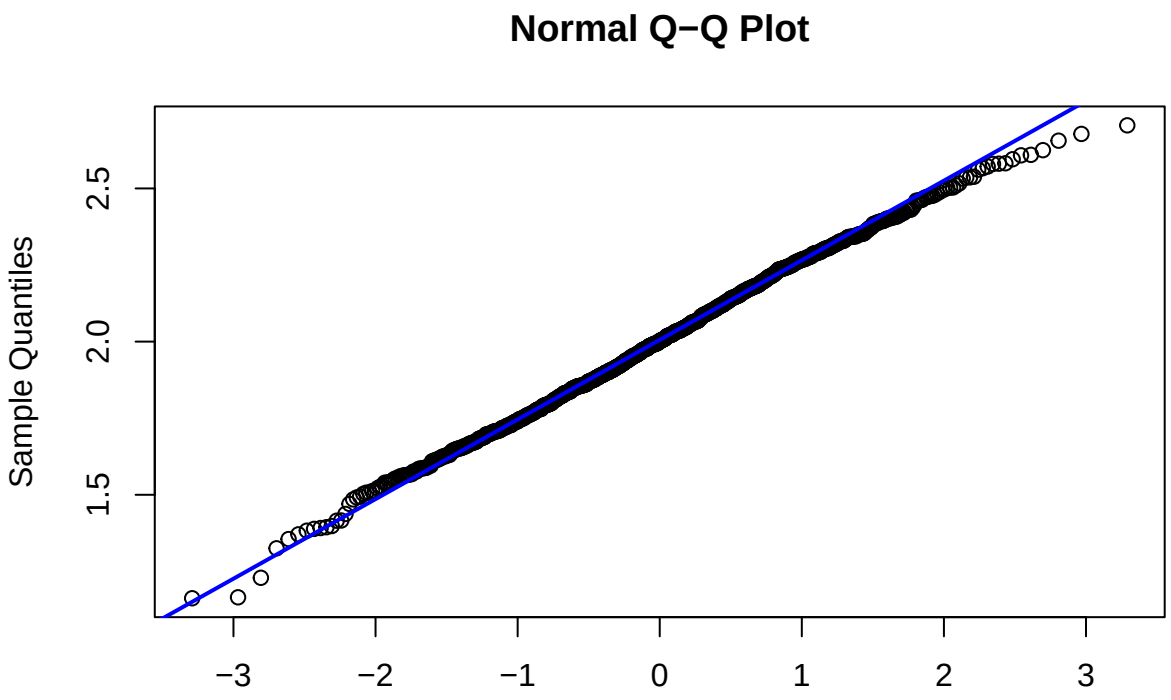

unlist(final_list_q2_slope)
unlist(final_list_q2_int)
2.25
2.00
1.75
0.8
0.9
1.0
1.1
1.2
count
10.0
7.5
5.0
2.5

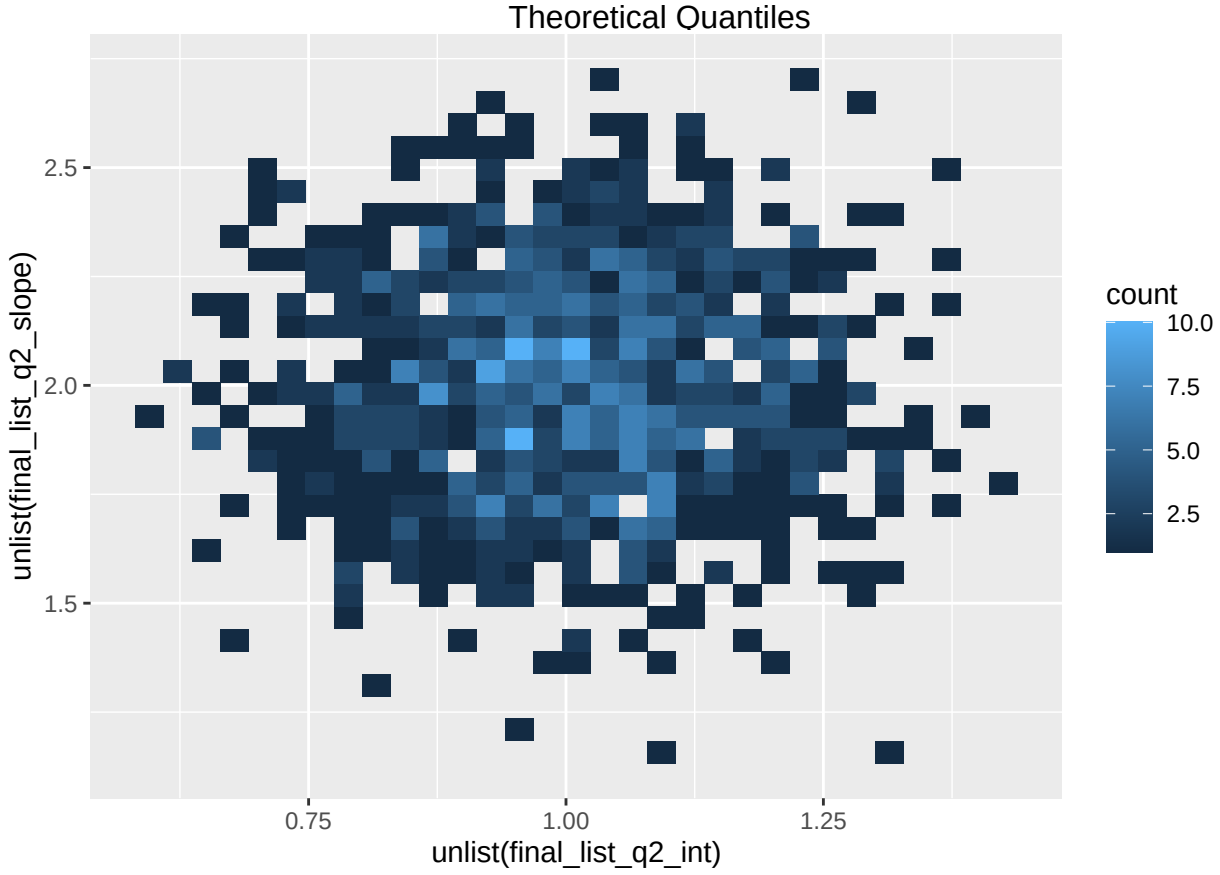

Normal Q−Q Plot
Sample Quantiles
0.85
0.95
1.05
1.15
−3
−2
−1
0
1
2
3
Theoretical Quantiles

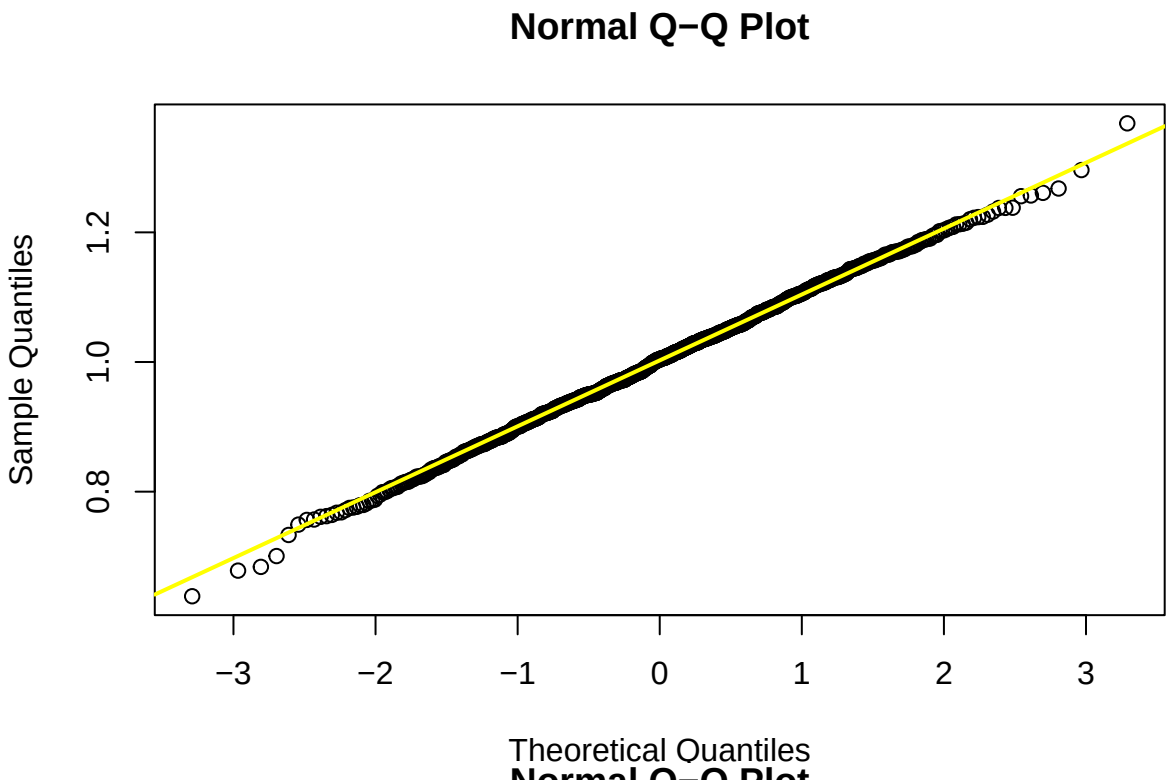

Normal Q−Q Plot
Sample Quantiles
1.8
1.9
2.0
2.1
2.2
2.3
−3
−2
−1
0
1
2
3
Theoretical Quantiles

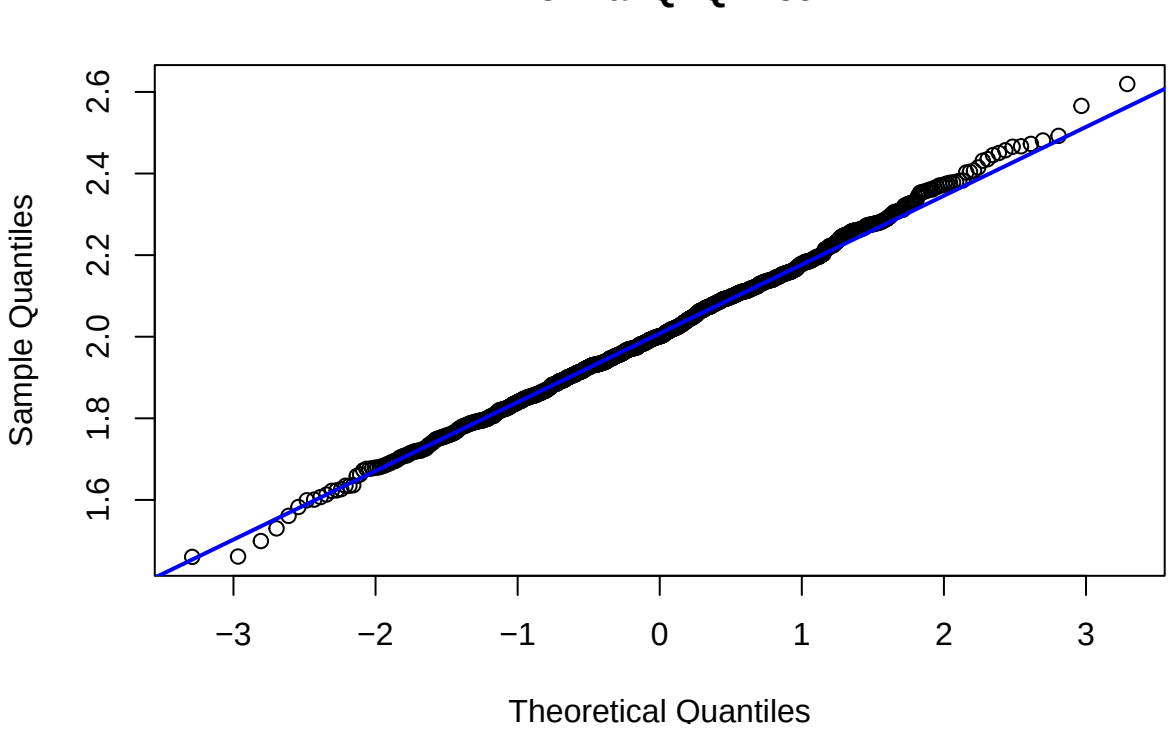

unlist(final_list_q2_slope)
2.2
2.0
1.8
0.9
1.0
1.1
unlist(final_list_q2_int)
count
12.5
10.0
7.5
5.0
2.5

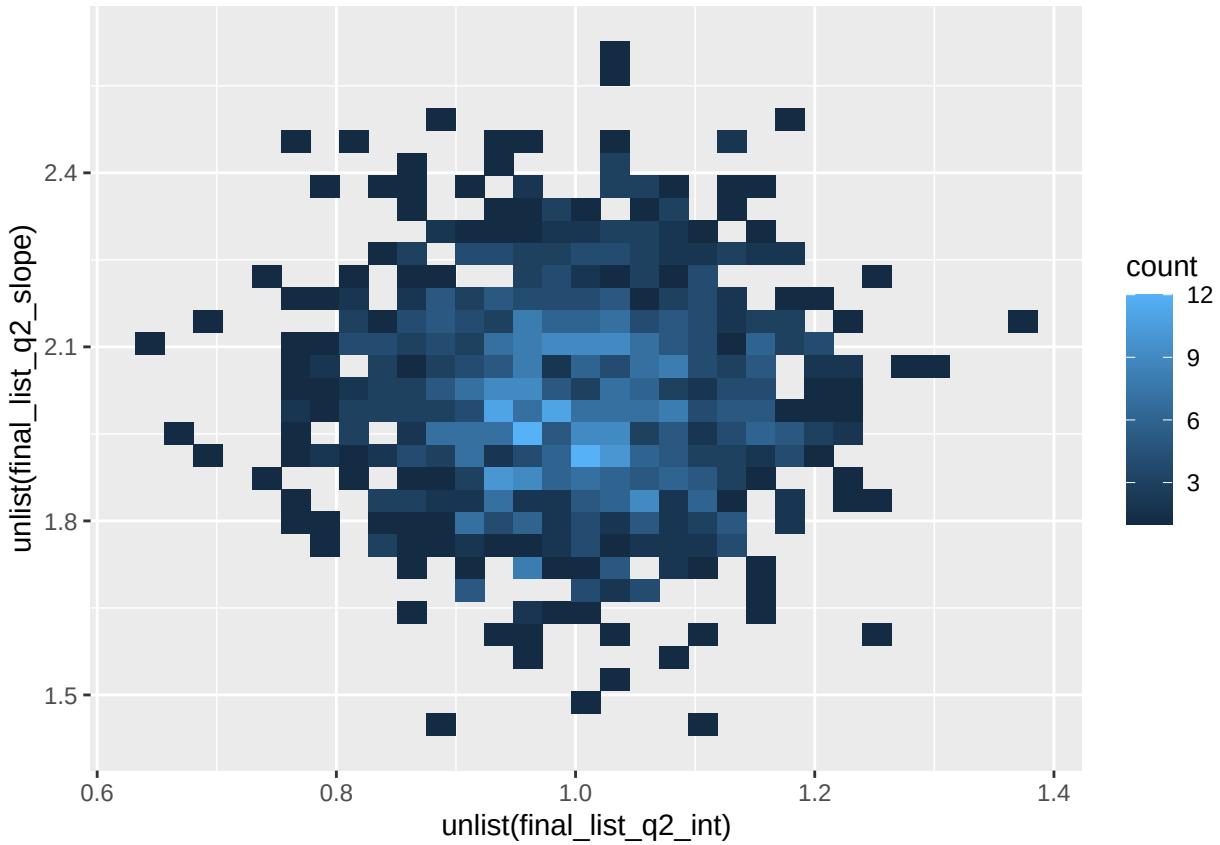

Normal Q−Q Plot
Sample Quantiles
0.6
0.8
1.0
1.2
1.4
−3
−2
−1
0
1
2
3
Theoretical Quantiles

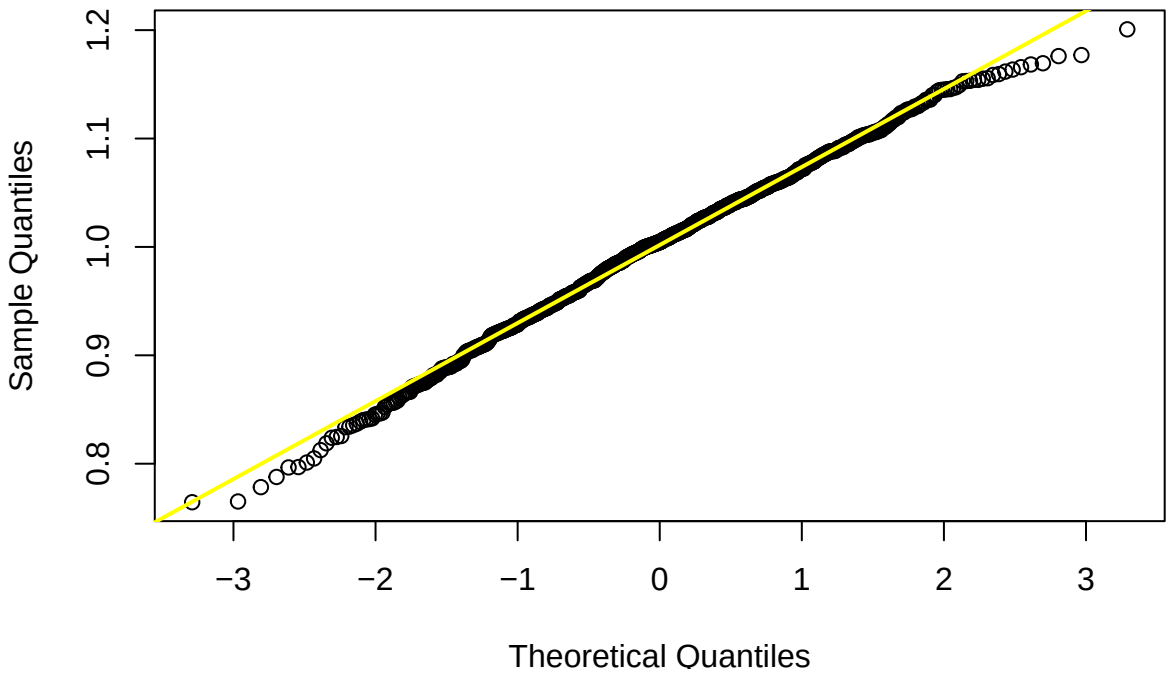

Normal Q−Q Plot
Sample Quantiles
Theoretical Quantiles
2.5
2.0
1.5
−3
−2
−1
0
1
2
3

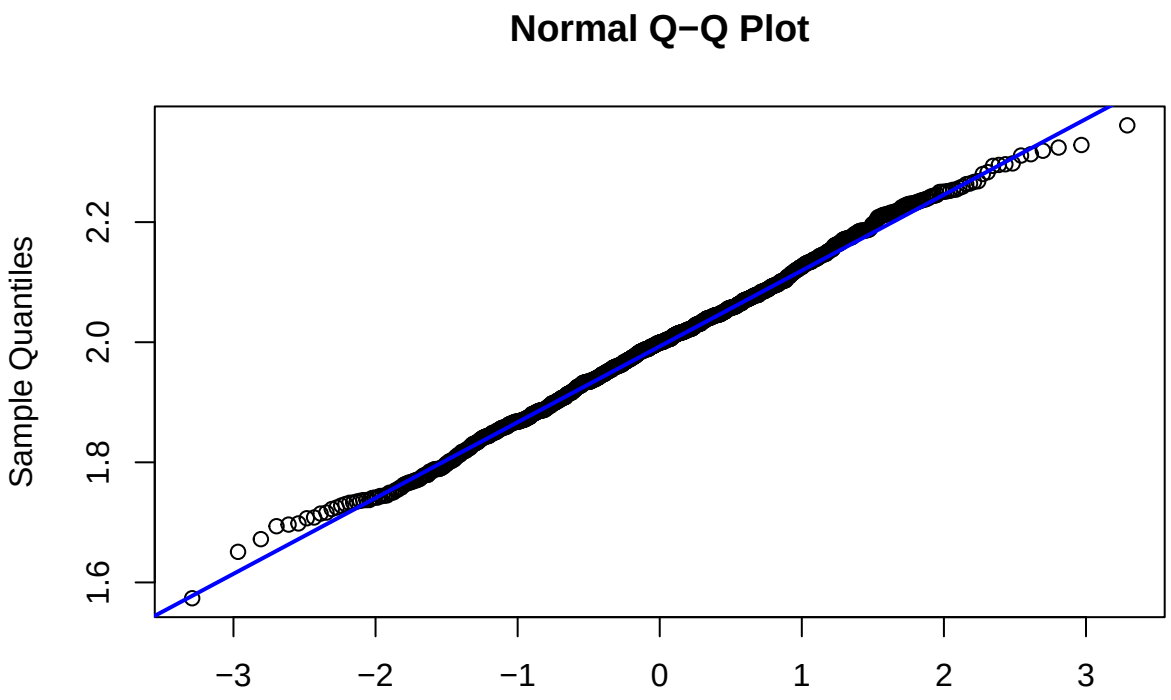

unlist(final_list_q2_slope)
unlist(final_list_q2_int)
2.5
2.0
1.5
0.75
1.00
1.25
count
10.0
7.5
5.0
2.5

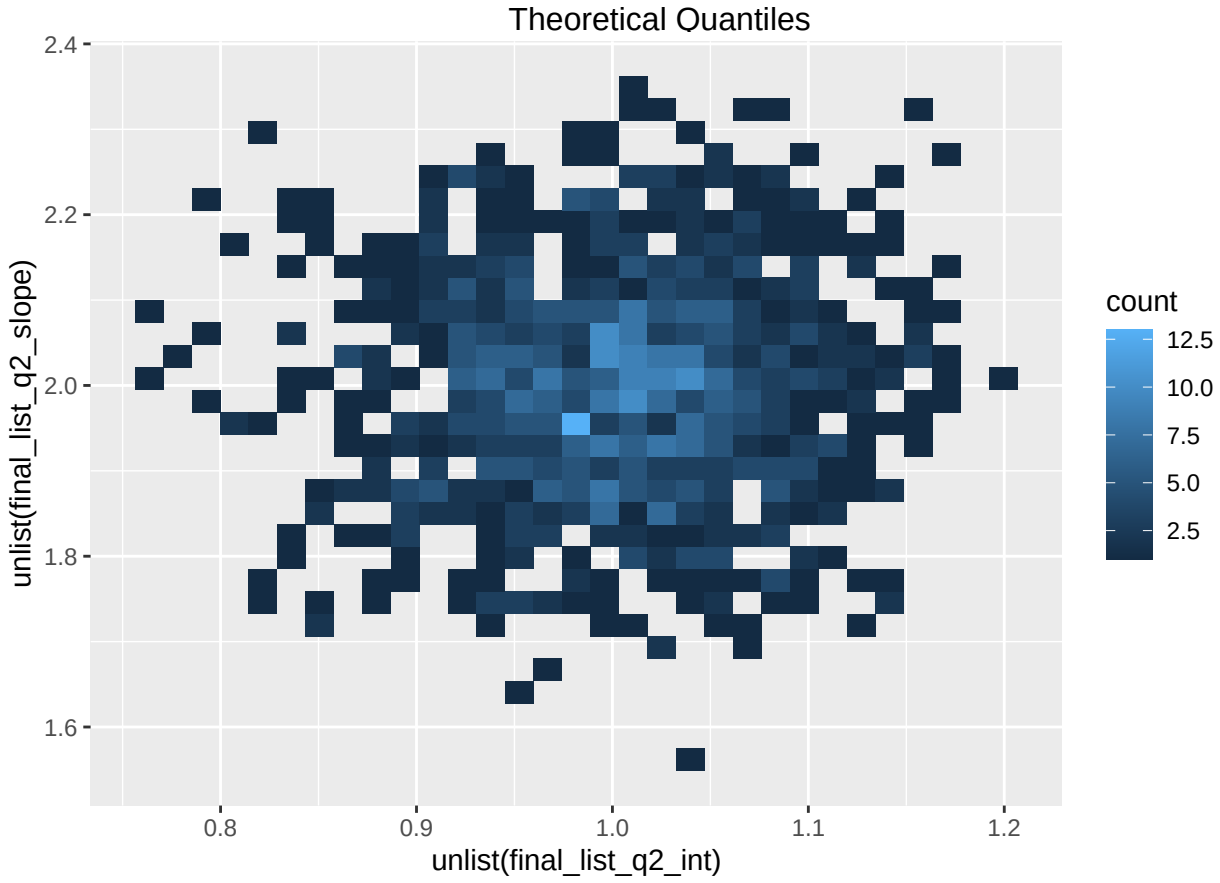

Normal Q−Q Plot
Sample Quantiles
0.8
1.0
1.2
−3
−2
−1
0
1
2
3
Theoretical Quantiles

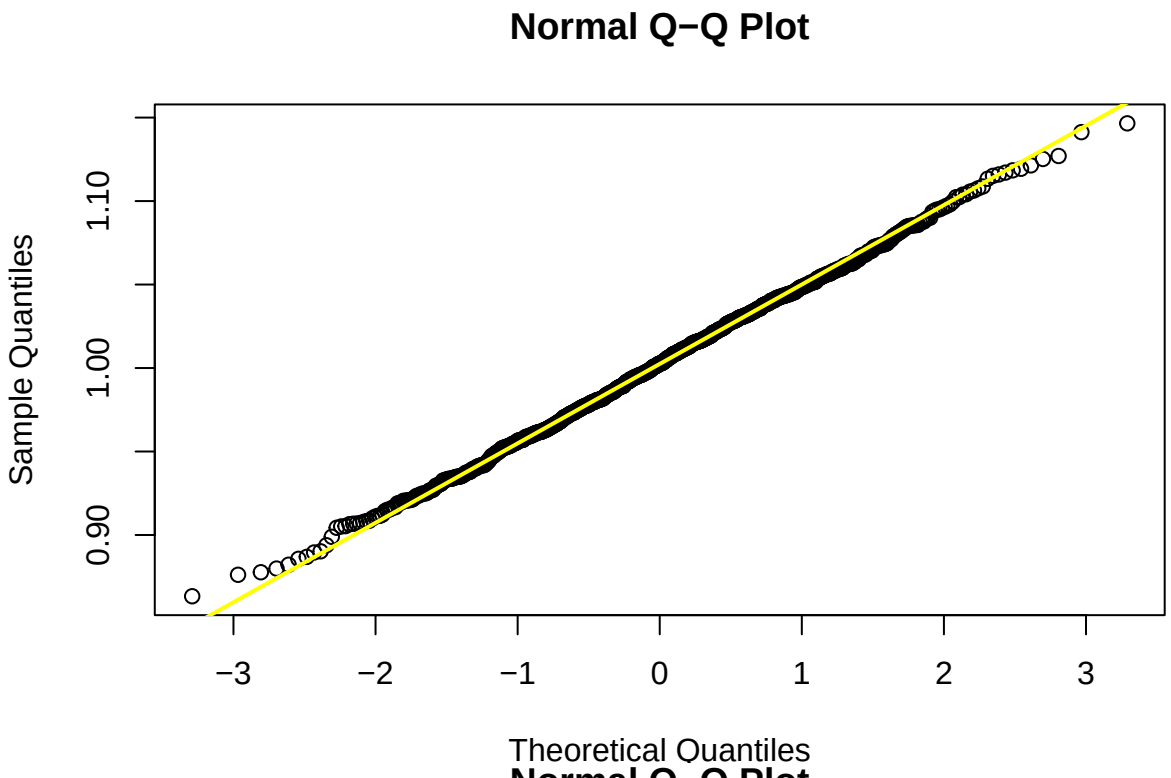

Normal Q−Q Plot
Sample Quantiles
1.6
1.8
2.0
2.2
2.4
2.6
−3
−2
−1
0
1
2
3
Theoretical Quantiles

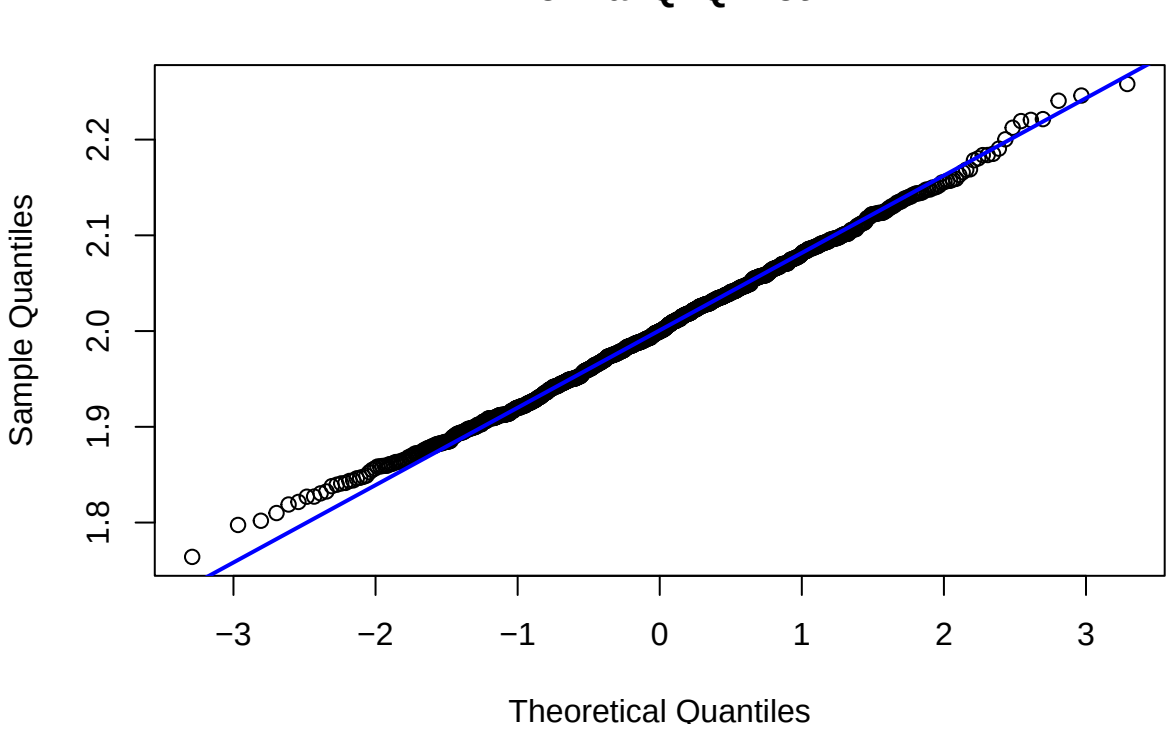

unlist(final_list_q2_slope)
2.4
2.1
1.8
1.5
0.6
0.8
1.0
1.2
1.4
unlist(final_list_q2_int)
count
12
9
6
3

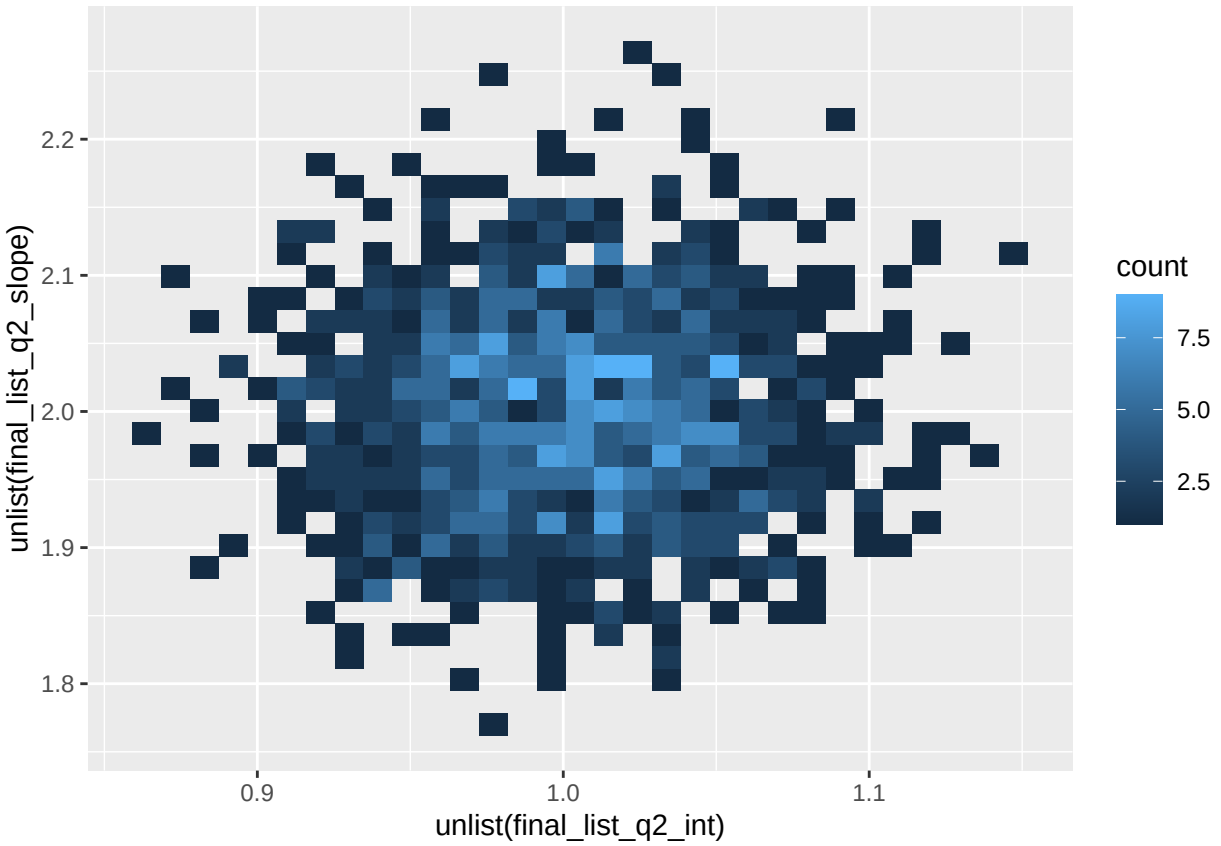

Normal Q−Q Plot
Sample Quantiles
0.8
0.9
1.0
1.1
1.2
−3
−2
−1
0
1
2
3
Theoretical Quantiles


Normal Q−Q Plot
Sample Quantiles
Theoretical Quantiles
1.6
1.8
2.0
2.2
−3
−2
−1
0
1
2
3


unlist(final_list_q2_slope)
unlist(final_list_q2_int)
count
12.5
10.0
7.5
5.0
2.5
2.4
2.2
2.0
1.8
1.6
0.8
0.9
1.0
1.1
1.2


Normal Q−Q Plot
Sample Quantiles
0.90
1.00
1.10
−3
−2
−1
0
1
2
3
Theoretical Quantiles


Normal Q−Q Plot
Sample Quantiles
1.8
1.9
2.0
2.1
2.2
−3
−2
−1
0
1
2
3
Theoretical Quantiles



It can be seen that the distribution of the least squares coefficients resembles the normal distribution more and more as the degrees of freedom rise. The distribution of the least squares coefficients. The distribution of the least squares coefficients also becomes more stable and less influenced by the selection of the distribution of errors as sample size rises.

Question 2

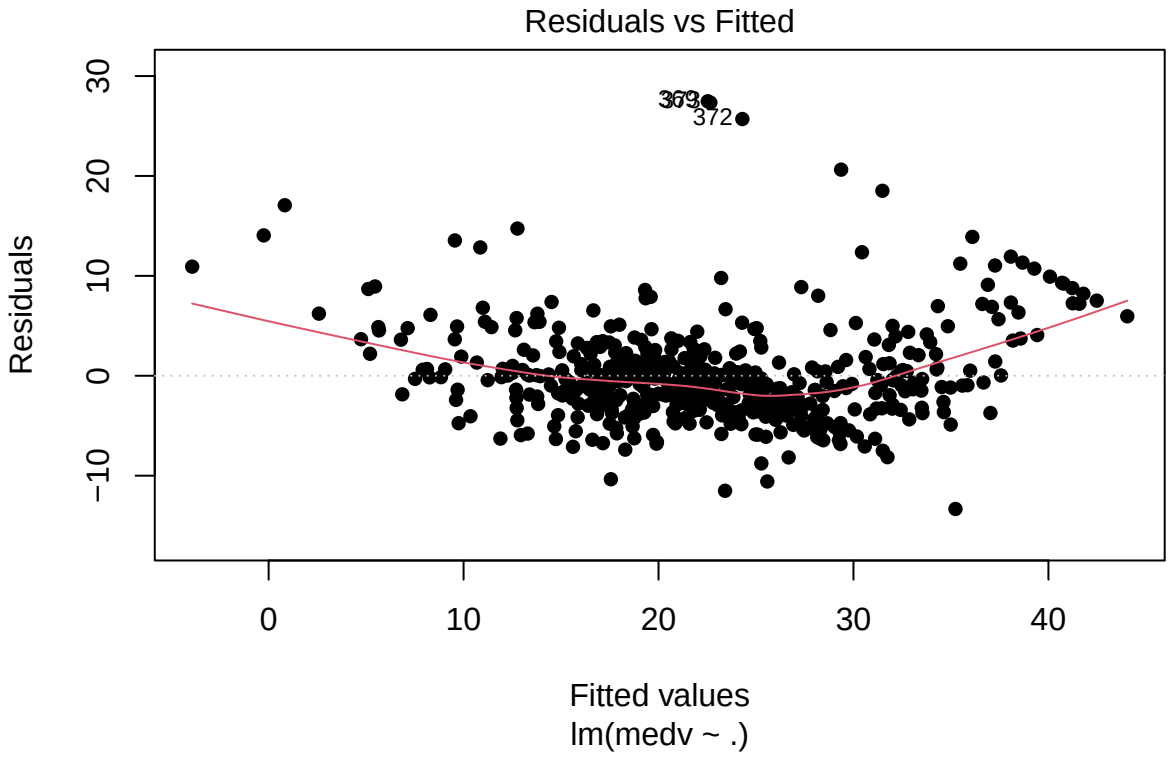
This problem is meant as a practice for performing diagnostics. Consider the Boston dataset in the package MASS. Fit a linear model with medv as response, omitting all the discrete predictor variables.

Solution: We first import the Boston data

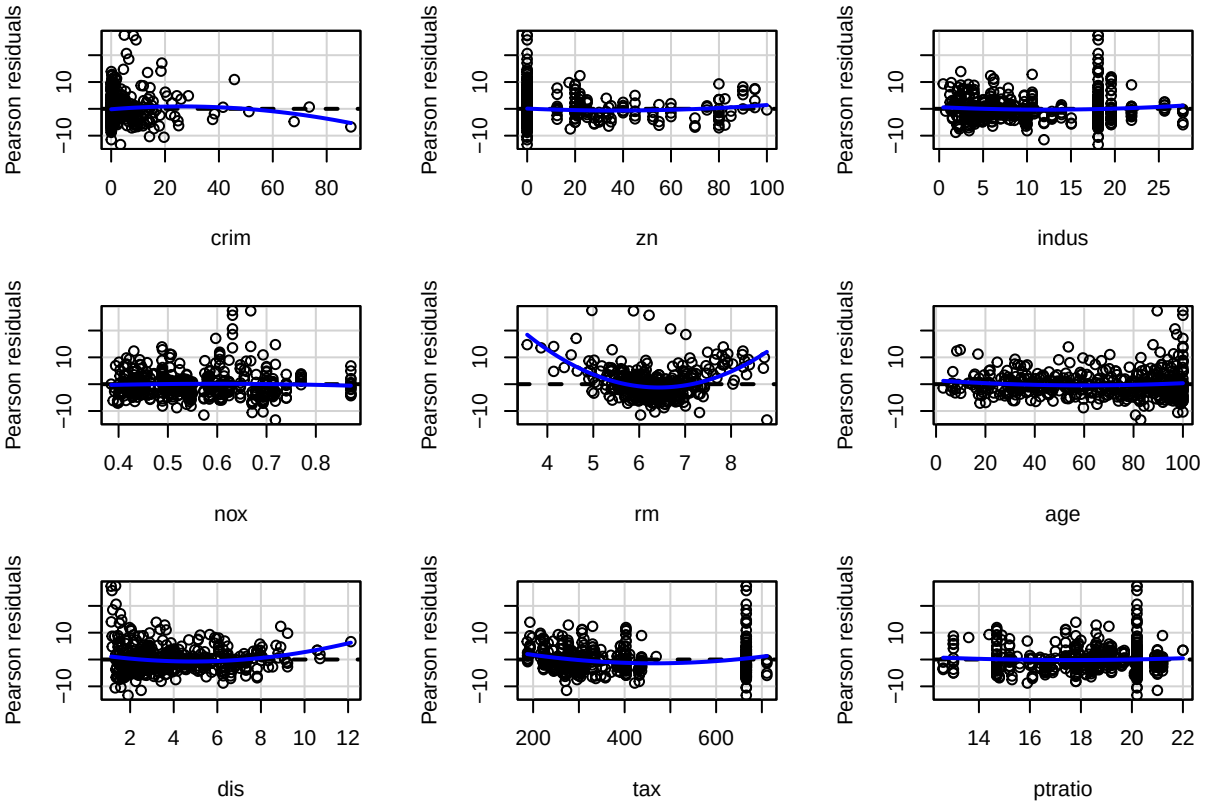
```
library(MASS)
data("Boston")
data_set_final<-subset(Boston,select =-c(4,9))
```

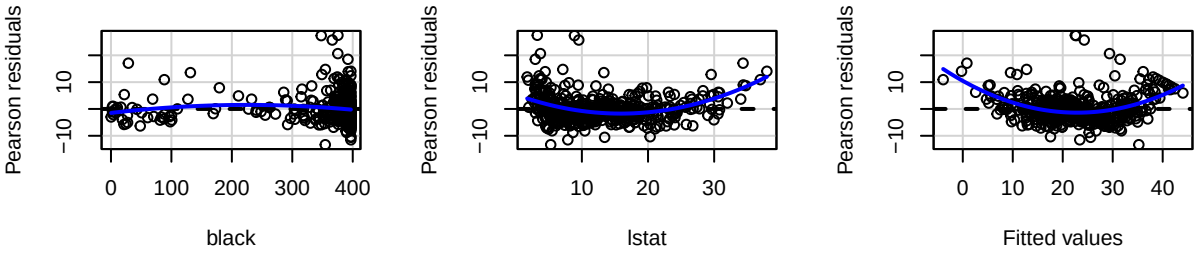
```
#Fitting the model
model_fit=lm(medv~.,data=data_set_final)
```

```
#plotingt the model
plot(model_fit ,which=1,pch=16)
```




```
# Seeing residuals
residualPlots(model_fit)
```





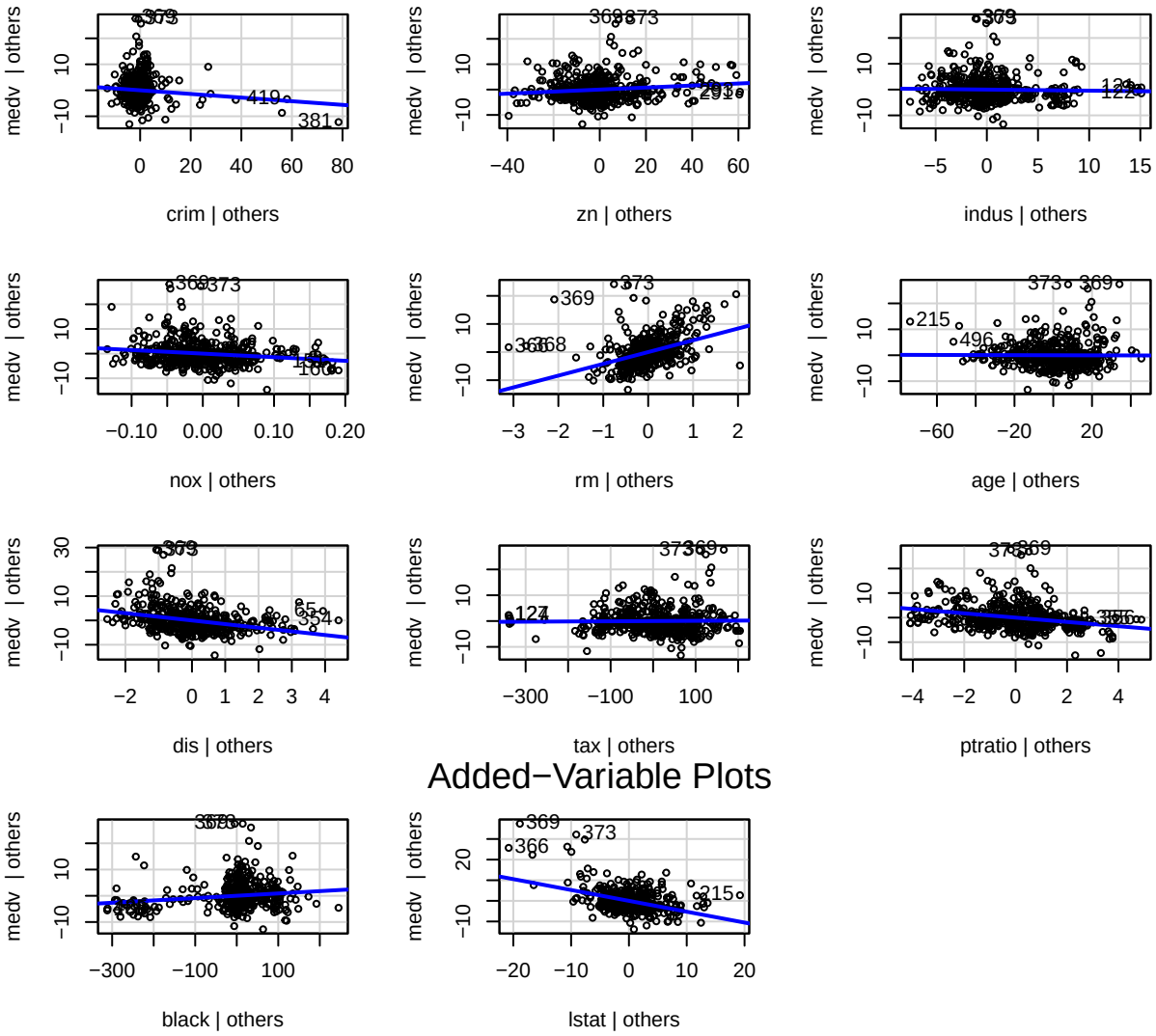

```
##            Test stat Pr(>|Test stat|)
## crim         -2.4854          0.01327 *
## zn            1.4242          0.15503
## indus         1.2836          0.19989
## nox          -1.0419          0.29799
## rm           12.6039        < 2.2e-16 ***
## age           1.7050          0.08883 .
## dis           4.3938        1.365e-05 ***
## tax           4.0370        6.275e-05 ***
## ptratio       0.9109          0.36282
## black        -2.1834          0.02948 *
## lstat        10.5122        < 2.2e-16 ***
## Tukey test   14.0829        < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

1) Mean

```
mean(model_fit$residuals)
```

```
## [1] 1.685082e-16
```

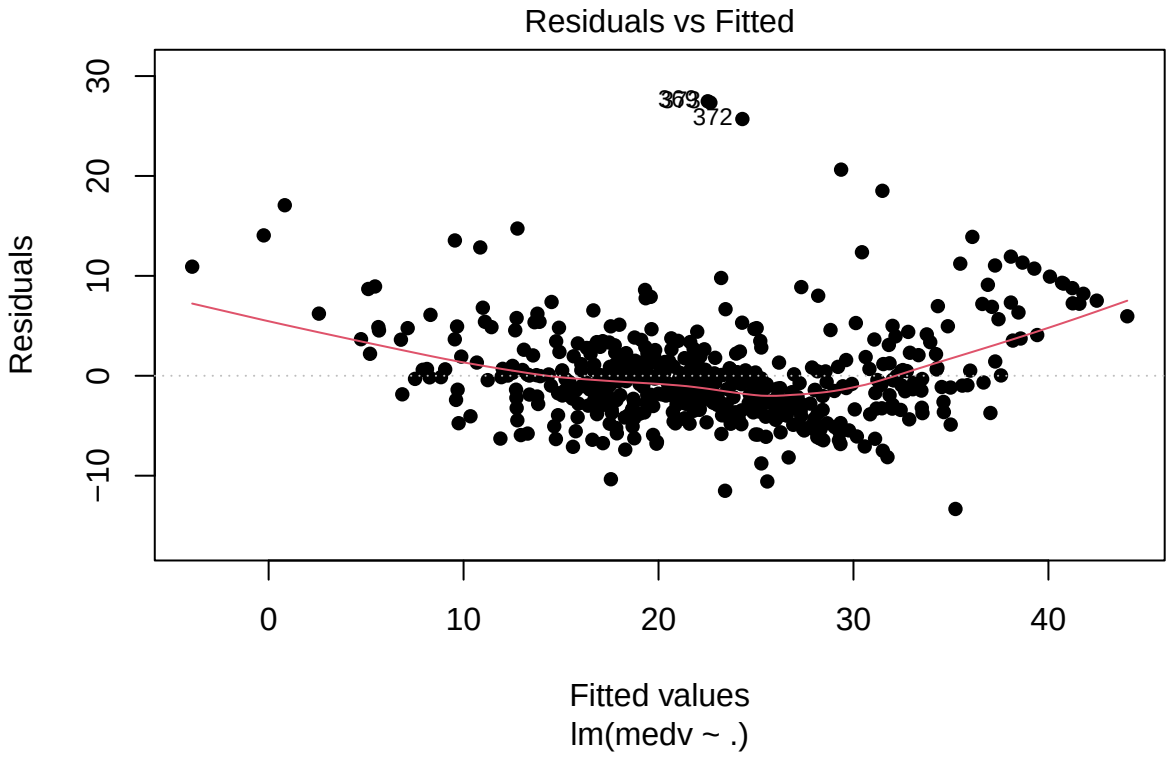
```
avPlots(model_fit)
```



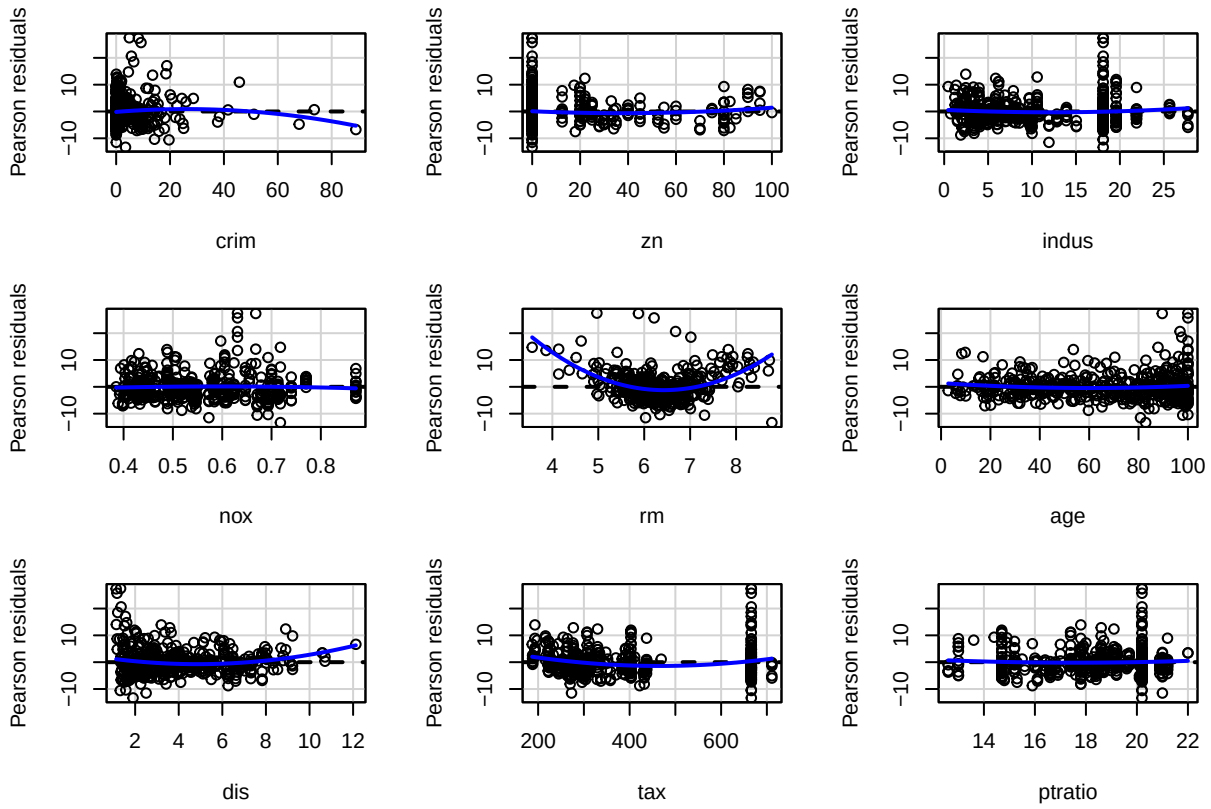
Inference: Since the mean is not centered around zero, we can see this from the first plot. Examining the residual vs. fitted values to determine the mean for each variables We see a curve in the graph even though the residuals' mean is generally heading toward zero. There is a curve in the mean line of the residual plot for the variables crim, dis,tax,black, and lstat.

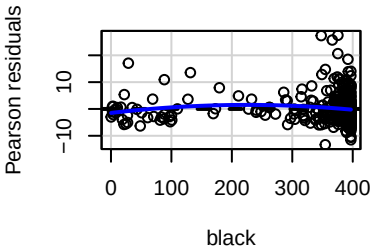
2. Variance

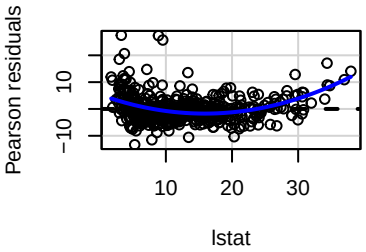
```
plot(model_fit ,which=1,pch=16)
```

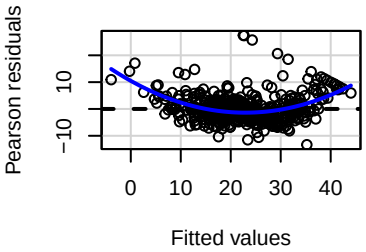



```
residualPlots(model_fit)
```



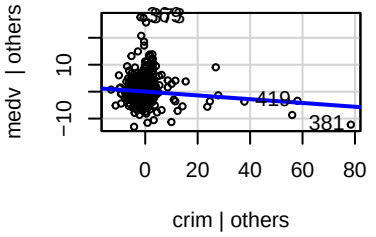


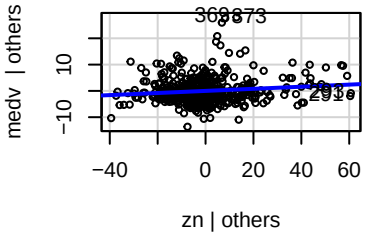


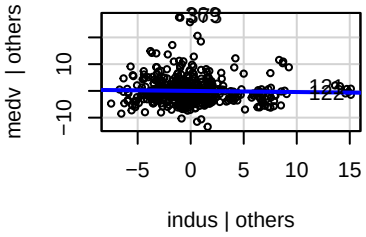


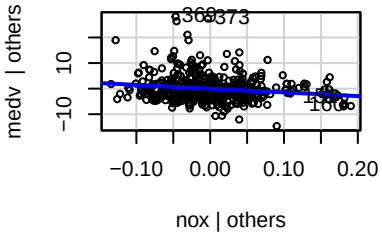

```
##             Test stat Pr(>|Test stat|)
## crim          -2.4854          0.01327 *
## zn             1.4242          0.15503
## indus          1.2836          0.19989
## nox           -1.0419          0.29799
## rm            12.6039        < 2.2e-16 ***
## age            1.7050          0.08883 .
## dis            4.3938        1.365e-05 ***
## tax            4.0370        6.275e-05 ***
## ptratio        0.9109          0.36282
## black         -2.1834          0.02948 *
## lstat         10.5122        < 2.2e-16 ***
## Tukey test    14.0829        < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

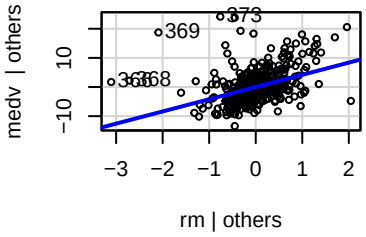
```
avPlots(model_fit) #Taking AV plot to take into consideration the added error by oth
```

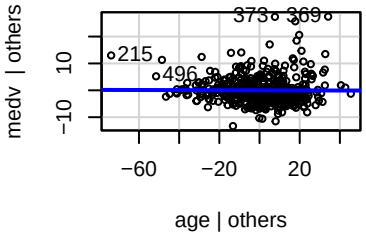


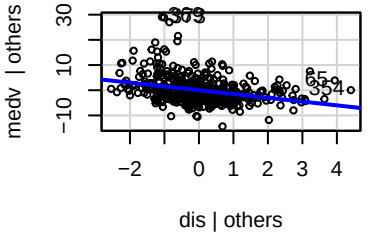


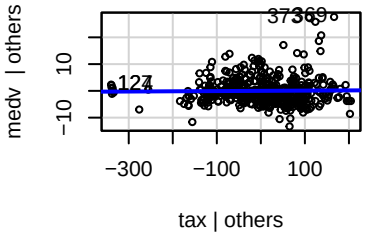


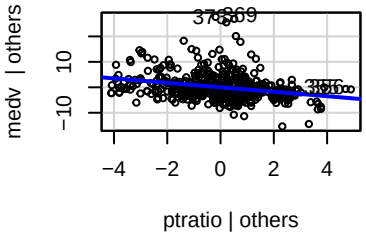


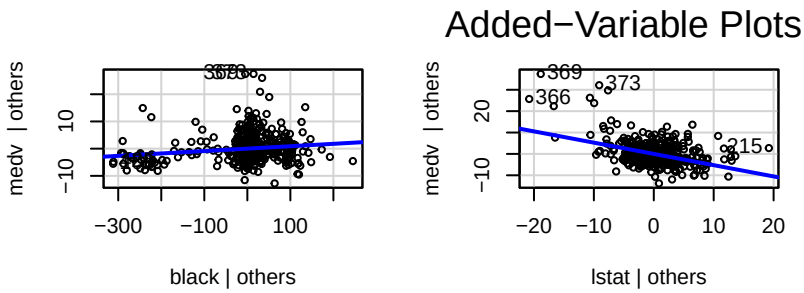








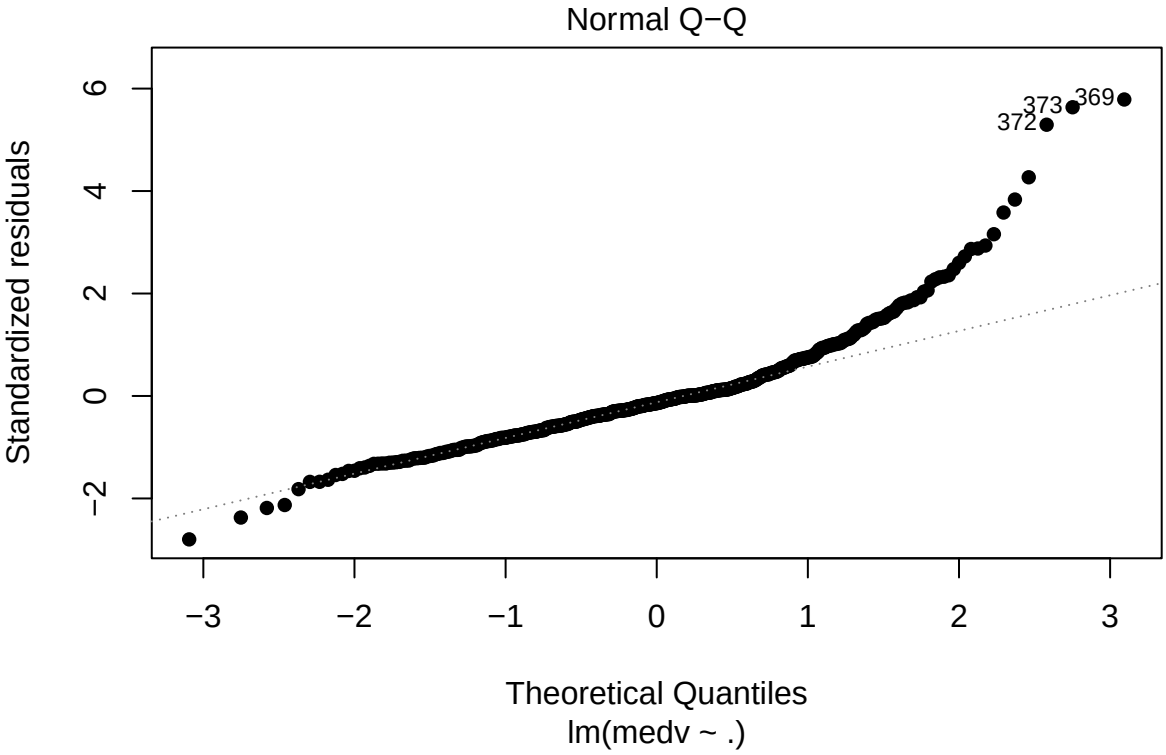




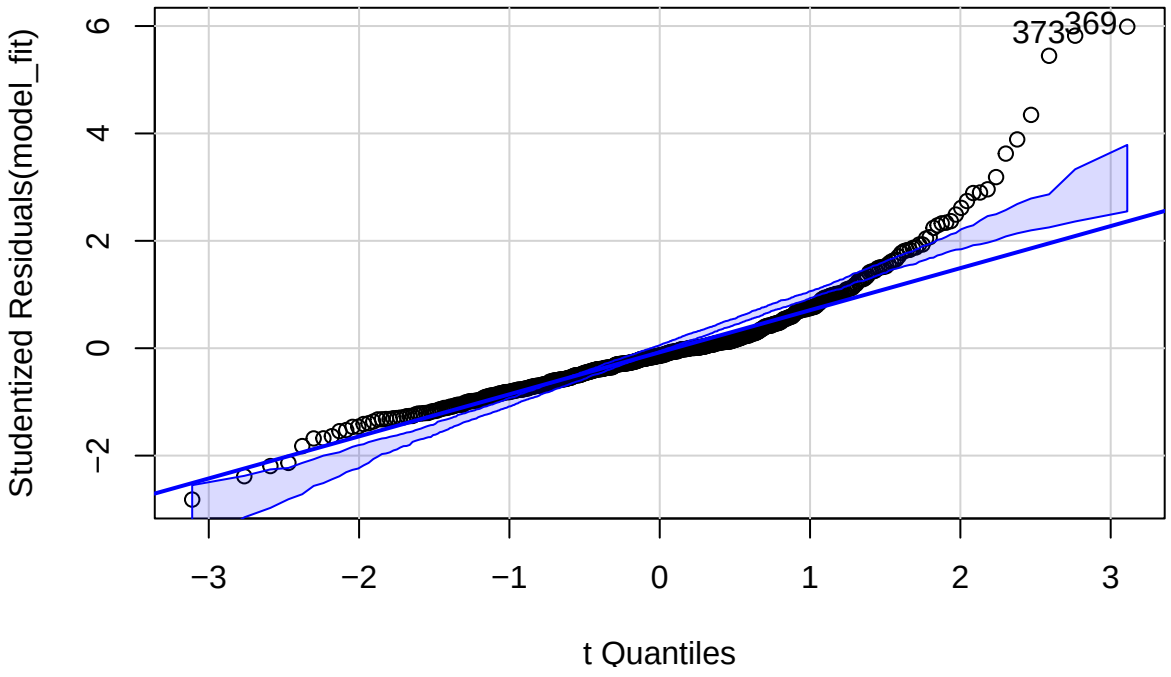
Inference: We notice that the distribution of the points is not equal; variance is not. We see a small fan shape in age, crim, zn, industries, dis, tax, and statistic data. Almost all of the characteristics' variance varies; it is not constant;

3. Normality

```
plot(model_fit,which=2,cex=1,pch=16)
```



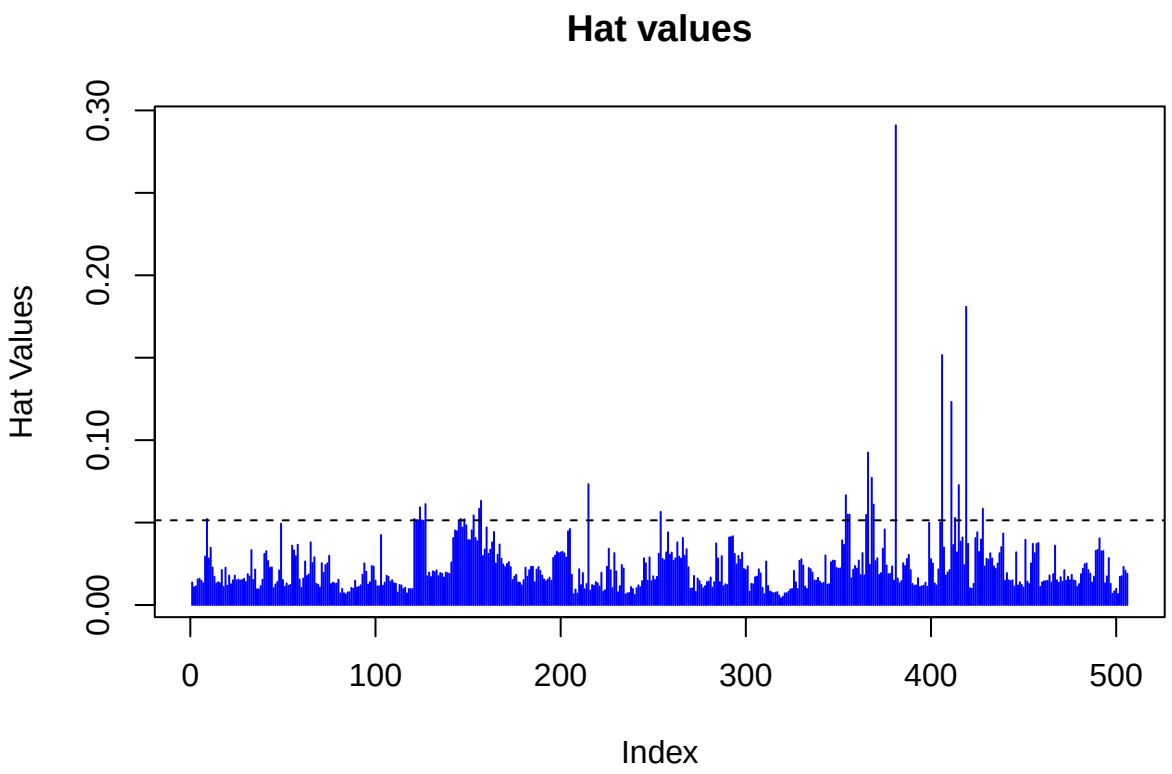

```
qqPlot(model_fit)
```




```
## [1] 369 373
```

It does not seem normal because the extreme values don't fit into the confidence band. All the standard assumptions are not followed, that is mean is not 0, variance is not homoscedatic and no normality/ #### Question 2 b) Outliers in predictor:

```
plot(hatvalues(model_fit),type='h', col="blue", ylab="Hat Values", main="Hat values")
p=12 ; n=506
abline(h = 2*(p+1)/n, lty=2)
```




```
sort(hatvalues(model_fit),decreasing = TRUE)[1]
```

```
##      381
## 0.290944
```

Inference : 381 is the most significant outlier in the predictor

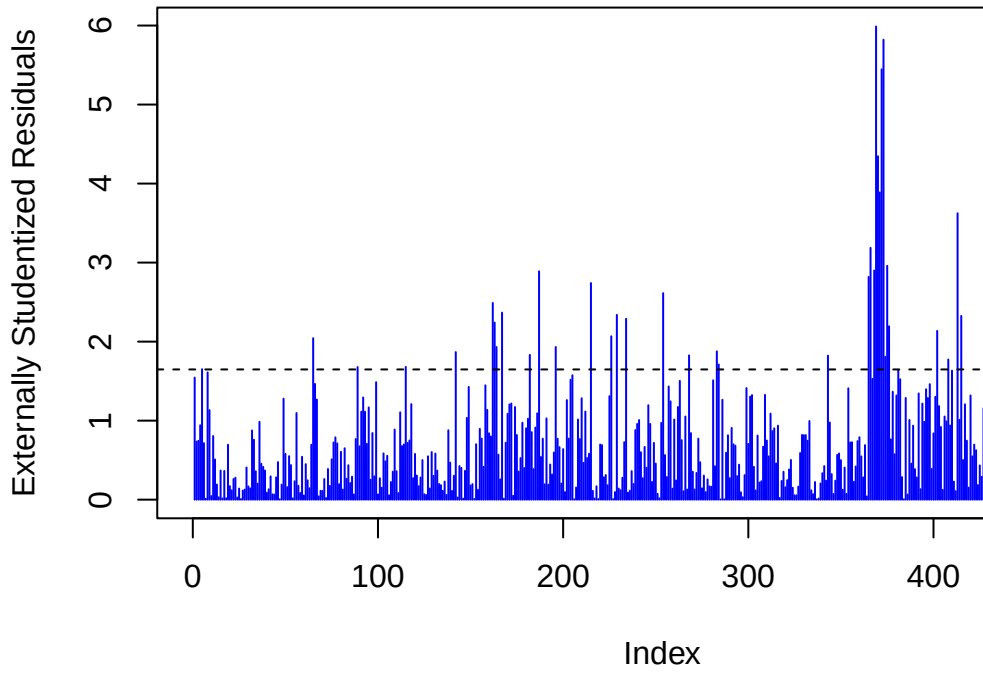
```
summary(data_set_final)
```

```
##       crim                zn              indus            nox
##  Min.   : 0.00632   Min.   :  0.00   Min.   : 0.46   Min.   :0.3850
##  1st Qu.: 0.08205   1st Qu.:  0.00   1st Qu.: 5.19   1st Qu.:0.4490
##  Median : 0.25651   Median :  0.00   Median : 9.69   Median :0.5380
##  Mean   : 3.61352   Mean   : 11.36   Mean   :11.14   Mean   :0.5547
##  3rd Qu.: 3.67708   3rd Qu.: 12.50   3rd Qu.:18.10   3rd Qu.:0.6240
##  Max.   :88.97620   Max.   :100.00   Max.   :27.74   Max.   :0.8710
##        rm              age              dis              tax
##  Min.   :3.561   Min.   :  2.90   Min.   : 1.130   Min.   :187.0
##  1st Qu.:5.886   1st Qu.: 45.02   1st Qu.: 2.100   1st Qu.:279.0
##  Median :6.208   Median : 77.50   Median : 3.207   Median :330.0
##  Mean   :6.285   Mean   : 68.57   Mean   : 3.795   Mean   :408.2
##  3rd Qu.:6.623   3rd Qu.: 94.08   3rd Qu.: 5.188   3rd Qu.:666.0
##  Max.   :8.780   Max.   :100.00   Max.   :12.127   Max.   :711.0
##     ptratio          black            lstat            medv
##  Min.   :12.60   Min.   :  0.32   Min.   : 1.73   Min.   : 5.00
##  1st Qu.:17.40   1st Qu.:375.38   1st Qu.: 6.95   1st Qu.:17.02
##  Median :19.05   Median :391.44   Median :11.36   Median :21.20
##  Mean   :18.46   Mean   :356.67   Mean   :12.65   Mean   :22.53
```

```
##  3rd Qu.:20.20   3rd Qu.:396.23   3rd Qu.:16.95   3rd Qu.:25.00
##  Max.   :22.00   Max.   :396.90   Max.   :37.97   Max.   :50.00
```

The crim,zn,black values for this data point take up extreme values that causes it to become an outlier.

```
plot(abs(rstudent(model_fit)), type='h', col="blue", ylab="Externally Studentized Residuals")
abline(h = qt(.95, n-p-2), lty=2) # threshold for suspects
```



Question 2 c) Outliers in response

```
sort(rstudent(model_fit),decreasing = TRUE)[1]
```

```
##      369
## 5.988068
```

most significant is 369

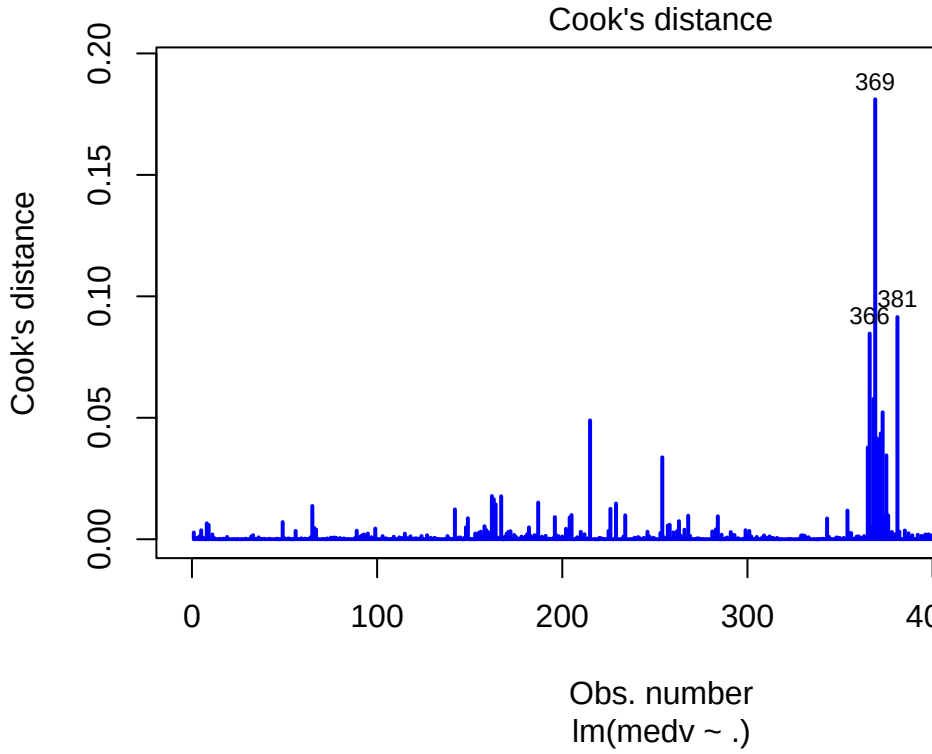
```
summary(data_set_final)
```

```
##       crim                zn              indus            nox
##  Min.   : 0.00632   Min.   :  0.00   Min.   : 0.46   Min.   :0.3850
##  1st Qu.: 0.08205   1st Qu.:  0.00   1st Qu.: 5.19   1st Qu.:0.4490
##  Median : 0.25651   Median :  0.00   Median : 9.69   Median :0.5380
##  Mean   : 3.61352   Mean   : 11.36   Mean   :11.14   Mean   :0.5547
##  3rd Qu.: 3.67708   3rd Qu.: 12.50   3rd Qu.:18.10   3rd Qu.:0.6240
##  Max.   :88.97620   Max.   :100.00   Max.   :27.74   Max.   :0.8710
##        rm              age              dis              tax
##  Min.   :3.561   Min.   :  2.90   Min.   : 1.130   Min.   :187.0
##  1st Qu.:5.886   1st Qu.: 45.02   1st Qu.: 2.100   1st Qu.:279.0
```

```
##  Median :6.208   Median : 77.50   Median : 3.207   Median :330.0
##  Mean   :6.285   Mean   : 68.57   Mean   : 3.795   Mean   :408.2
##  3rd Qu.:6.623   3rd Qu.: 94.08   3rd Qu.: 5.188   3rd Qu.:666.0
##  Max.   :8.780   Max.   :100.00   Max.   :12.127   Max.   :711.0
##     ptratio          black            lstat            medv
##  Min.   :12.60   Min.   :  0.32   Min.   : 1.73   Min.   : 5.00
##  1st Qu.:17.40   1st Qu.:375.38   1st Qu.: 6.95   1st Qu.:17.02
##  Median :19.05   Median :391.44   Median :11.36   Median :21.20
##  Mean   :18.46   Mean   :356.67   Mean   :12.65   Mean   :22.53
##  3rd Qu.:20.20   3rd Qu.:396.23   3rd Qu.:16.95   3rd Qu.:25.00
##  Max.   :22.00   Max.   :396.90   Max.   :37.97   Max.   :50.00
```

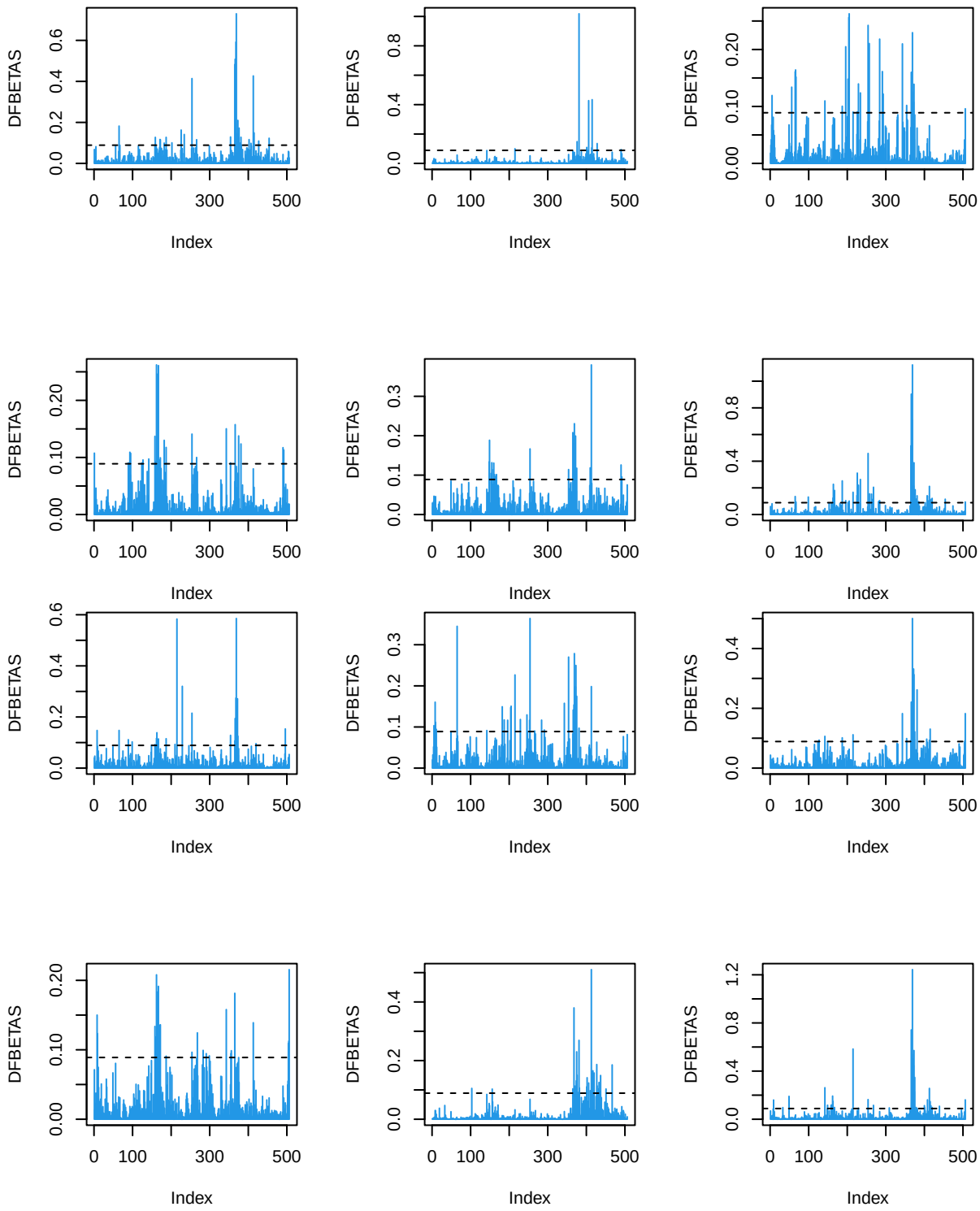
The zn,age values for this data point take up extreme values that causes it to become an outlier.

```
# Cook's distances
plot(model_fit, which=4, col="blue", lwd=2)
abline(h = 1, lty=2) # threshold for suspects (not visible on this plot)
```

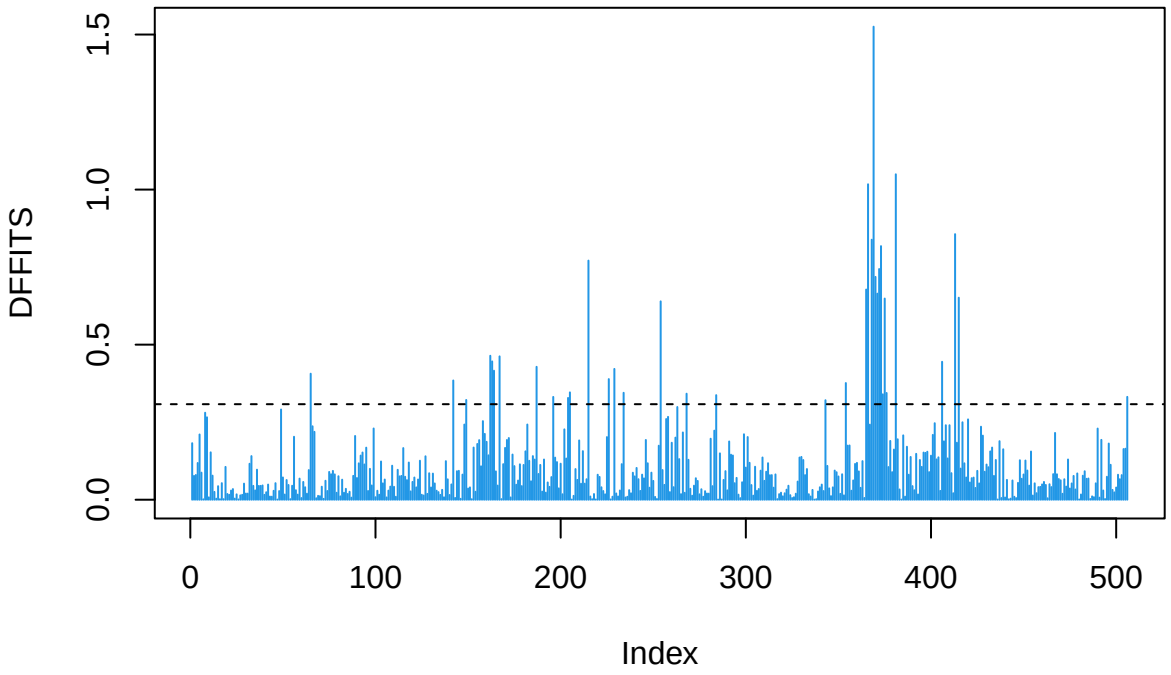


Question 2 d) Influential observations

```
# DFBETAS
par(mfrow=c(2,3))
for (j in 1:12){
plot(abs(dfbetas(model_fit)[,j]), col=4, type='h', ylab='DFBETAS')
abline(h = 2/sqrt(n), lty=2) # threshold for suspects
}
```




```
# DFFITS
par(mfrow=c(1,1))
plot(abs(dffits(model_fit)), typ='h', col=4, ylab='DFFITS')
abline(h = 2*sqrt(p/n), lty=2)
```




```
require(ellipse)
```

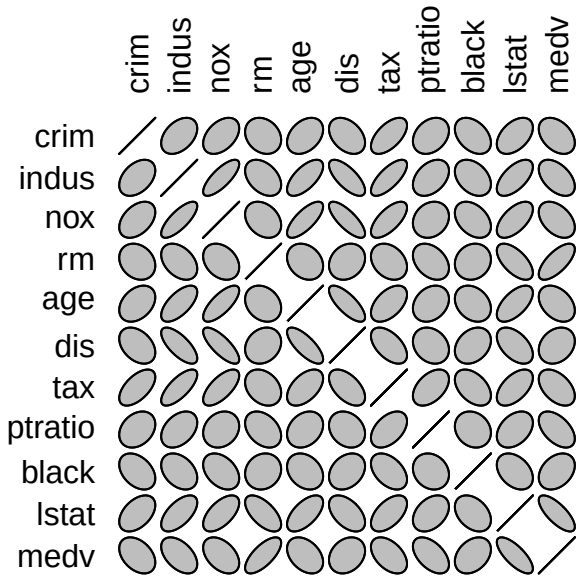
```
## Loading required package: ellipse
```

```
##
## Attaching package: 'ellipse'
```

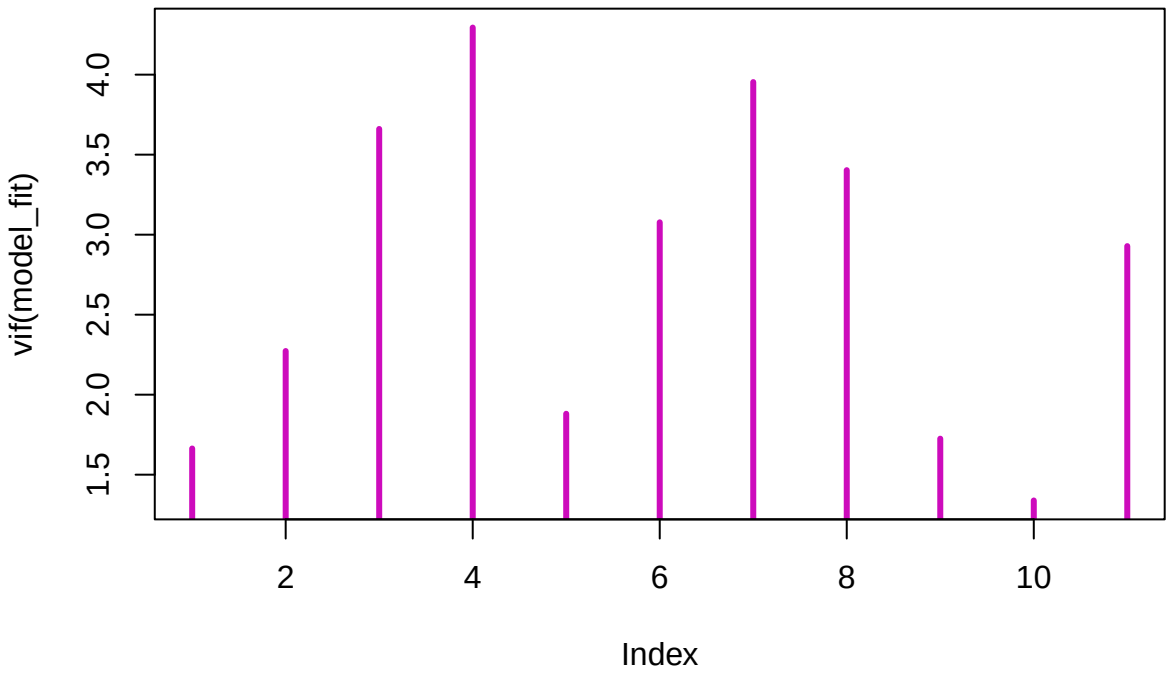
```
## The following object is masked from 'package:car':
##
##     ellipse
```

```
## The following object is masked from 'package:graphics':
##
##     pairs
```

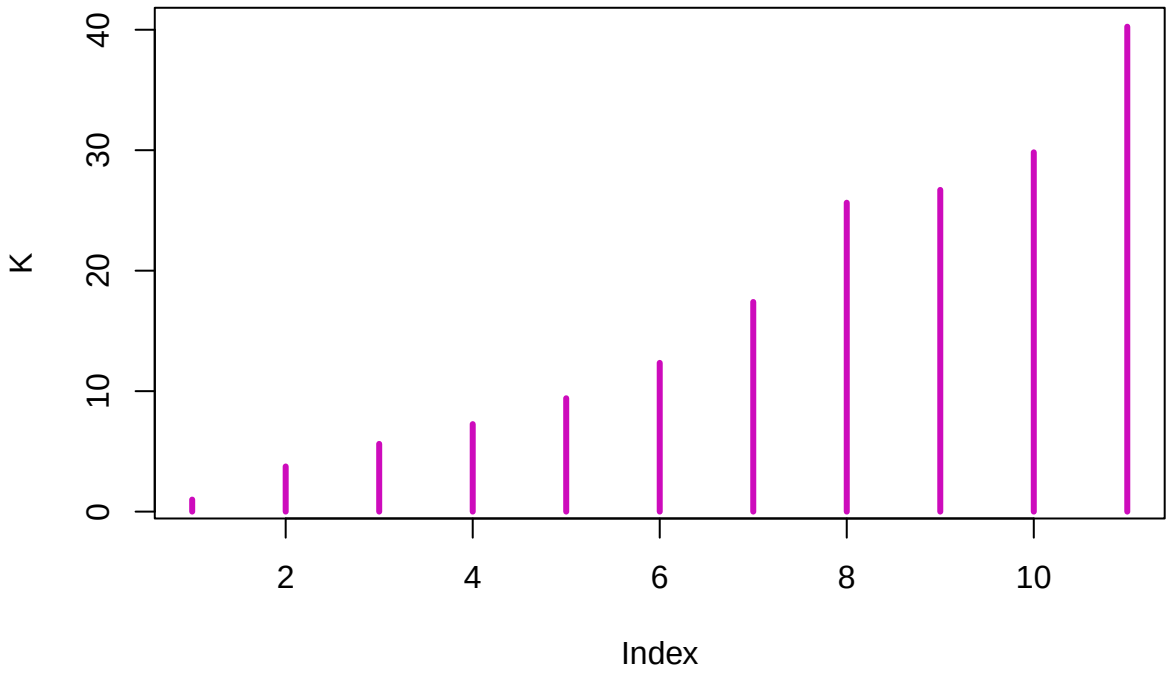
```
plotcorr(cor(data_set_final[, -2]))
```




```
plot(vif(model_fit), type='h', col=6, lwd=3)
abline(h = 10, lty=2)
```




```
C = cor(data_set_final[, -2]) # correlation matrix for the predictors
L = eigen(C) # eigenvalues
K = max(L$val)/L$val # condition indices
plot(K, type='h', col=6, lwd=3)
abline(h = 1000, lty=2)
```



A high VIF indicates stronger correlation. None of the features are highly correlated for this model. Contributions: Both the team members Sourabh Prakash and Priyanshi Shah have contributed equally to the homework by discussing the key points and logic together and doing pair programming. For the implementation part question 1 was contributed by Sourabh Prakash and question 2 by Priyanshi Shah.