New York City Stop-Question-Frisk Analysis

Jacky Lee, Tyler Sam, Sam Tan, Julianne Lin

The report should have the following structure:

1. Introduction (overview in your own words of the data; summary of your team's chosen question, why it's important, and what you found; roadmap for the remainder of the report.)
2. Responses to required questions: Please list each question/answer separately, but write formal responses to the questions.
3. Analysis of your chosen questions: Be sure to describe your questions and why they are interesting/important, and provide suitable graphical and statistical analysis of the data to answer the questions.
4. Conclusions and Recommendations
5. Bibliography if you use any external sources
6. Appendices as necessary

Your report will be evaluated based on the following criteria: - Writing: Structure of report; clarity of writing; quality of editing and polishing the document

- Correctness of methodology: Appropriate graphical and statistical techniques are used to answer the questions
- Validity of interpretation: The results are interpreted correctly
- Significance: The team-chosen question(s) leverage the available data to provide insight into the stop-question-frisk policy.

```
#1.
testedHispanic = length(sqf2010$race[sqf2010$race=='WHITE-HISPANIC'])
testedAsian = length(sqf2010$race[sqf2010$race=='ASIAN/PACIFIC ISLANDER'])
trueHispanic = testedHispanic*19/18 - testedAsian/18
trueAsian = testedAsian*19/18 - testedHispanic/18
total = testedHispanic + testedAsian

probHispanic = trueHispanic / total
probAsian = trueAsian / total
probActualHispanic = 0.95*trueHispanic / testedHispanic
probActualAsian = 0.95*trueAsian / testedAsian
```

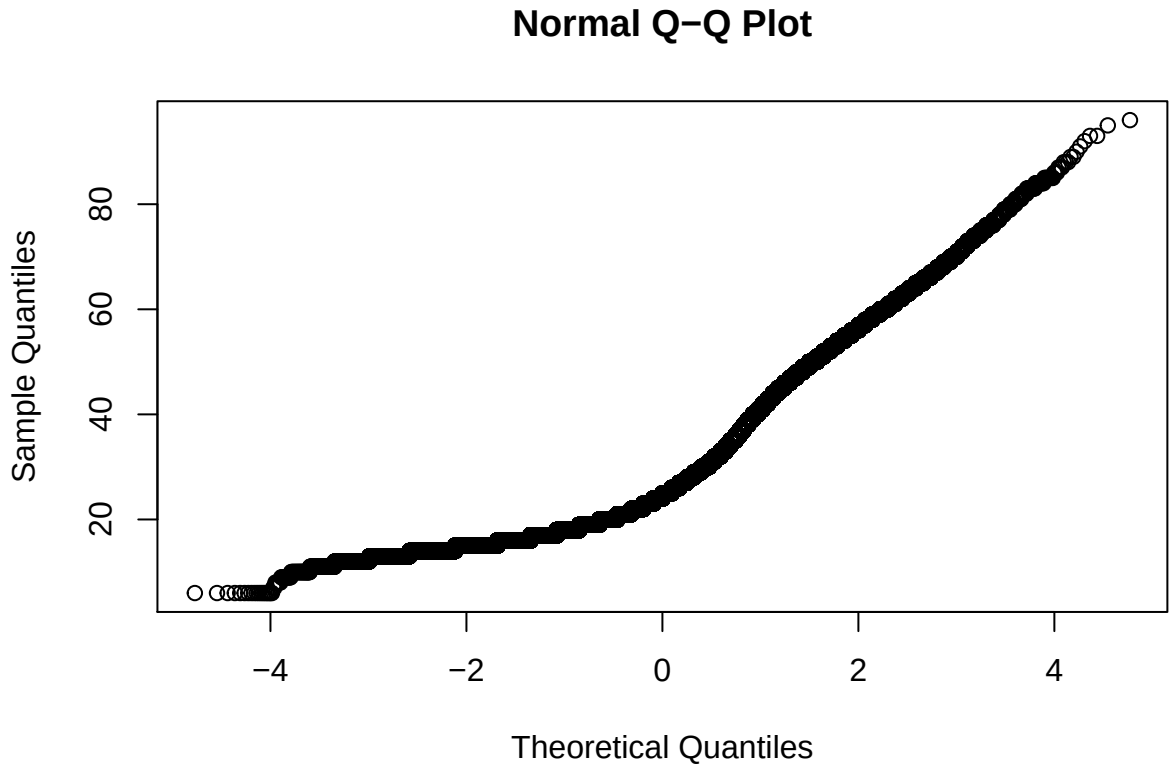
```
#2.
friskedWhite <-sqf2010$frisked[sqf2010$race =="WHITE"]
friskedBlack <-sqf2010$frisked[sqf2010$race =="BLACK"]
friskedWhiteHispanic <-sqf2010$frisked[sqf2010$race =="WHITE-HISPANIC"]
friskedBlackHispanic <-sqf2010$frisked[sqf2010$race =="BLACK-HISPANIC"]
friskedAsianPI <-sqf2010$frisked[sqf2010$race =="ASIAN/PACIFIC ISLANDER"]
friskedNatAmer <- sqf2010$frisked[sqf2010$race == "AMERICAN INDIAN/ALASKAN NATIVE"]

t.test(x=friskedAsianPI, y = friskedNatAmer, conf.level = .95)
```

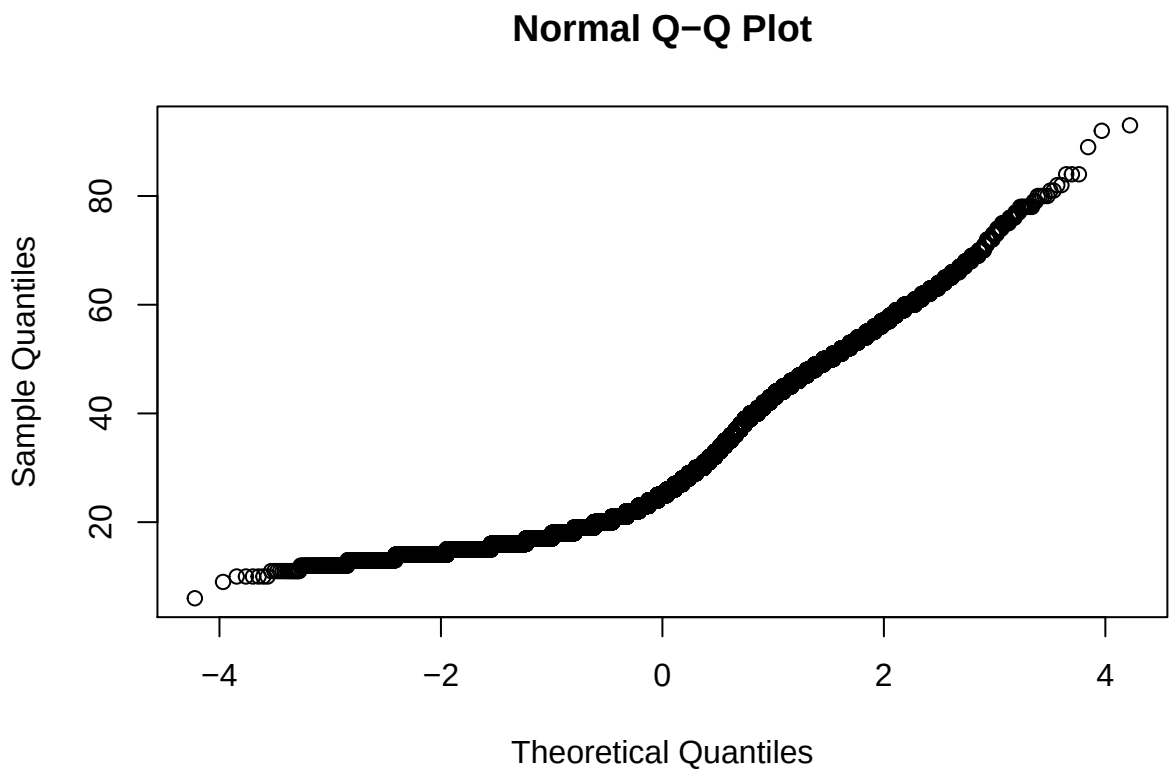
```
##
##  Welch Two Sample t-test
##
## data:  friskedAsianPI and friskedNatAmer
```

```
## t = -0.73591, df = 3286.7, p-value = 0.4618
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.02817925  0.01279882
## sample estimates:
## mean of x mean of y
## 0.4614839 0.4691741
```

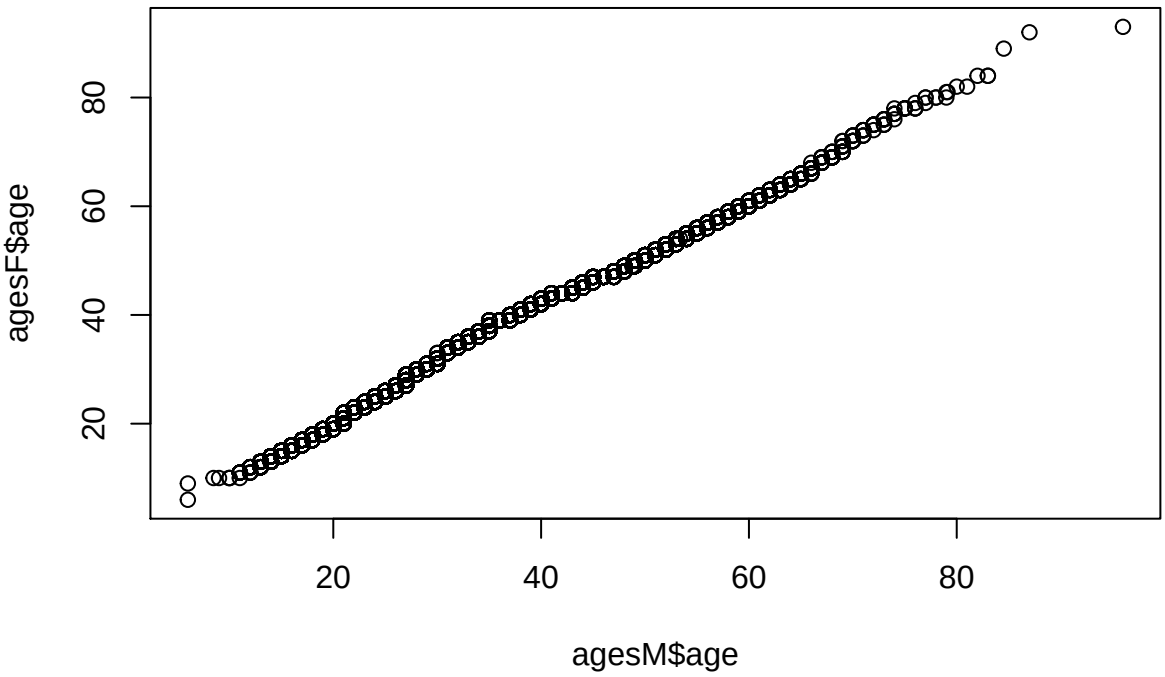
```
#3.
agesM = subset(sqf2010, sqf2010$age>5&sqf2010$age<98&sqf2010$sex =="M")
agesF = subset(sqf2010, sqf2010$age>5&sqf2010$age<98&sqf2010$sex =="F")
qqnorm(agesM$age)
```




```
qqnorm(agesF$age)
```

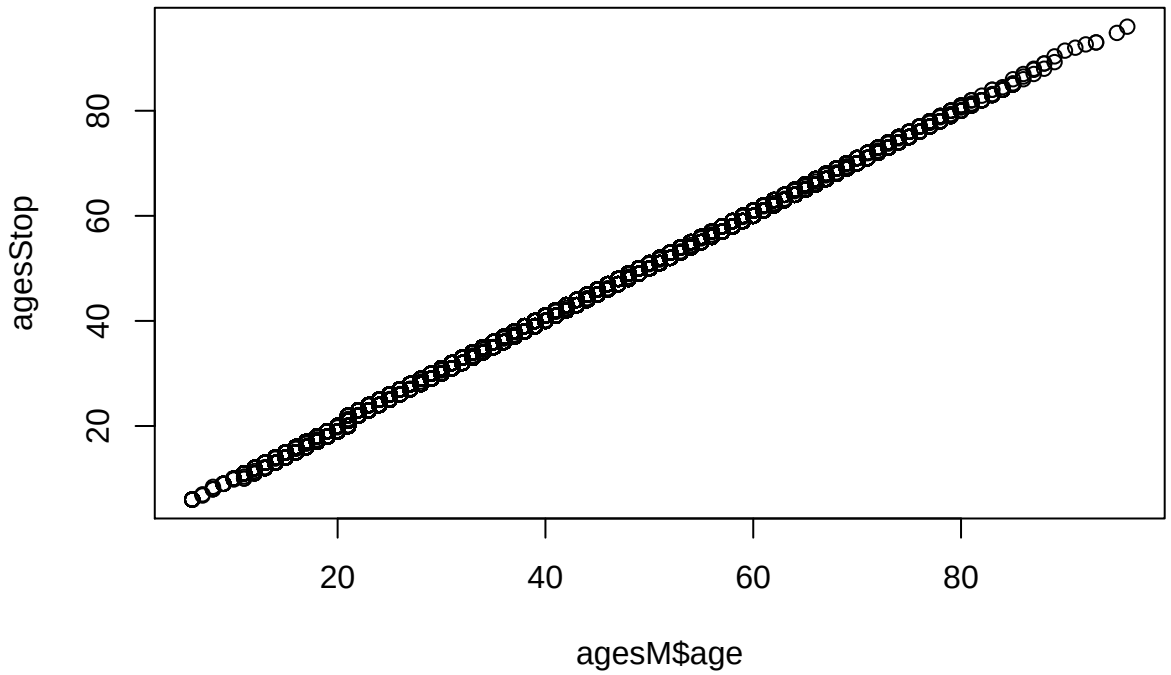



```
# both subsets of the age appear normal
qqplot(agesM$age, agesF$age)
```




```
# since the qqplot is a line, the distributions are the same.

agesStop = sqf2010$age[sqf2010$age<98&sqf2010$age>5]
qqplot(agesM$age, agesStop)
```




```
# since the qqplot is a line, the distributions are the same.
```

```
#4.
sample_size = 22563
# the sample size of the 2015 data set is 22563
sample_mean <- mean(sqf2015$frisked)
z_critical <- qnorm(0.975)
pop_stdev <- sd(sqf2015$frisked)
margin_of_error <- z_critical * (pop_stdev / sqrt(sample_size))
confidence_interval <- c(sample_mean - margin_of_error,
                         sample_mean + margin_of_error)
print("Confidence interval:")
```

```
## [1] "Confidence interval:"
```

```
confidence_interval
```

```
## [1] 0.6700898 0.6823013
```

```
# Confidence Interval for 2015 rates for being frisked(0.6700898 0.6823013)

summary(sqf2015)
```

```
##     datestop          timestop    sex
##  Min.   : 1012015   Min.   :   0   F: 1515
##  1st Qu.: 3112015   1st Qu.: 505   M:20853
##  Median : 5282015   Median :1623   Z:  195
##  Mean   : 5957353   Mean   :1377
##  3rd Qu.: 9012015   3rd Qu.:2052
##  Max.   :12312015   Max.   :2359
##
##                      race            age             weight
##  BLACK                 :11950   Min.   :  0.00   Min.   :  1.0
##  WHITE-HISPANIC        : 5090   1st Qu.: 19.00   1st Qu.:150.0
##  WHITE                 : 2514   Median : 24.00   Median :170.0
```

```
##  BLACK-HISPANIC        : 1409   Mean   : 28.96   Mean   :171.4
##  ASIAN/PACIFIC ISLANDER: 1103   3rd Qu.: 33.00   3rd Qu.:185.0
##  OTHER                 :  298   Max.   :999.00   Max.   :999.0
##  (Other)               :  199
##     haircolr        eyecolor          build              city
##  BLACK  :16221   BROWN  :20023   HEAVY   : 2167   BRONX    :4754
##  BROWN  : 4742   BLACK  : 1458   MEDIUM  :10986   BROOKLYN :6354
##  BALD   :  549   BLUE   :  419   MUSCULAR:  194   MANHATTAN:3941
##  BLOND  :  320   HAZEL  :  216   THIN    : 8880   QUEENS   :5718
##  UNKNOWN:  240   GREEN  :  162   UNKNOWN :  336   STATEN IS:1796
##  GRAY   :  183   UNKNOWN:  156
##  (Other):  308   (Other):  129
##      inside                    location        typeofid
##  Min.   :0.0000   HOUSING AUTHORITY: 3371   OTHER  :  433
##  1st Qu.:0.0000   NEITHER          :18291   PHOTO  :12978
##  Median :0.0000   TRANSIT AUTHORITY:  901   REFUSED:  639
##  Mean   :0.1872                             VERBAL : 8513
##  3rd Qu.:0.0000
##  Max.   :1.0000
##
##      perobs           perstop        arstmade         sumissue
##  Min.   :  0.000   5      :8003   Min.   :0.0000   Min.   :0.00000
##  1st Qu.:  1.000   10     :4792   1st Qu.:0.0000   1st Qu.:0.00000
##  Median :  1.000   3      :1638   Median :0.0000   Median :0.00000
##  Mean   :  2.639   15     :1518   Mean   :0.1759   Mean   :0.02606
##  3rd Qu.:  2.000   2      :1505   3rd Qu.:0.0000   3rd Qu.:0.00000
##  Max.   :535.000   1      : 925   Max.   :1.0000   Max.   :1.00000
##                    (Other):4182
##     frisked          searched         contrabn           radio
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.00000   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:0.0000
##  Median :1.0000   Median :0.0000   Median :0.00000   Median :0.0000
##  Mean   :0.6762   Mean   :0.1863   Mean   :0.04986   Mean   :0.4161
##  3rd Qu.:1.0000   3rd Qu.:0.0000   3rd Qu.:0.00000   3rd Qu.:1.0000
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.00000   Max.   :1.0000
##
##      height            pf              weap
##  Min.   :36.00   Min.   :0.0000   Min.   :0.00000
##  1st Qu.:67.00   1st Qu.:0.0000   1st Qu.:0.00000
##  Median :69.00   Median :0.0000   Median :0.00000
##  Mean   :68.83   Mean   :0.3317   Mean   :0.04822
##  3rd Qu.:71.00   3rd Qu.:1.0000   3rd Qu.:0.00000
##  Max.   :95.00   Max.   :1.0000   Max.   :1.00000
##
```

```r
sample_size = 639
# the sample size of the 2015 data set is 22563
sample_noId = subset(sqf2015, sqf2015$typeofid =="REFUSED" )
sample_mean <- mean(sample_noId$frisked)
z_critical <- qnorm(0.975)
pop_stdev <- sd(sample_noId$frisked)
margin_of_error <- z_critical * (pop_stdev / sqrt(sample_size))
confidence_interval <- c(sample_mean - margin_of_error,
                         sample_mean + margin_of_error)
```

```r
print("Confidence interval:")
```

```
## [1] "Confidence interval:"
```

```r
confidence_interval
```

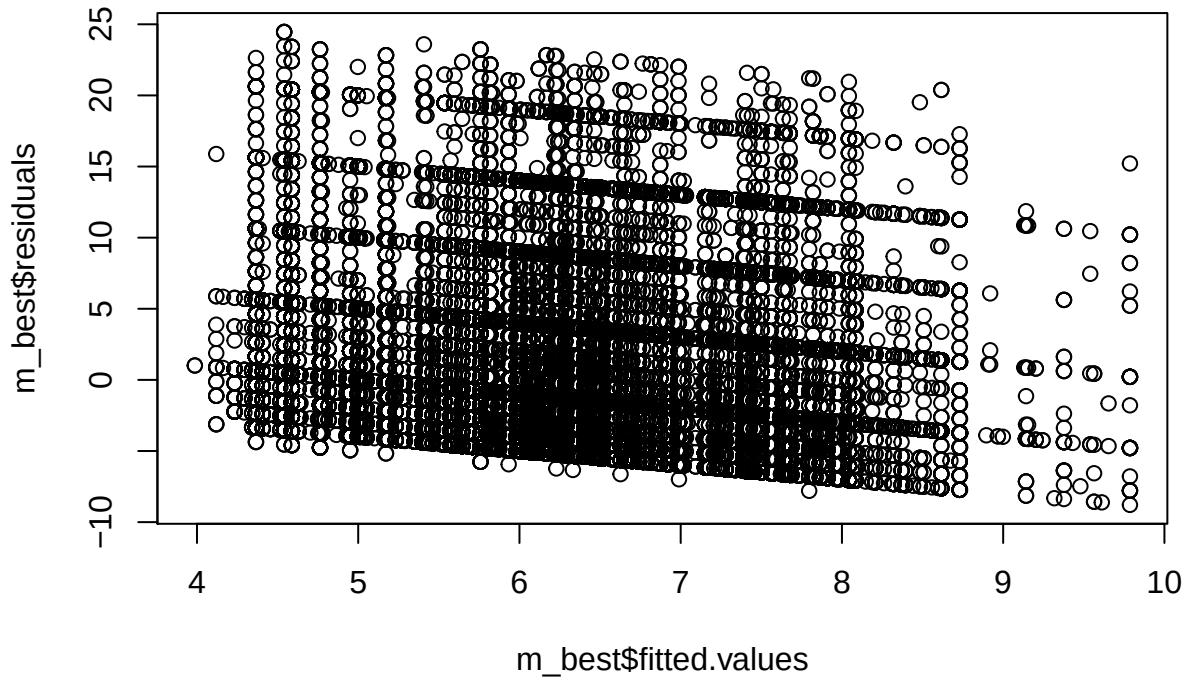
```
## [1] 0.5724871 0.6481702
```

```r
# Confidence Interval for 2015 rates for being frisked given that
# one refused to show ID(0.5724871 0.6481702)
```

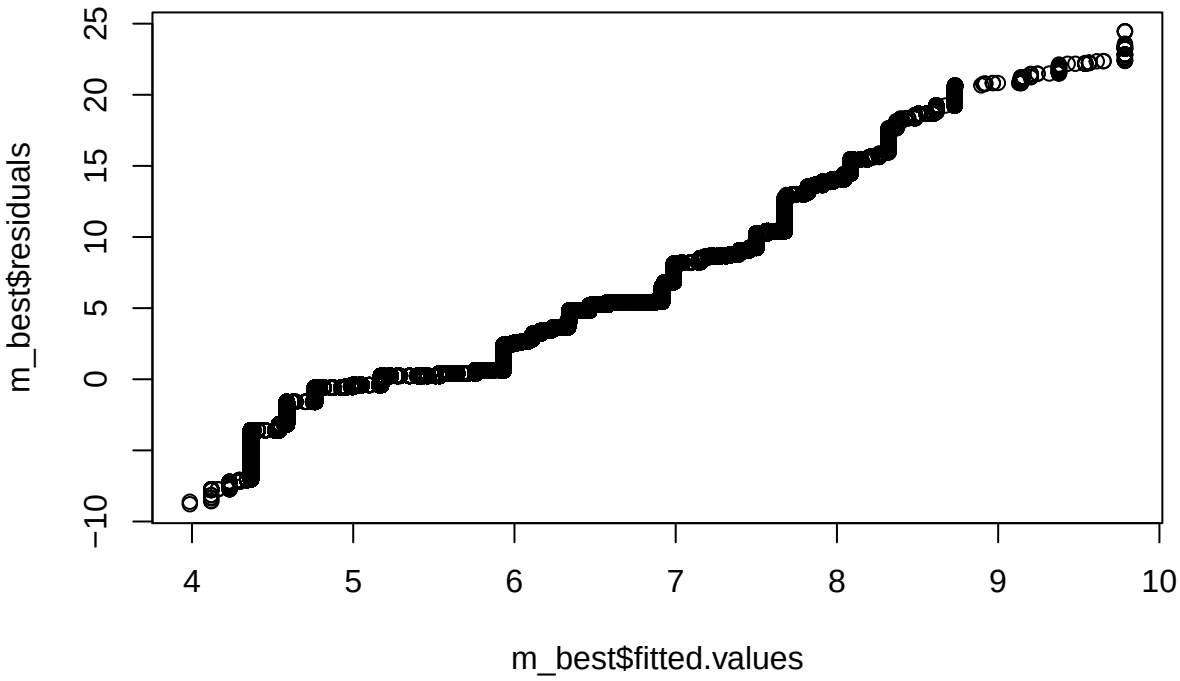
```r
#5.
m_best = step(lm(formula = perstop ~ arstmade + searched + inside + sumissue + frisked + weap + contrabn
```

```
## Start:  AIC=1430692
## perstop ~ arstmade + searched + inside + sumissue + frisked +
##     weap + contrabn + radio + pf
##
##             Df Sum of Sq     RSS     AIC
## <none>                   6554616 1430692
## - weap       1       119 6554735 1430701
## - contrabn   1       547 6555163 1430740
## - frisked    1      3524 6558140 1431011
## - inside     1      4933 6559549 1431139
## - searched   1     14079 6568695 1431971
## - pf         1     14561 6569177 1432015
## - arstmade   1     26271 6580887 1433079
## - sumissue   1    117471 6672087 1441298
## - radio      1    151127 6705743 1444303
```

```r
plot(m_best$fitted.values, m_best$residuals)
```




```r
qqplot(m_best$fitted.values, m_best$residuals)
```




```
# m_full = lm(complied ~ pop + hdi + dem + internet + freepress, data = goog_sub)
# m_best = step(m_full)
# plot(m_best$fitted.values, m_best$residuals)
```