ECON403A PROJECT 3

Ziwen Gu, Yining Quan, Mohammed Ibraaz Syed, Minxuan Wang

December 11, 2017

GROUP MEMBERS: Ziwen Gu, Yining Quan, Mohammed Ibraaz Syed, Minxuan Wang

Setting working directory:

```
setwd("C:/Users/Ibraaz/Desktop/ECON403A HW FOLDER/")
```

Getting data

```
Pr3datafile <- "german_healthcare_usage.csv"
Pr3 <- read.csv(Pr3datafile)
```

The following are the names of the variables:

```
names(Pr3)
```

```
##  [1] "ID"       "FEMALE"   "YEAR"     "AGE"      "HANDDUM"  "ALC"
##  [7] "FAMHIST"  "HANDPER"  "HHKIDS"   "EDUC"     "MARRIED"  "HAUPTS"
## [13] "REALS"    "FACHHS"   "ABITUR"   "UNIV"     "WORKING"  "BLUEC"
## [19] "WHITEC"   "SELF"     "BEAMT"    "DOCVIS"   "HOSPVIS"  "UNEMPLOY"
## [25] "PUBLIC"   "ADDON"    "NUMOBS"   "HSAT"     "DOCTOR"   "HEALTHY"
## [31] "YEAR1984" "YEAR1985" "YEAR1986" "YEAR1987" "YEAR1988" "YEAR1991"
## [37] "YEAR1994" "LOGINC"   "TI"       "HOSPITAL" "HHNINC"   "NEWHSAT"
## [43] "PRESCRIP"
```

We first observe that the variable **NUMOBS** indicates how many observations there are for each person.

We also note that we are informed in the Data Description section to ignore the variables **TI** and **INCOME**. We observe that **INCOME** is not a variable in the provided dataset.

The Data Description section also notes that the variable **HSAT** has 40 coding errors and that the variable **NEWHSAT** fixes them.

I. Exploratory Analysis

Exploring the dataset

```
describe(Pr3, skew = FALSE)
```

```
##           vars     n    mean      sd     min     max   range    se
## ID           1 27326 3517.44 2014.63    1.00 7293.00 7292.00 12.19
## FEMALE       2 27325    0.48    0.50    0.00    1.00    1.00  0.00
## YEAR         3 27326 1987.82    3.17 1984.00 1994.00   10.00  0.02
## AGE          4 27324   43.52   11.33   25.00   64.00   39.00  0.07
## HANDDUM      5 27326    0.21    0.41    0.00    1.00    1.00  0.00
## ALC          6 27323   21.41   11.56    1.45   41.45   40.00  0.07
## FAMHIST      7 27325    0.50    0.50    0.00    5.00    5.00  0.00
## HANDPER      8 27326    7.01   19.26    0.00  100.00  100.00  0.12
## HHKIDS       9 27325    0.40    0.49    0.00    1.00    1.00  0.00
## EDUC        10 27325   11.32    2.32    7.00   18.00   11.00  0.01
## MARRIED     11 27325    0.76    0.43    0.00    1.00    1.00  0.00
## HAUPTS      12 27326    0.62    0.48    0.00    1.00    1.00  0.00
## REALS       13 27326    0.20    0.40    0.00    1.00    1.00  0.00
## FACHHS      14 27326    0.04    0.20    0.00    1.00    1.00  0.00
## ABITUR      15 27325    0.12    0.32    0.00    1.00    1.00  0.00
## UNIV        16 27326    0.07    0.26    0.00    1.00    1.00  0.00
## WORKING     17 27326    0.68    0.47    0.00    1.00    1.00  0.00
## BLUEC       18 27326    0.24    0.43    0.00    1.00    1.00  0.00
## WHITEC      19 27326    0.30    0.46    0.00    1.00    1.00  0.00
## SELF        20 27326    0.06    0.24    0.00    1.00    1.00  0.00
## BEAMT       21 27326    0.07    0.26    0.00    1.00    1.00  0.00
## DOCVIS      22 27326    3.18    5.69    0.00  121.00  121.00  0.03
## HOSPVIS     23 27326    0.14    0.88    0.00   51.00   51.00  0.01
## UNEMPLOY    24 27326    0.32    0.47    0.00    1.00    1.00  0.00
## PUBLIC      25 27326    0.89    0.32    0.00    1.00    1.00  0.00
## ADDON       26 27323    0.02    0.14    0.00    1.00    1.00  0.00
## NUMOBS      27 27326    4.90    1.79    1.00    7.00    6.00  0.01
## HSAT        28 27326    6.79    2.29    0.00   10.00   10.00  0.01
## DOCTOR      29 27325    0.63    0.48    0.00    1.00    1.00  0.00
## HEALTHY     30 27326    0.61    0.49    0.00    1.00    1.00  0.00
## YEAR1984    31 27326    0.14    0.35    0.00    1.00    1.00  0.00
## YEAR1985    32 27326    0.14    0.35    0.00    1.00    1.00  0.00
## YEAR1986    33 27326    0.14    0.35    0.00    1.00    1.00  0.00
## YEAR1987    34 27325    0.13    0.34    0.00    1.00    1.00  0.00
## YEAR1988    35 27322    0.16    0.37    0.00    1.00    1.00  0.00
## YEAR1991    36 27325    0.16    0.37    0.00    1.00    1.00  0.00
## YEAR1994    37 27326    0.12    0.33    0.00    1.00    1.00  0.00
## LOGINC      38 27321   -1.16    0.49   -6.50    1.12    7.62  0.00
## TI          39 27325    4.90    1.79    1.00    7.00    6.00  0.01
## HOSPITAL    40 27321    0.09    0.28    0.00    1.00    1.00  0.00
## HHNINC      41 27325    0.35    0.18    0.00    3.07    3.07  0.00
## NEWHSAT     42 27326    6.79    2.29    0.00   10.00   10.00  0.01
## PRESCRIP    43 27326    2.50    3.15    0.00   70.00   70.00  0.02
```

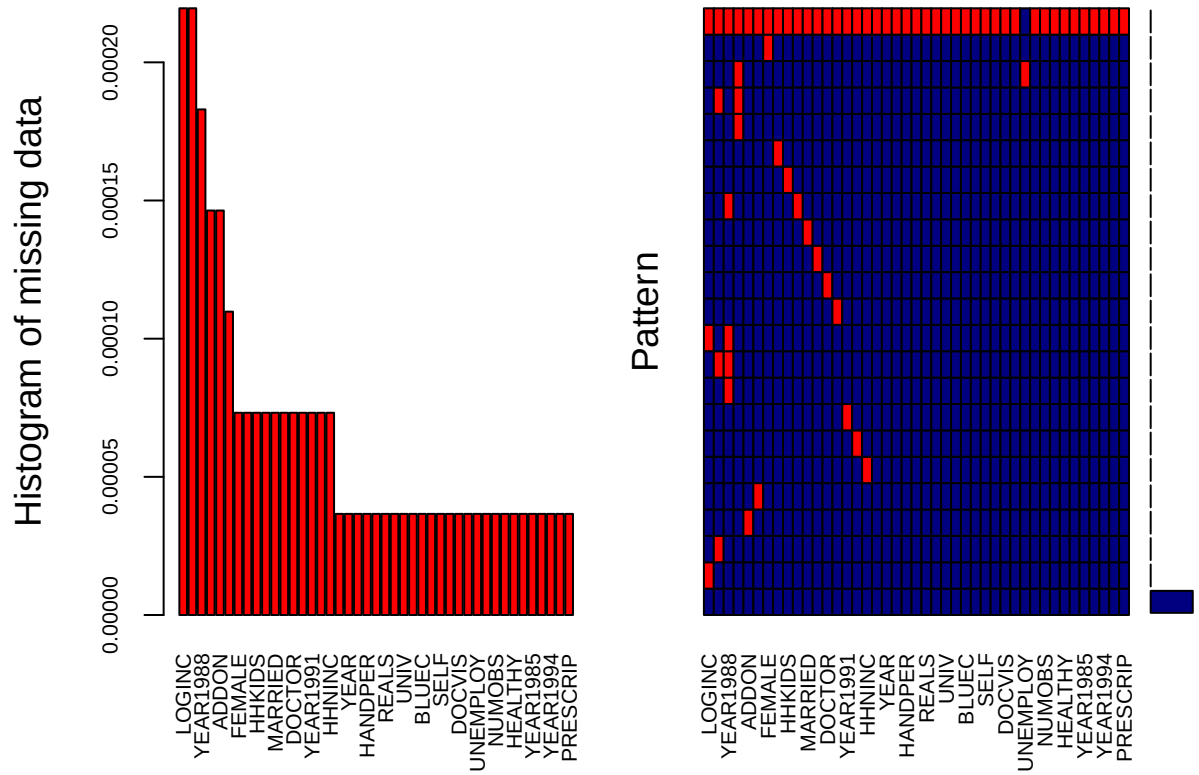
Looking for any missing values:

```
summary(Pr3)[7,]
```

```
##            ID        FEMALE          YEAR           AGE       HANDDUM
## "NA's   :1  " "NA's   :2  " "NA's   :1  " "NA's   :3  " "NA's   :1  "
##           ALC       FAMHIST       HANDPER        HHKIDS          EDUC
## "NA's   :4  " "NA's   :2  " "NA's   :1  " "NA's   :2  " "NA's   :2  "
##       MARRIED        HAUPTS         REALS        FACHHS        ABITUR
## "NA's   :2  " "NA's   :1  " "NA's   :1  " "NA's   :1  " "NA's   :2  "
##          UNIV       WORKING         BLUEC        WHITEC          SELF
## "NA's   :1  " "NA's   :1  " "NA's   :1  " "NA's   :1  " "NA's   :1  "
##         BEAMT        DOCVIS       HOSPVIS      UNEMPLOY        PUBLIC
## "NA's   :1  " "NA's   :1  " "NA's   :1  " "NA's   :1  " "NA's   :1  "
##         ADDON        NUMOBS          HSAT        DOCTOR       HEALTHY
## "NA's   :4  " "NA's   :1  " "NA's   :1  " "NA's   :2  " "NA's   :1  "
##      YEAR1984      YEAR1985      YEAR1986      YEAR1987      YEAR1988
## "NA's   :1  " "NA's   :1  " "NA's   :1  " "NA's   :2  " "NA's   :5  "
##      YEAR1991      YEAR1994        LOGINC            TI      HOSPITAL
## "NA's   :2  " "NA's   :1  " "NA's   :6  " "NA's   :2  " "NA's   :6  "
##        HHNINC       NEWHSAT      PRESCRIP
## "NA's   :2  " "NA's   :1  " "NA's   :1  "
```

We note the existence of missing values in our dataset. We explore the pervasiveness of missing values:

```
aggr_plot <- aggr(Pr3, col=c('navyblue','red'),
            numbers=TRUE, sortVars=TRUE, labels = names(Pr3),
            cex.axis=.7, gap=3, ylab=c("Histogram of missing data", "Pattern"))
```



```
##
##  Variables sorted by number of missings:
##  Variable        Count
##    LOGINC 2.195631e-04
##  HOSPITAL 2.195631e-04
##  YEAR1988 1.829692e-04
##       ALC 1.463754e-04
##     ADDON 1.463754e-04
##       AGE 1.097815e-04
##    FEMALE 7.318769e-05
##   FAMHIST 7.318769e-05
##    HHKIDS 7.318769e-05
##      EDUC 7.318769e-05
##   MARRIED 7.318769e-05
##    ABITUR 7.318769e-05
##    DOCTOR 7.318769e-05
##  YEAR1987 7.318769e-05
##  YEAR1991 7.318769e-05
##        TI 7.318769e-05
##    HHNINC 7.318769e-05
##        ID 3.659384e-05
##      YEAR 3.659384e-05
##   HANDDUM 3.659384e-05
##   HANDPER 3.659384e-05
##    HAUPTS 3.659384e-05
##     REALS 3.659384e-05
##    FACHHS 3.659384e-05
##      UNIV 3.659384e-05
##   WORKING 3.659384e-05
##     BLUEC 3.659384e-05
##    WHITEC 3.659384e-05
##      SELF 3.659384e-05
##     BEAMT 3.659384e-05
##    DOCVIS 3.659384e-05
##   HOSPVIS 3.659384e-05
##  UNEMPLOY 3.659384e-05
##    PUBLIC 3.659384e-05
##    NUMOBS 3.659384e-05
##      HSAT 3.659384e-05
##   HEALTHY 3.659384e-05
##  YEAR1984 3.659384e-05
##  YEAR1985 3.659384e-05
##  YEAR1986 3.659384e-05
##  YEAR1994 3.659384e-05
##   NEWHSAT 3.659384e-05
##  PRESCRIP 3.659384e-05
```

Looking at what percent of data has missing values:

```
NAvalues <- summary(aggr_plot)
def <- data.frame(NAvalues$combinations$Count, NAvalues$combinations$Percent)
```

Around 99.89% of the data does not have missing values so we make the decision to only work with the data that have no missing values.

Removing data with missing values from working dataset to form new dataet:

```
Pr3c <- Pr3[complete.cases(Pr3),]
```

Missing Values Original Dataset:

```
sum(is.na(Pr3))
```

```
## [1] 76
```

Missing Values in Dataset After Removing Missing Values:

```
sum(is.na(Pr3c))
```

```
## [1] 0
```

We confirm that there are no longer any missing values in the data and thus we can proceed.

Part I (a)

Note from page 2 that the quantitative variables are YEAR, AGE, ALC, HANDPER, EDUC, DOCVIS, HOSPVIS, NUMOBS, HSAT, LOGINC, HHNINC, NEWHSAT, and PRESCRIP.

Note that we will include **YEAR**, **NUMOBS**, and **ID** as they do happen to quantitative variables.

Note also that we will *not include* the variable **HSAT** as the Data Description tells us that **NEWHSAT** is the same as **HSAT** - just with the 40 coding errors fixed.

We will proceed by first plotting the following variables:

- AGE
- ALC
- HANDPER
- EDUC
- DOCVIS
- HOSPVIS
- LOGINC
- HHNINC
- NEWHSAT
- PRESCRIP

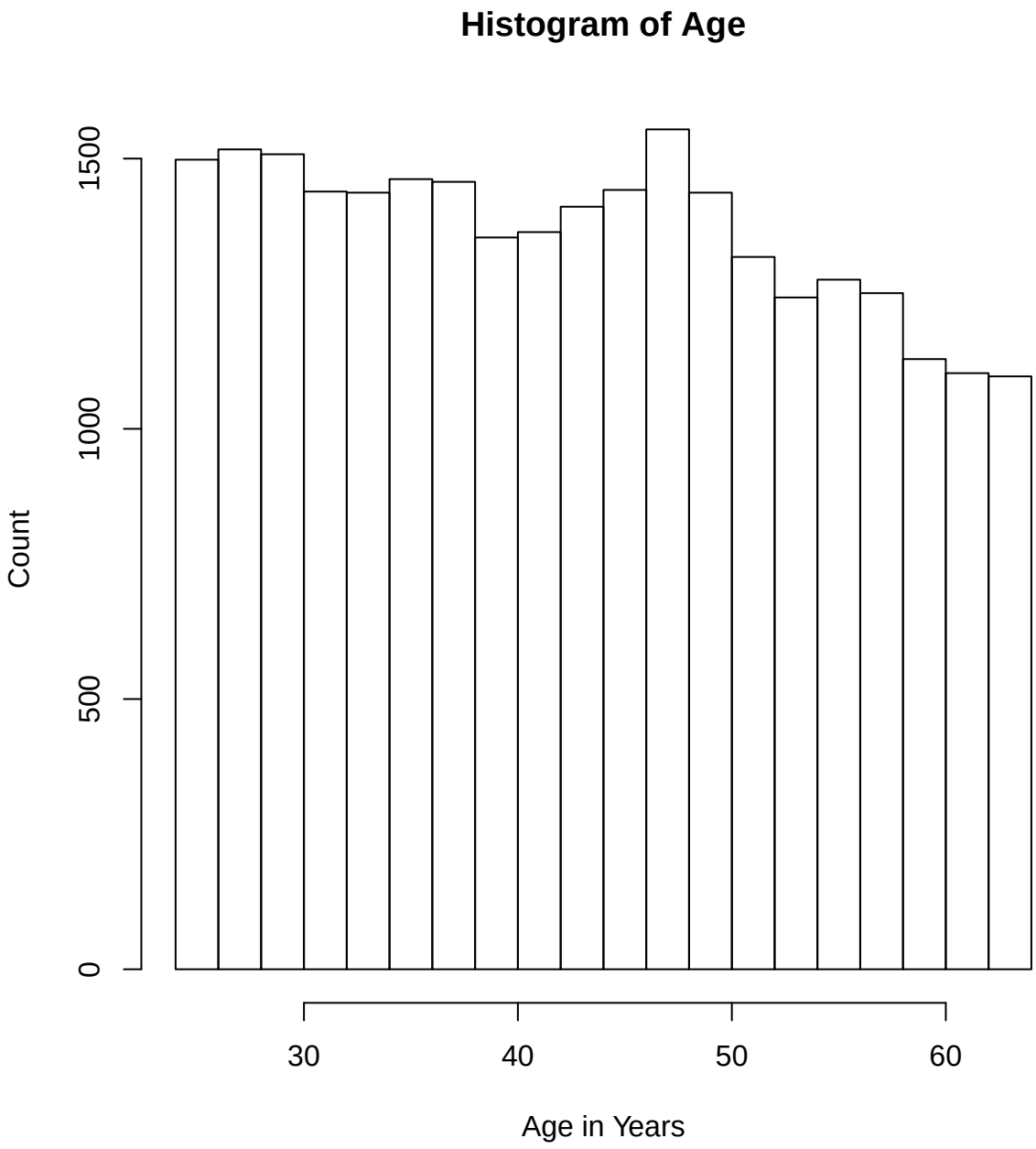
We will then plot the last two variables:

- YEAR
- NUMOBS
- ID

Histogram and Density Curve for AGE

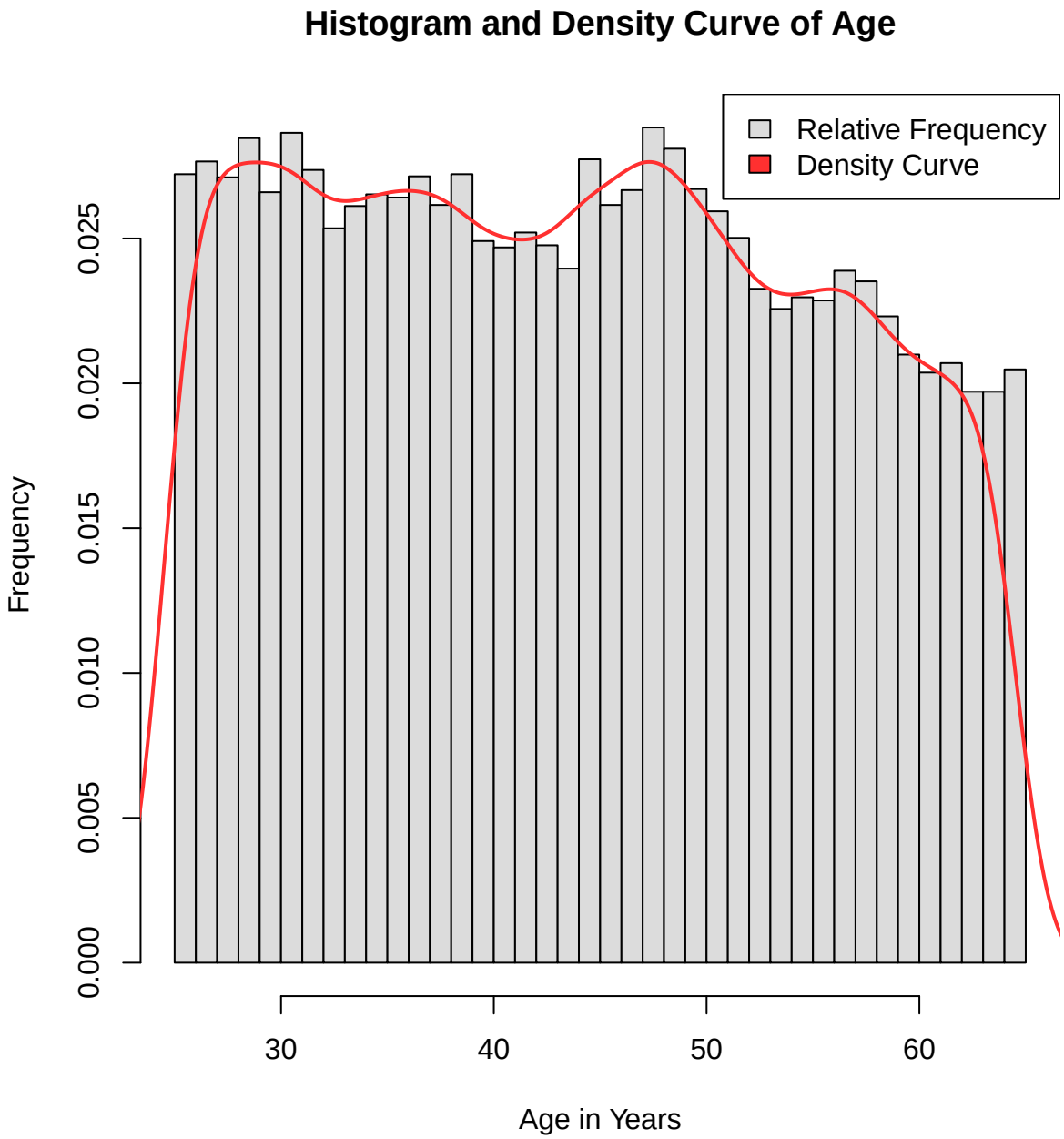
Histogram

```
hist(Pr3c$AGE, xlab= "Age in Years", ylab= "Count", main= "Histogram of Age")
```



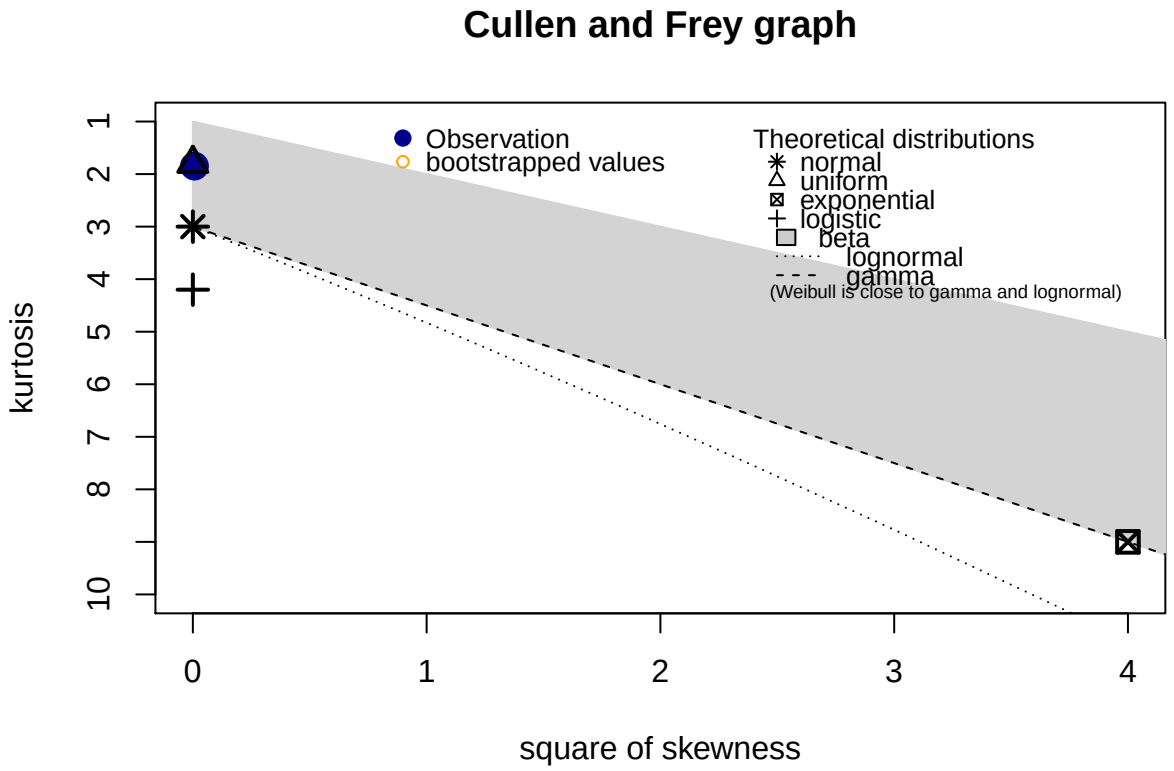
Histogram and Density Curve

```
truehist(Pr3c$AGE,col="gainsboro", ylab="Frequency", xlab="Age in Years",
         main="Histogram and Density Curve of Age")
lines(density(Pr3c$AGE), lwd=2,col="firebrick1")
legend('topright', c("Relative Frequency", "Density Curve"),
       fill=c("gainsboro", "firebrick1"))
```



We plot the Cullen and Frey graph to give us an idea of what distributions to try out:

```
descdist(Pr3c$AGE, boot = 1000)
```




```
## summary statistics
## ------
## min:  25   max:  64
## median:  43
## mean:  43.53163
## estimated sd:  11.32851
## estimated skewness:  0.08367055
## estimated kurtosis:  1.848404
```

It appears that this is very likely to be a uniform distribution.

We will test for a uniform distribution. We will also test for a normal distribution as logic tells us that the age distribution of 7293 individuals should realistically follow a normal distribution.

Testing fits for distributions

Testing for a uniform distribution

```
AGEunif <- fitdist(Pr3c$AGE, "unif")
```

Testing for a normal distribution

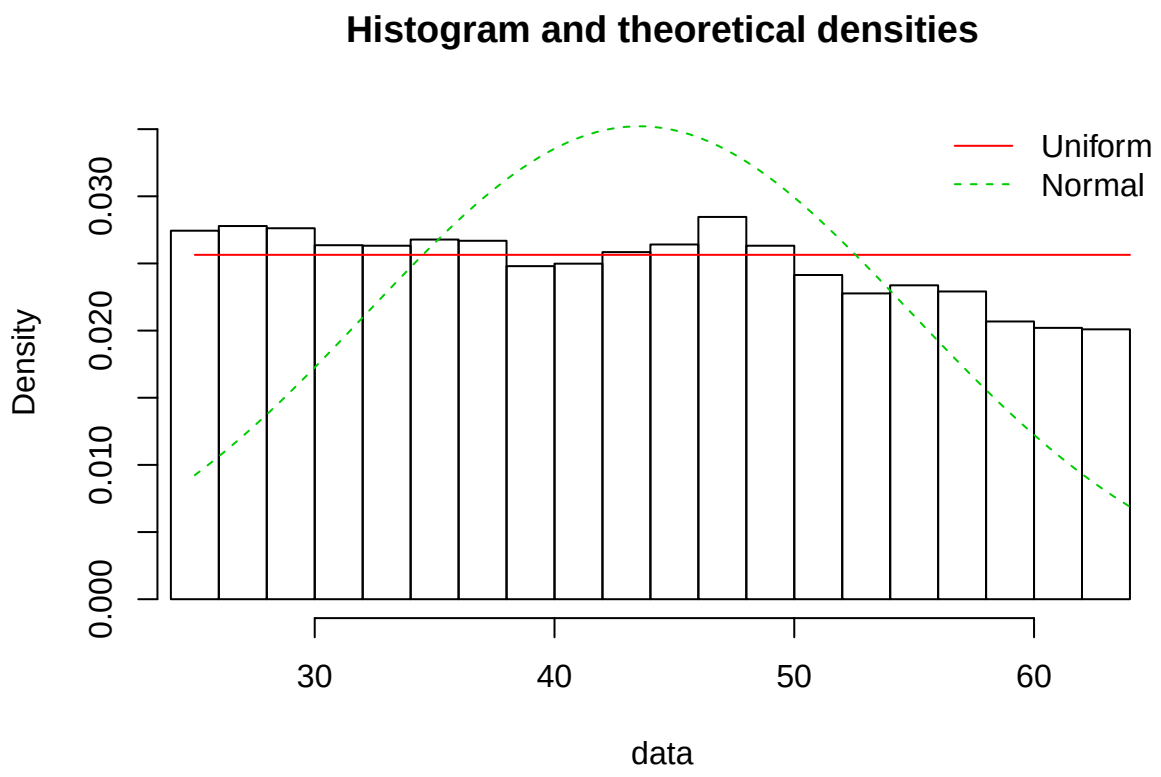
```
AGEnorm <- fitdist(Pr3c$AGE, "norm")
```

Setting Legend

```
plot.legend <- c("Uniform", "Normal")
```

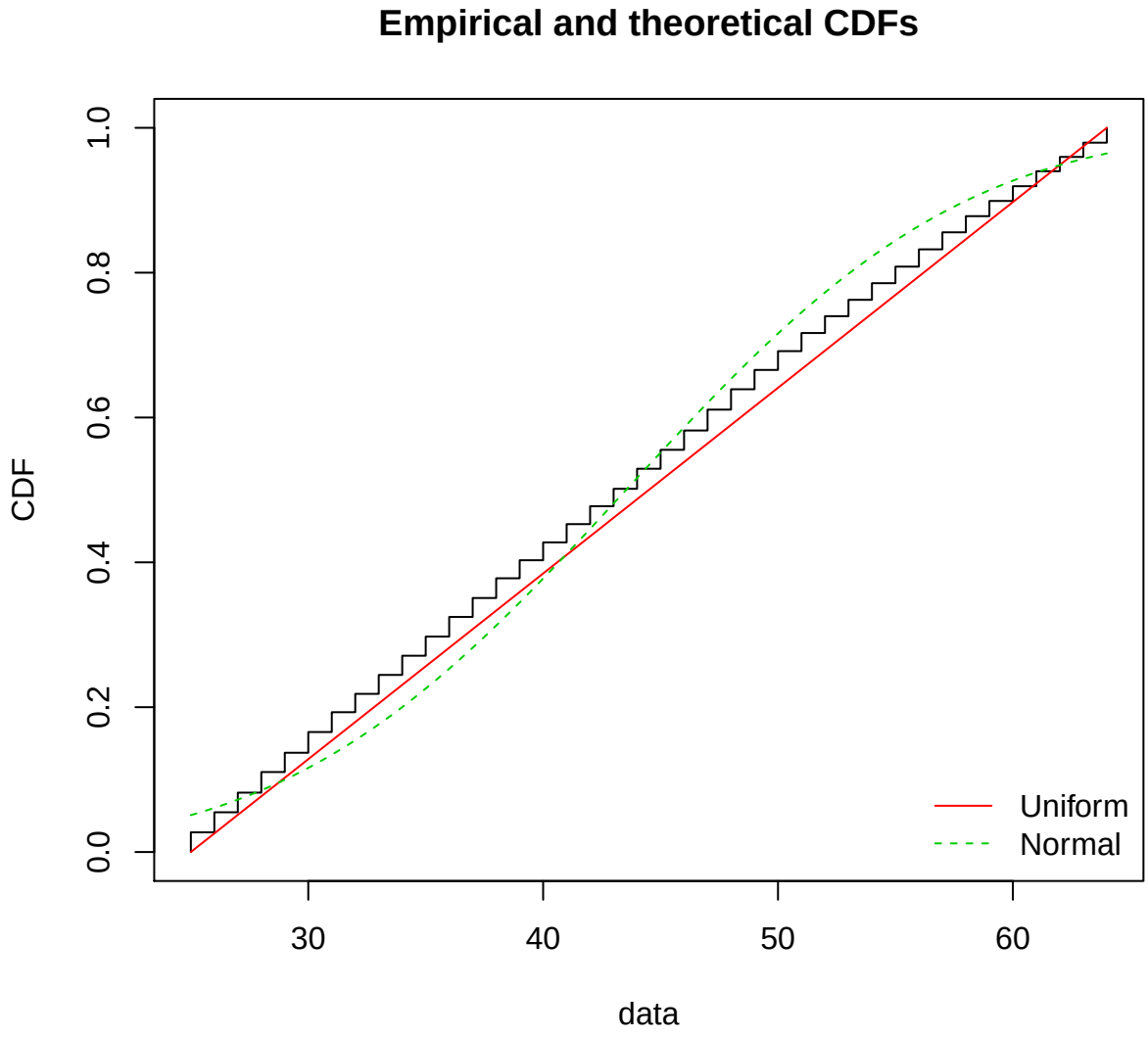
Comparing Histogram and Theoretical Densities

```
denscomp(list(AGEunif, AGEnorm), legendtext = plot.legend)
```



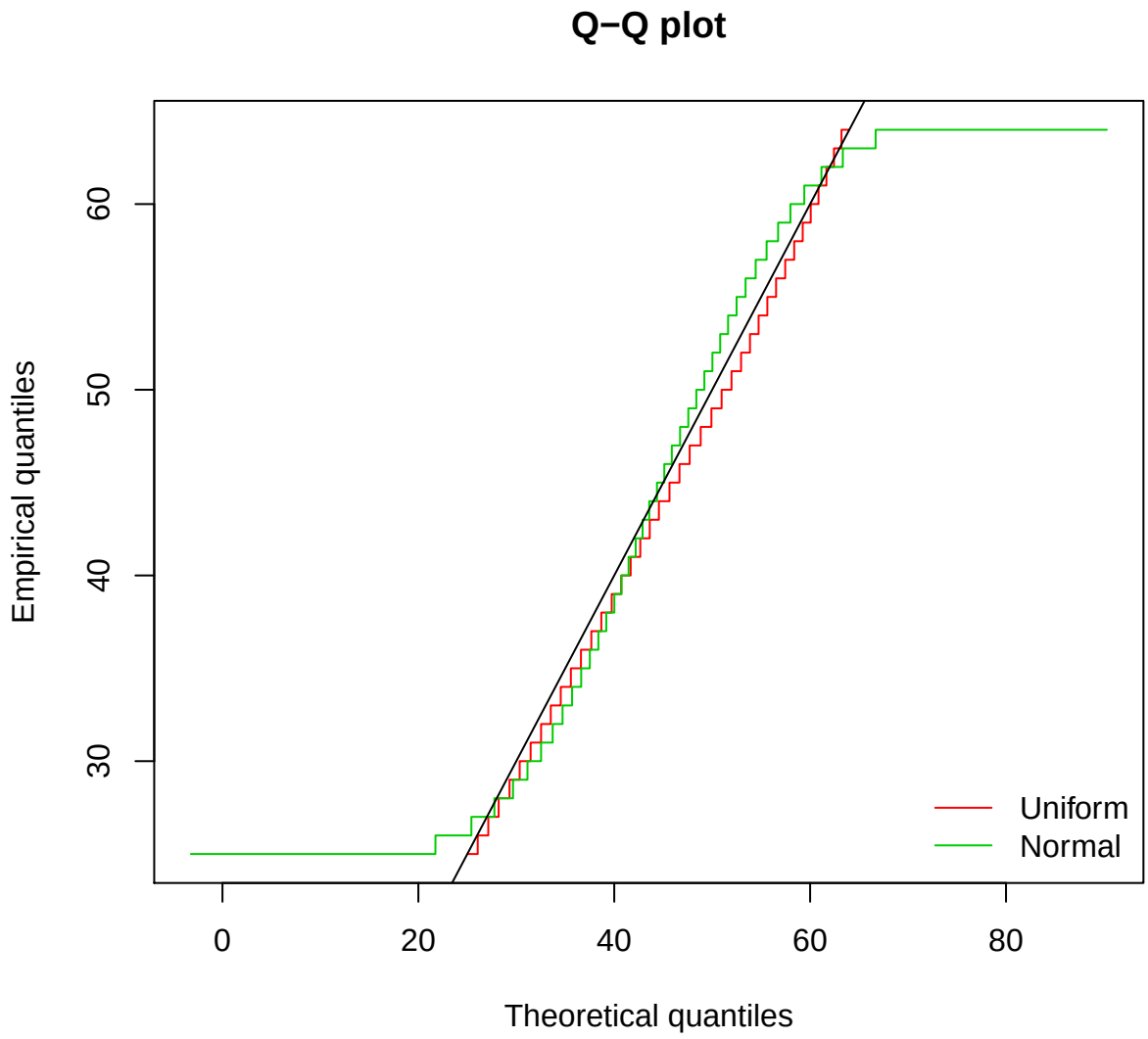
Observe that the uniform distribution seems to be the best fit based on the theoretical densities of the distributions.

```
cdfcomp(list(AGEunif, AGEnorm), legendtext = plot.legend)
```



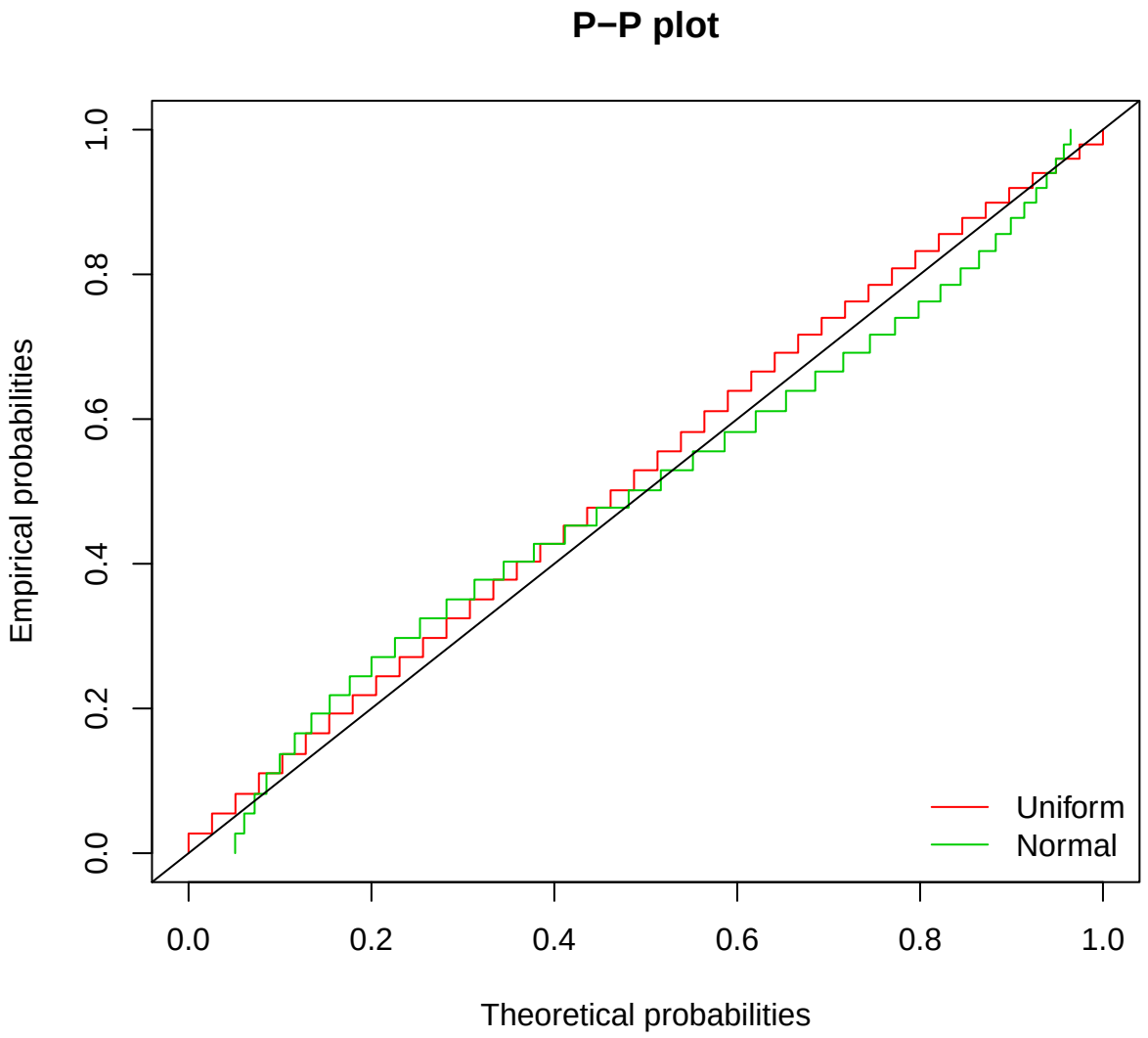
Observe that the uniform distribution appears much more appropriate than the normal distribution.

```
qqcomp(list(AGEunif, AGEnorm), legendtext = plot.legend)
```



As far as the empirical quantiles compared to the theoretical quantiles, the uniform distribution is much better than the normal distribution - which according to the Q-Q plot is not appropriate at all for ages less than around 20 and greater than around 60.

```
ppcomp(list(AGEunif, AGEnorm), legendtext = plot.legend)
```



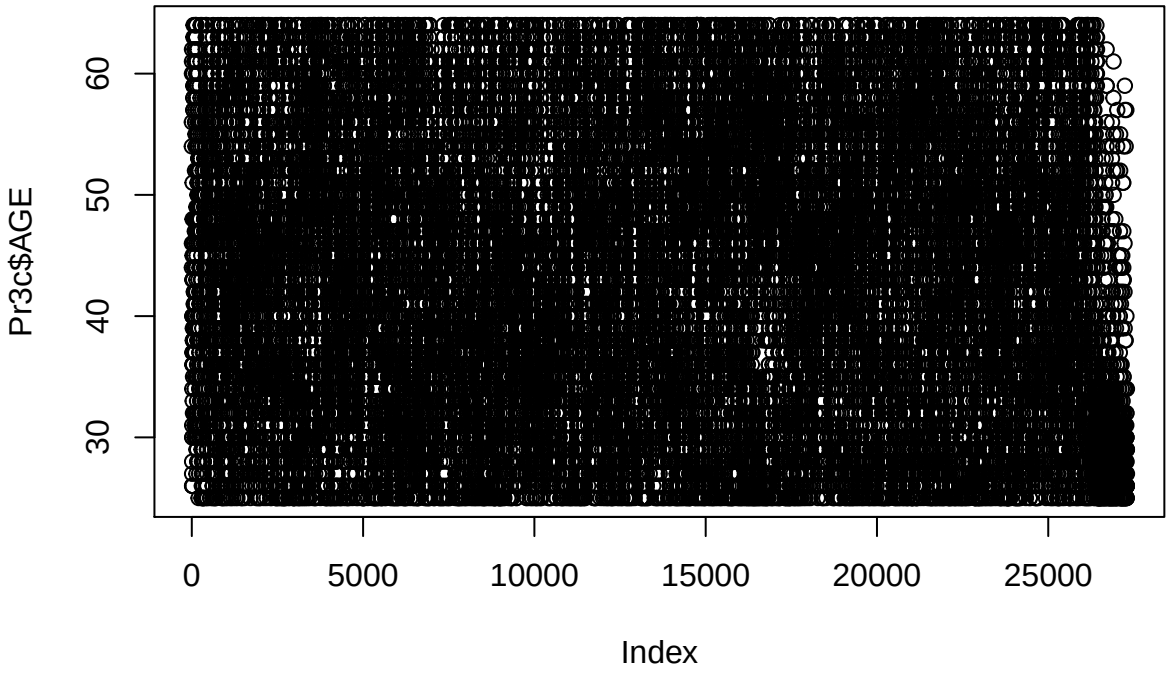
The P-P plot looks very similar to the CDF plot and confirms that the uniform distribution is better than the normal distribution.

Conclusion about AGE

We conclude that the AGE variable is best approximated by a uniform distribution. This is a highly unusual finding.

We examine a plot of the variable to confirm our conclusion:

```
plot(Pr3c$AGE)
```

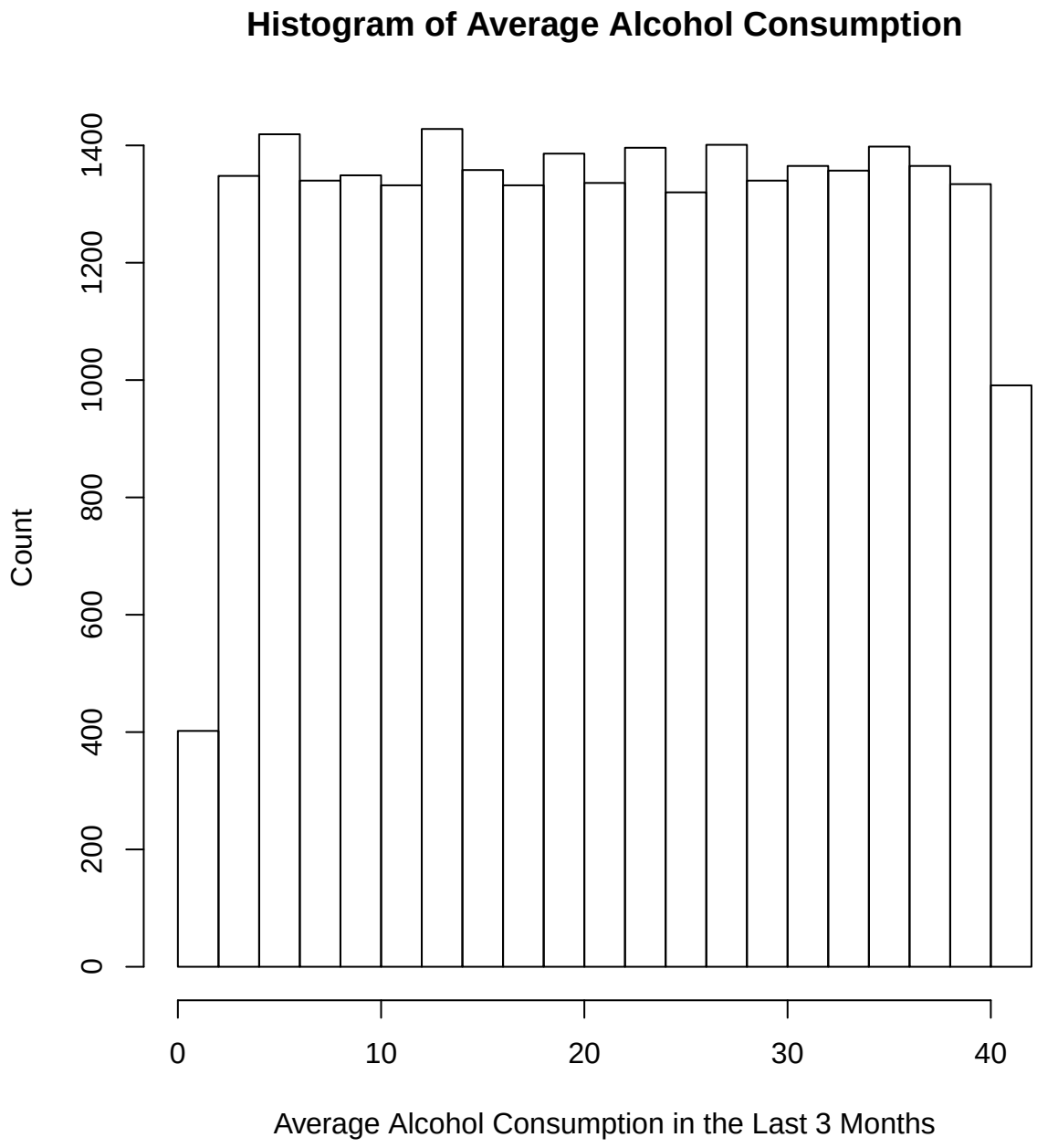


It does indeed appear that AGE follows a relatively uniform distribution.

Histogram and Density Curve for ALC

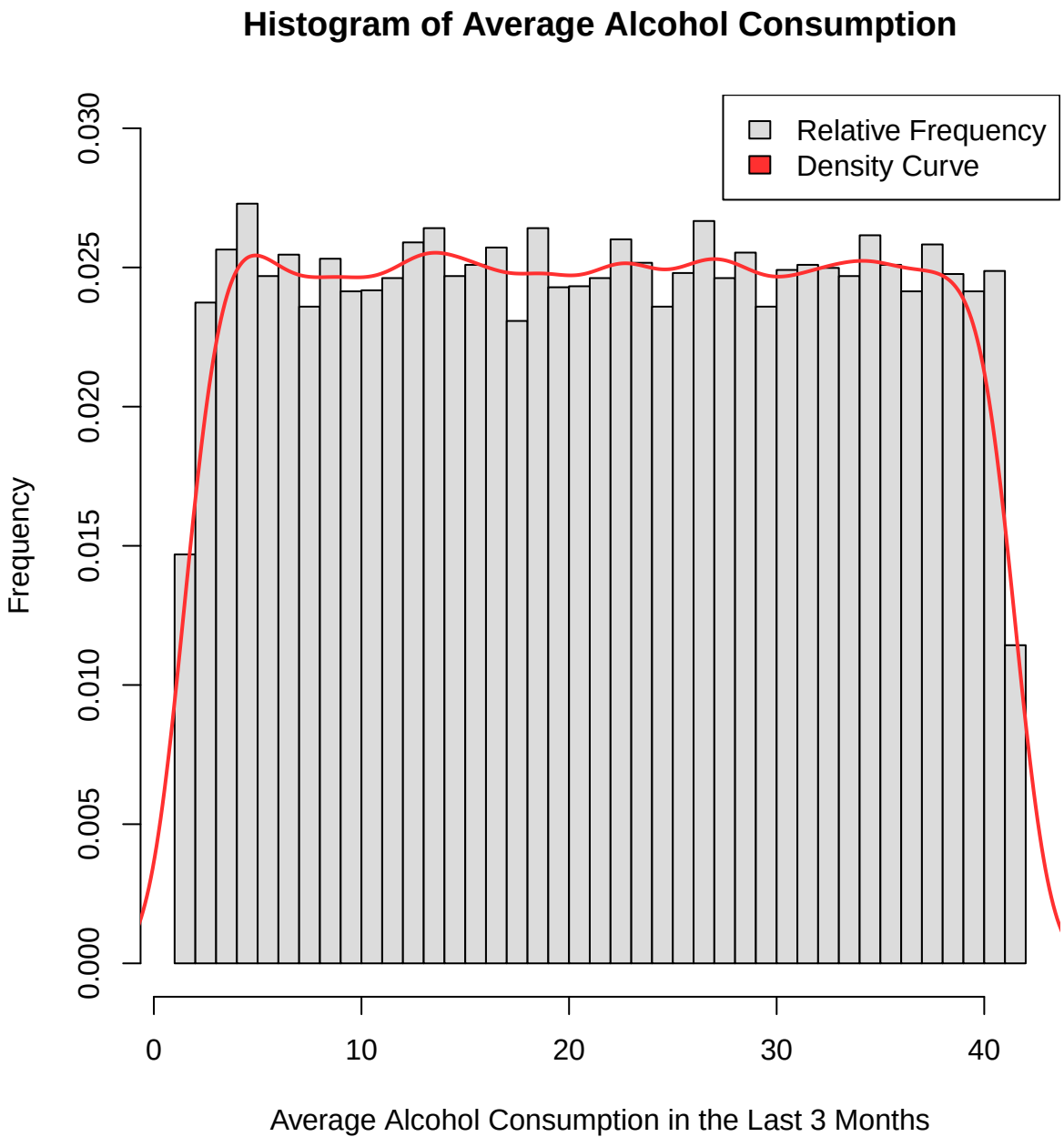
Histogram

```
hist(Pr3c$ALC, xlab= "Average Alcohol Consumption in the Last 3 Months",
     ylab= "Count", main= "Histogram of Average Alcohol Consumption")
```



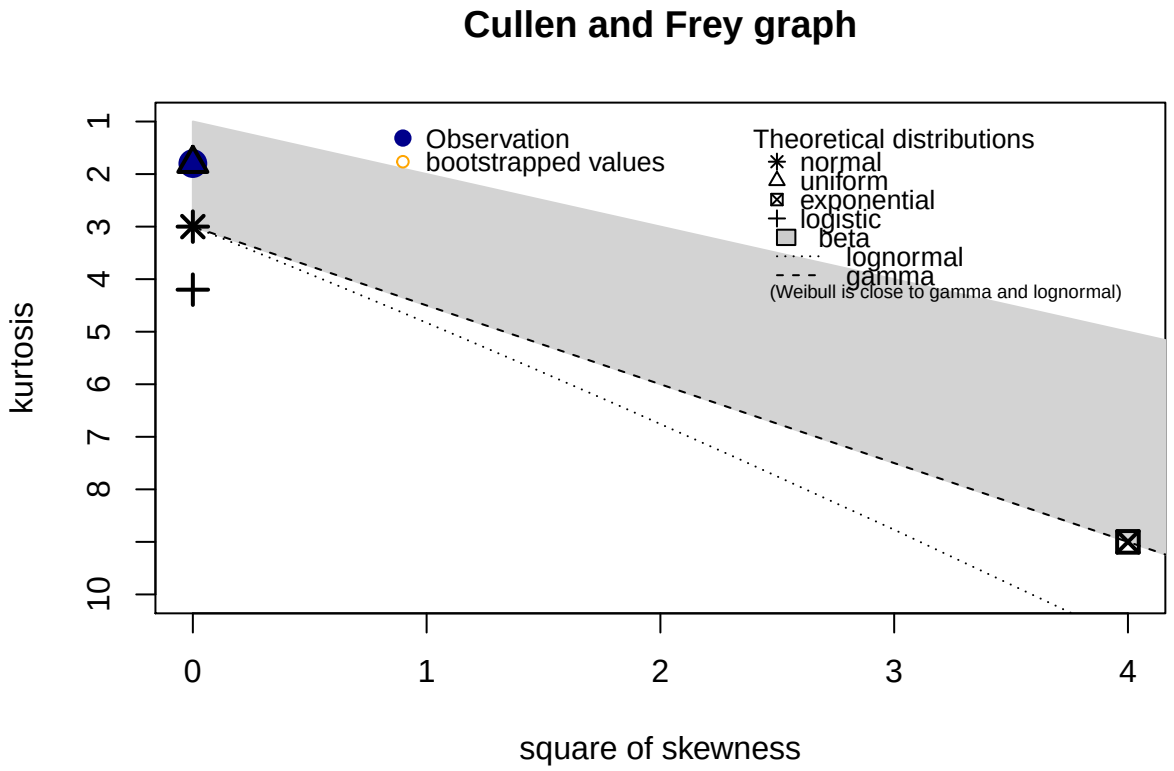
Histogram and Density Curve

```
truehist(Pr3c$ALC,col="gainsboro", ylab="Frequency",
         xlab= "Average Alcohol Consumption in the Last 3 Months",
         main= "Histogram of Average Alcohol Consumption", ylim = c(0,0.03))
lines(density((Pr3c$ALC)), lwd=2,col="firebrick1")
legend("topright", c("Relative Frequency", "Density Curve"),
       fill=c("gainsboro", "firebrick1"))
```



We plot the Cullen and Frey graph to give us an idea of what distributions to try out:

```
descdist(Pr3c$ALC, boot = 1000)
```




```
## summary statistics
## ------
## min:  1.448   max:  41.446
## median:  21.452
## mean:  21.41312
## estimated sd:  11.56067
## estimated skewness:  0.001096402
## estimated kurtosis:  1.799469
```

It appears that this is very likely to be a uniform distribution.

We will test for a uniform distribution. We will also test for a normal distribution as logic tells us that the alcohol consumption distribution of 7293 individuals should realistically follow a normal distribution.

Testing fits for distributions

Testing for a uniform distribution

```
ALCunif <- fitdist(Pr3c$ALC, "unif")
```

Testing for a normal distribution

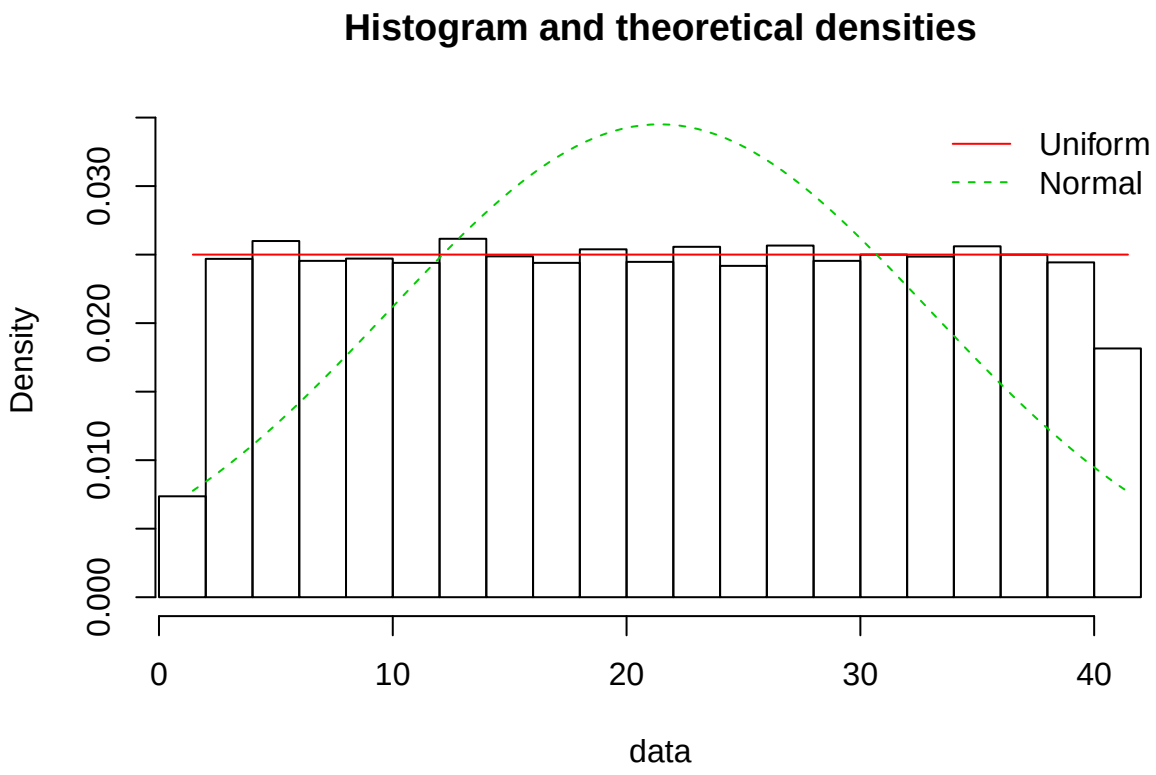
```
ALCnorm <- fitdist(Pr3c$ALC, "norm")
```

Setting Legend

```
plot.legend <- c("Uniform", "Normal")
```

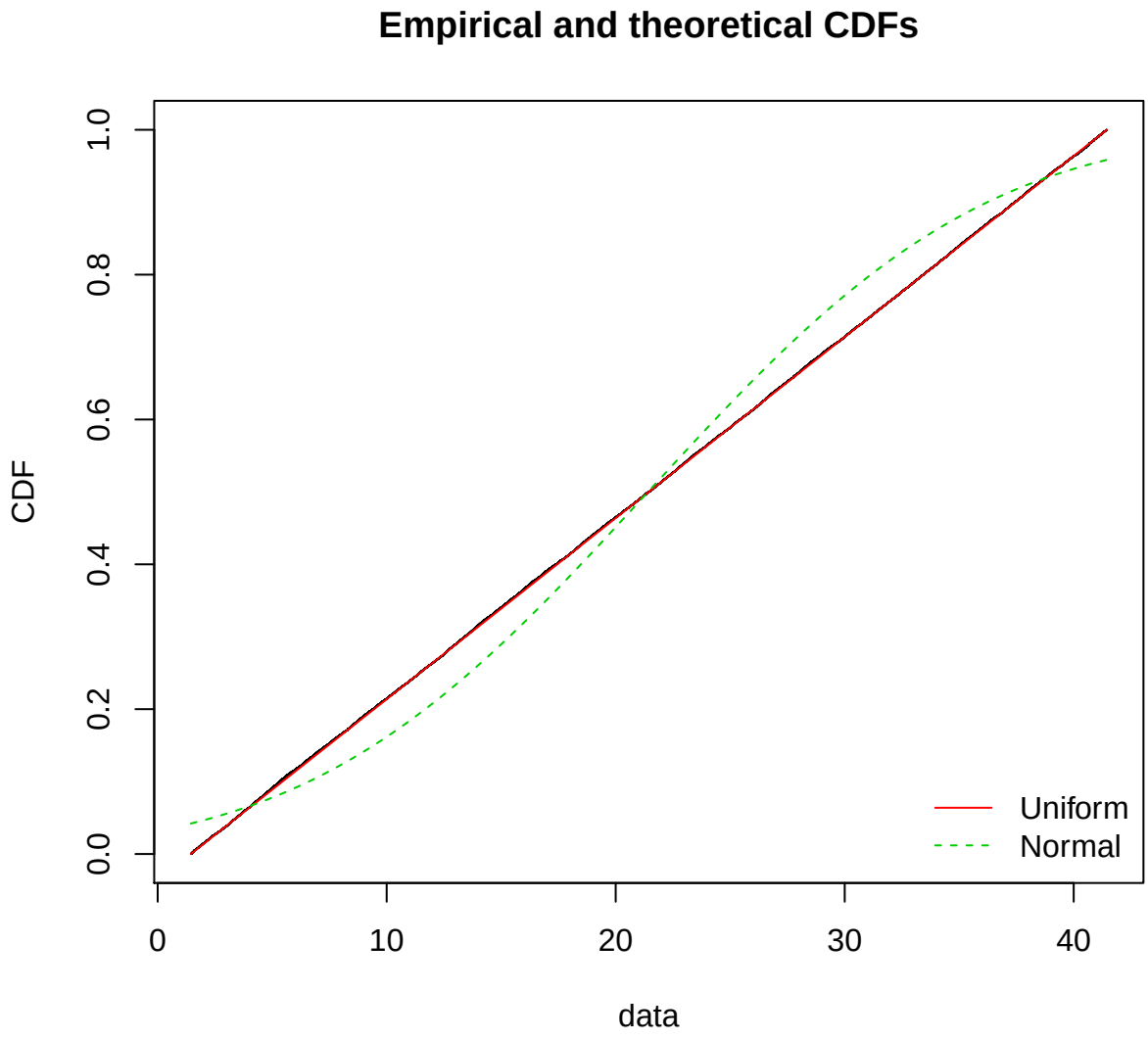
Comparing Histogram and Theoretical Densities

```
denscomp(list(ALCunif, ALCnorm), legendtext = plot.legend)
```



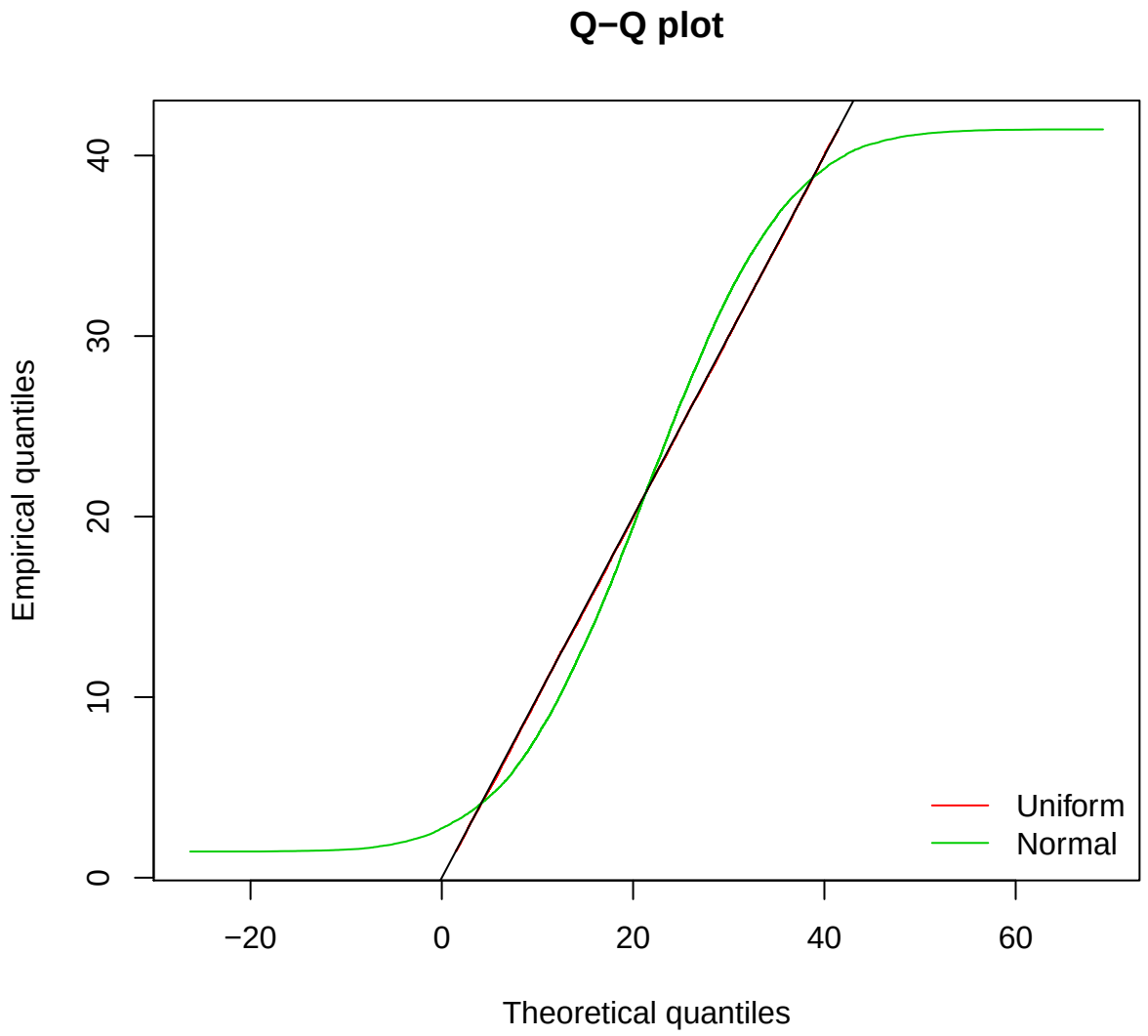
Observe that the uniform distribution seems to be the best fit based on the theoretical densities of the distributions.

```
cdfcomp(list(ALCunif, ALCnorm), legendtext = plot.legend)
```



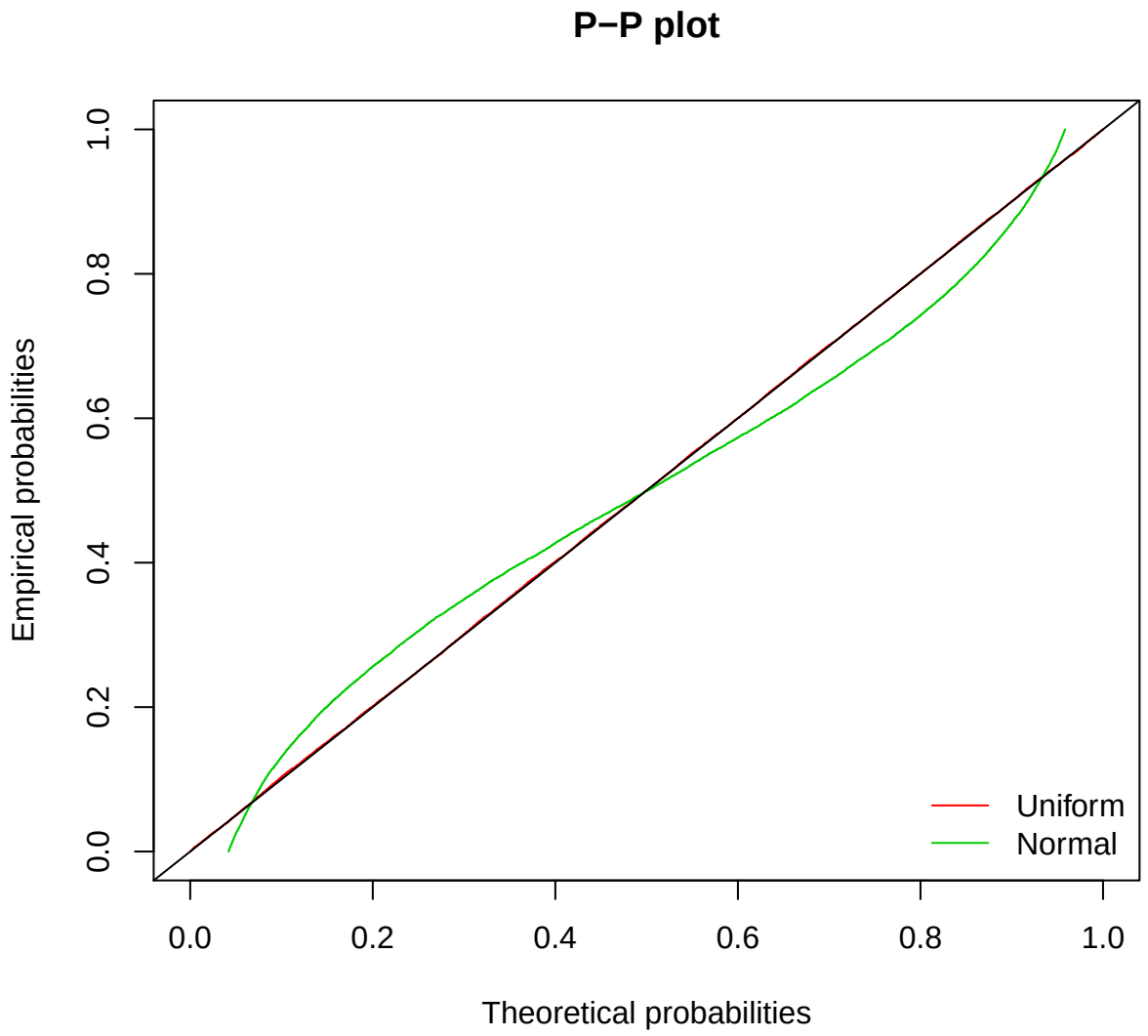
Observe that the uniform distribution appears to be the best distribution.

```
qqcomp(list(ALCunif, ALCnorm), legendtext = plot.legend)
```



As far as the empirical quantiles compared to the theoretical quantiles, the uniform distribution is much better than the normal distribution - which according to the Q-Q plot is not appropriate at all. In fact, the normal distribution makes no sense based on the Q-Q plot.

```
ppcomp(list(ALCunif, ALCnorm), legendtext = plot.legend)
```



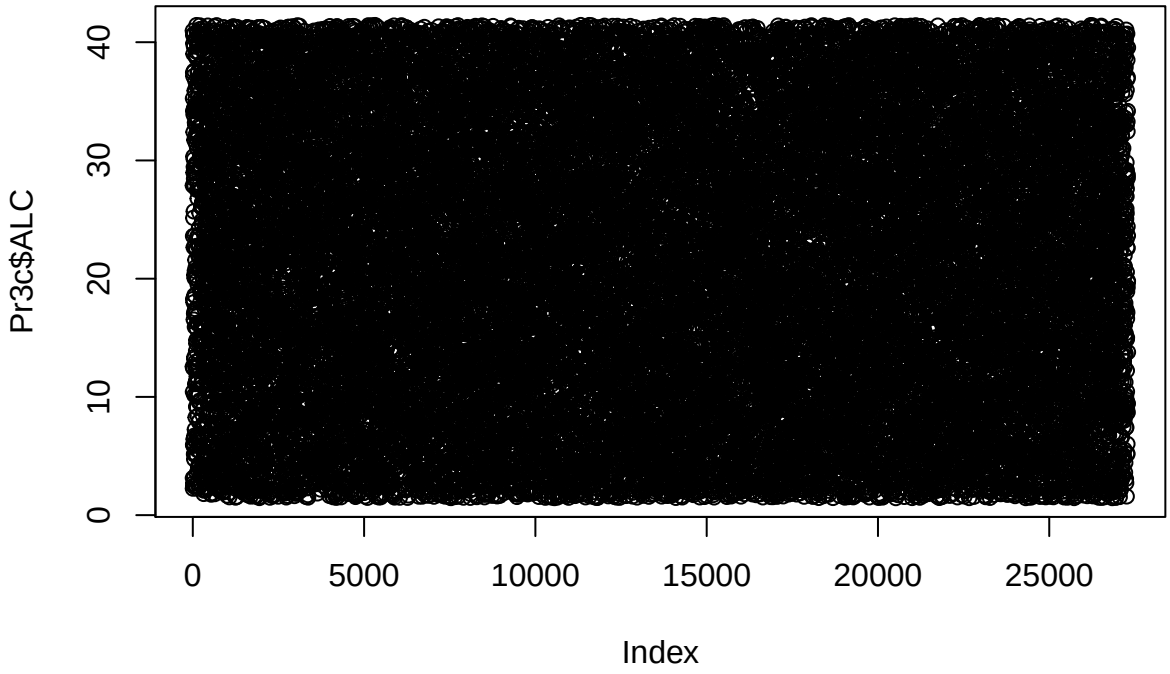
The P-P plot confirms our conclusion from the Q-Q plot.

Conclusion about ALC

We conclude that the ALC variable is best approximated by a uniform distribution. This is a highly unusual finding as one would expect the alcohol consumption of individuals to look approximate a normal distribution.

We examine a plot of the variable to confirm our conclusion:

```
plot(Pr3c$ALC)
```

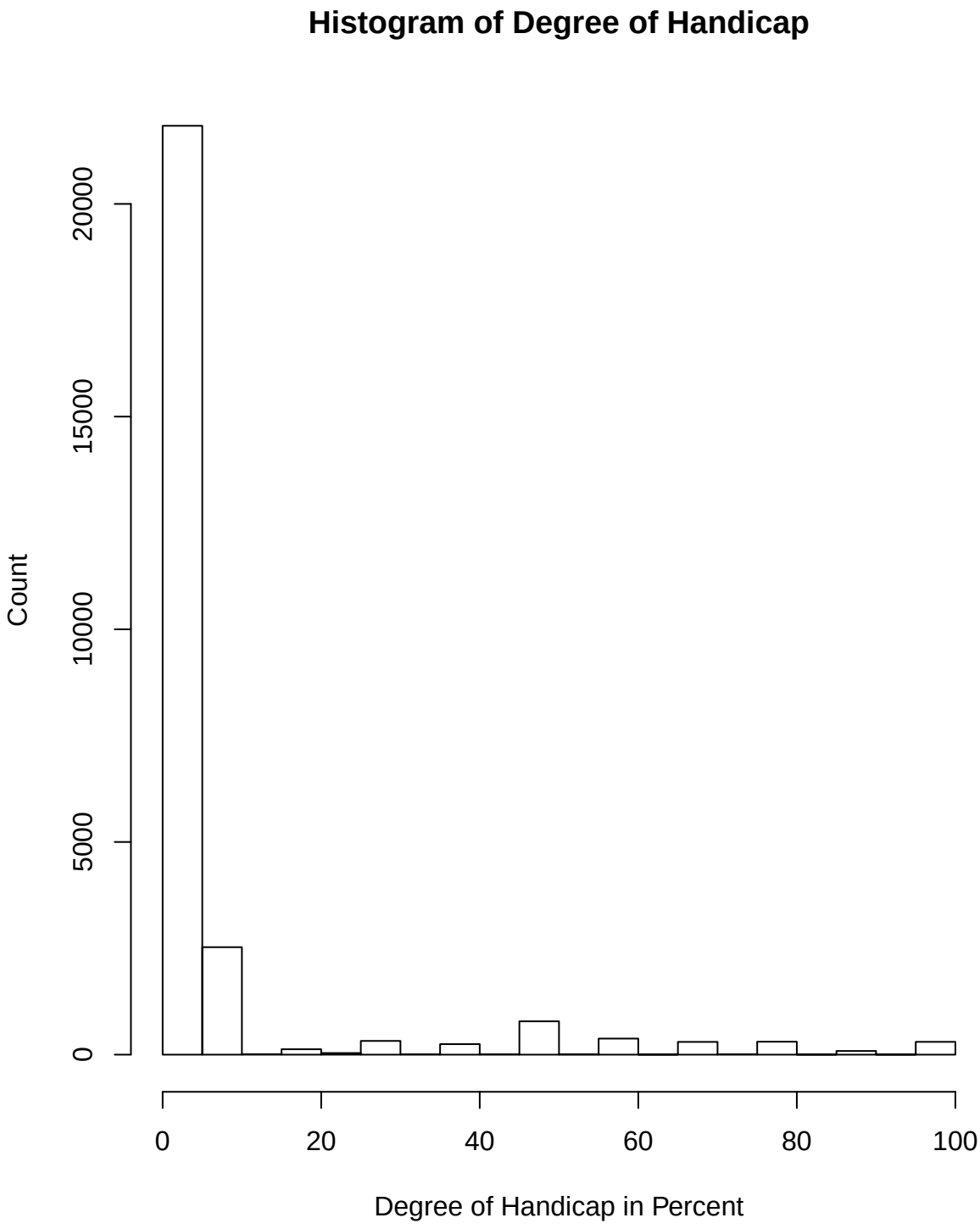


It does indeed appear that our data for ALC follows a uniform distribution.

Histogram and Density Curve for HANDPER

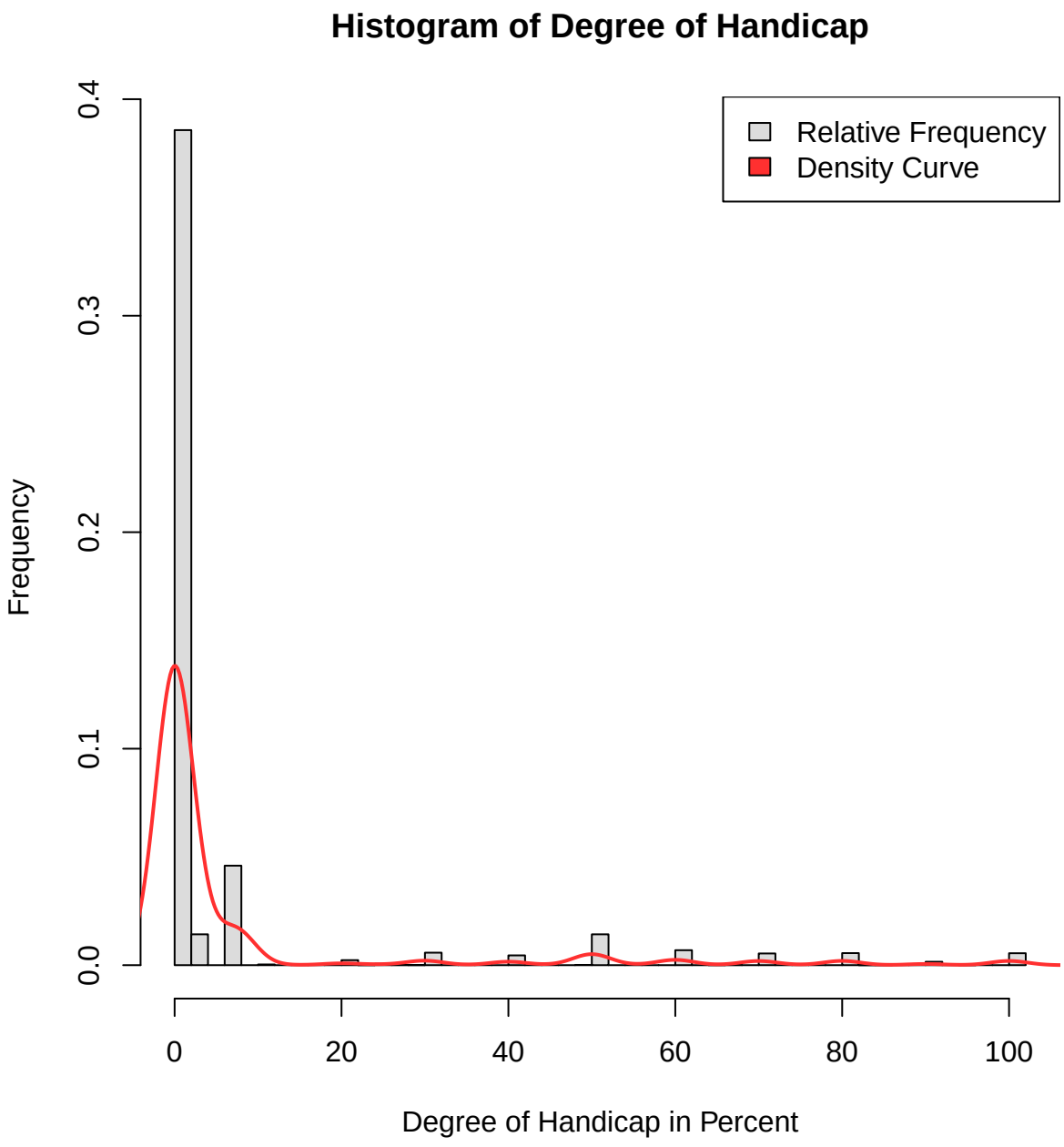
Histogram

```
hist(Pr3c$HANDPER, xlab= "Degree of Handicap in Percent", ylab= "Count",
     main= "Histogram of Degree of Handicap")
```



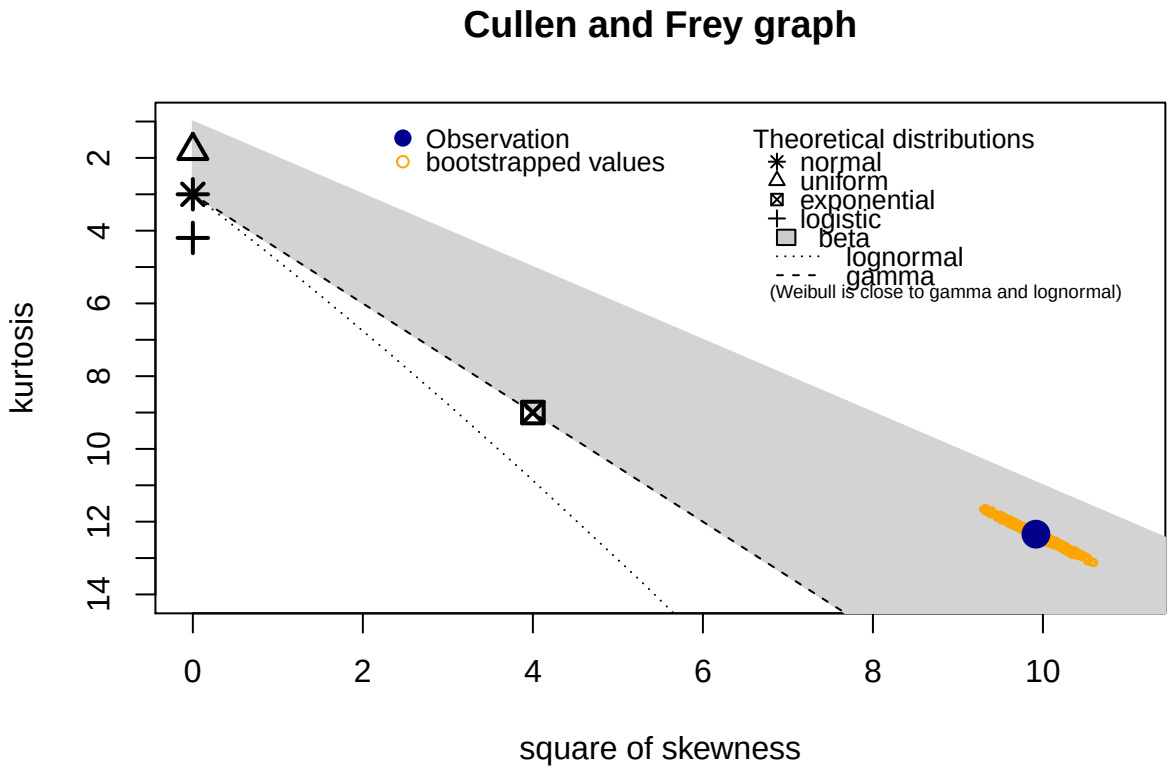
Histogram and Density Curve

```
truehist(Pr3c$HANDPER,col="gainsboro", ylab="Frequency",
         xlab= "Degree of Handicap in Percent",
         main= "Histogram of Degree of Handicap")
lines(density((Pr3c$HANDPER)), lwd=2,col="firebrick1")
legend("topright", c("Relative Frequency", "Density Curve"),
       fill=c("gainsboro", "firebrick1"))
```



We plot the Cullen and Frey graph to give us an idea of what distributions to try out:

```
descdist(Pr3c$HANDPER, boot = 1000)
```




```
## summary statistics
## ------
## min:  0   max:  100
## median:  0
## mean:  7.015266
## estimated sd:  19.26947
## estimated skewness:  3.14922
## estimated kurtosis:  12.34326
```

Observe that normal, uniform, and logistic distributions have to have the square of skewness equal to 0.

Thus the HANDPER data does not fit any one of those distributions.

Also, observe that the exponential distribution has to have a square of skewness value equal to 4.

Since the data for HANDPER has a square of skewness equal to 10, the best distribution is very unlikely to be exponential.

Also observe that a lognormal distribution can not have any values equal to zero so that is also not a possibility.

We will attempt to fit various gamma distributions.

Testing fits for distributions

Testing fit for gamma distribution

```
HANDPERgamma1 <- fitdist(Pr3c$HANDPER, distr = "gamma", method = "mme")
```

Testing fit for gamma distribution with different parameters

```
HANDPERgamma2 <- fitdist(Pr3c$HANDPER, distr = "gamma", method = "mle",
                         lower = c(0, 0), start = list(scale = 1, shape = 1))
```

Testing fit for gamma distribution with different parameters

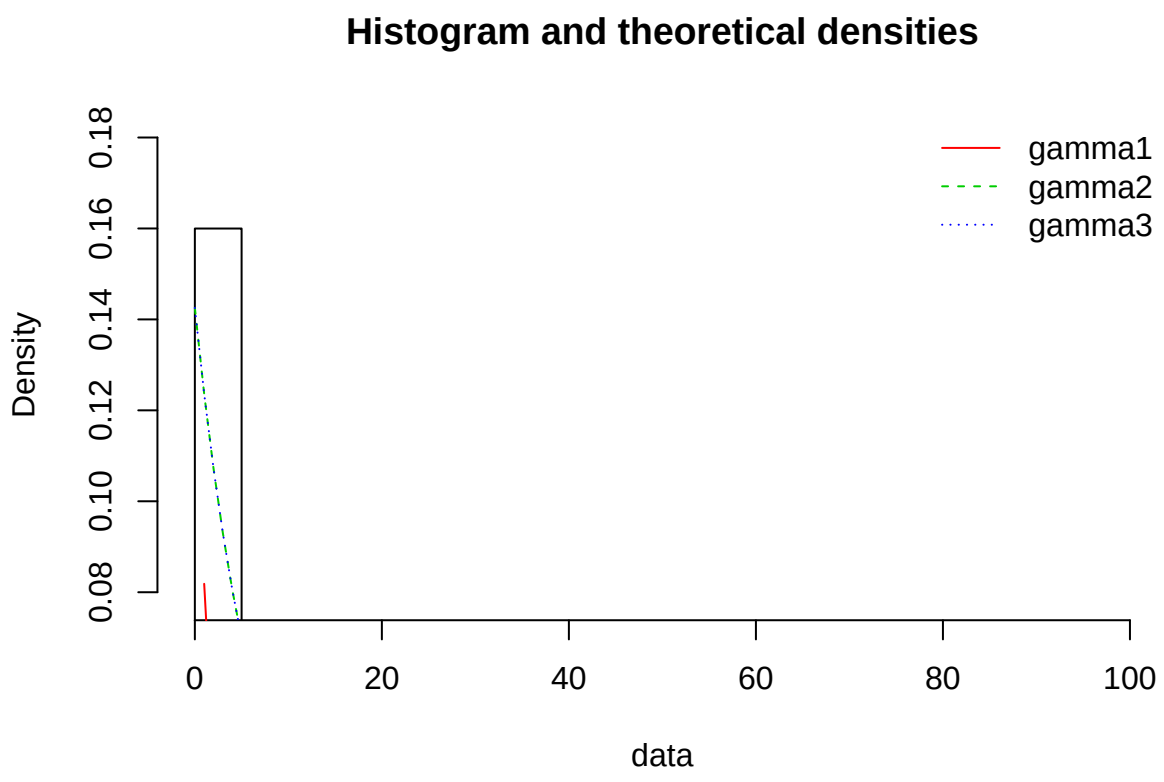
```
HANDPERgamma3 <- fitdist(Pr3c$HANDPER, distr = "gamma", method = "mle",
                         lower = c(0, 0), start = list(scale = 0.5, shape = 1))
```

Setting Legend

```
plot.legend <- c("gamma1", "gamma2", "gamma3")
```

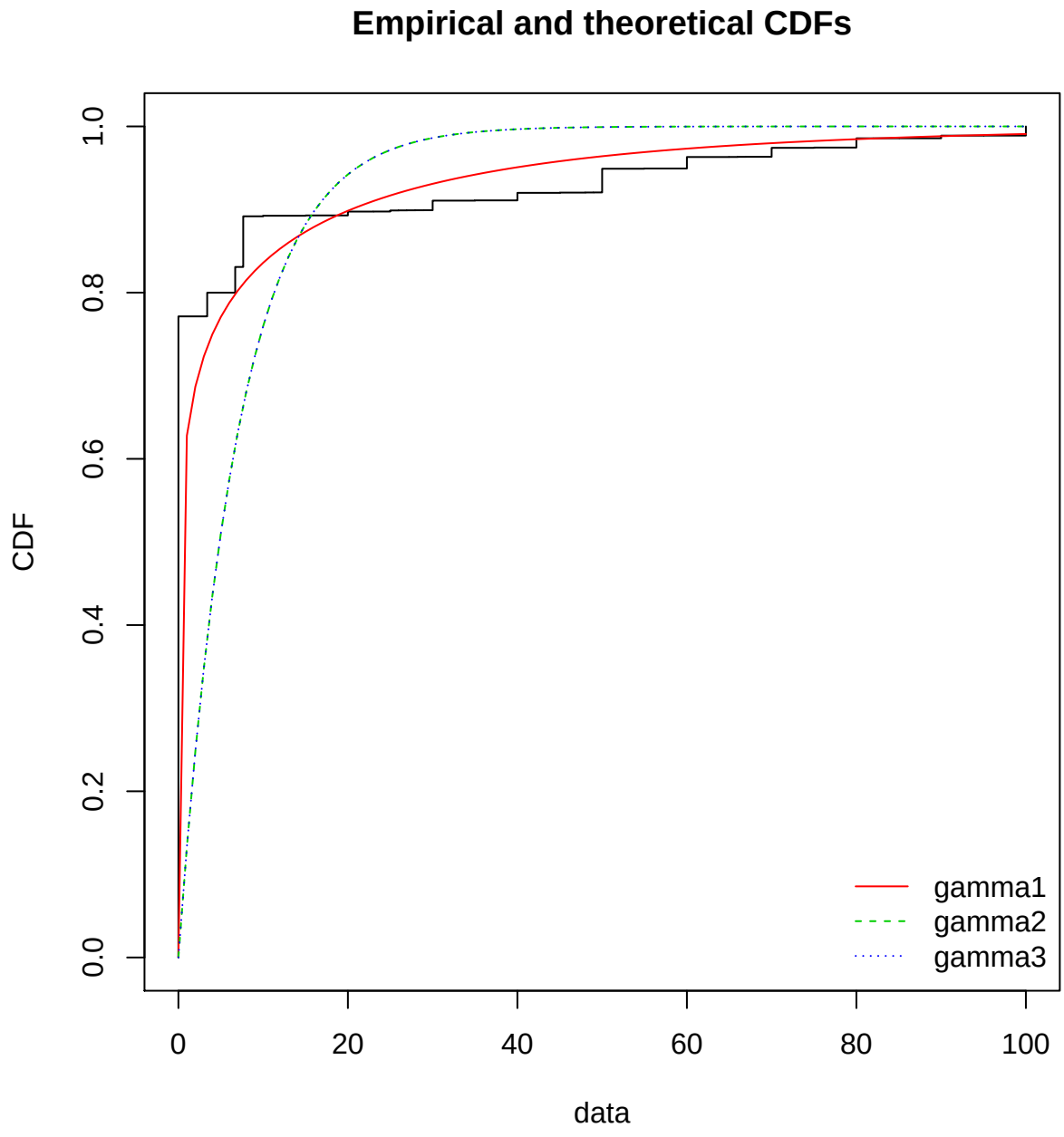
Comparing Histogram and Theoretical Densities

```
denscomp(list(HANDPERgamma1, HANDPERgamma2, HANDPERgamma3),
         legendtext = plot.legend, ylim = 0.13)
```



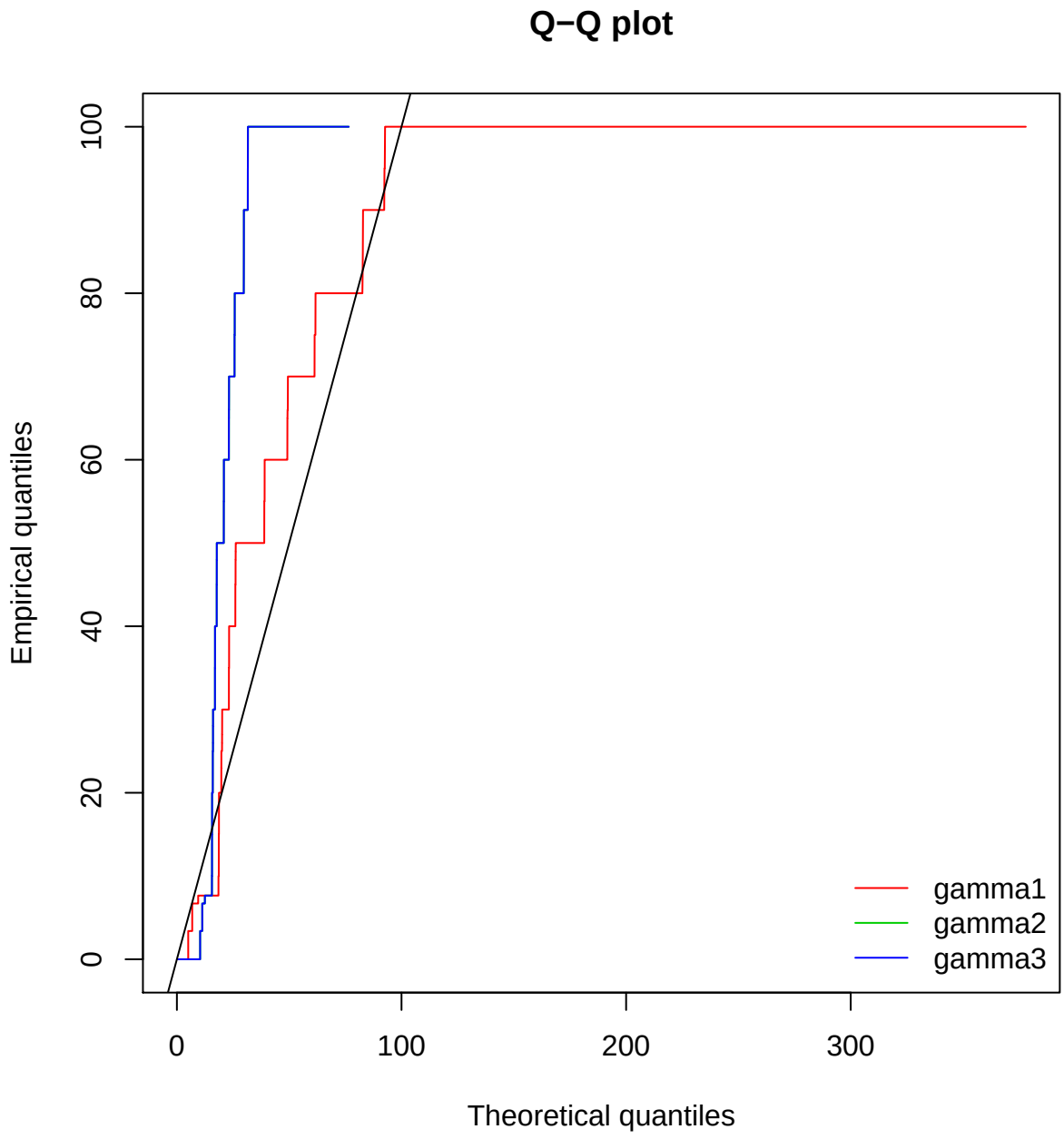
Observe that the gamma distributions do appear to be the best fits based on the theoretical densities of the distributions.

```
cdfcomp(list(HANDPERgamma1, HANDPERgamma2, HANDPERgamma3),
        legendtext = plot.legend)
```



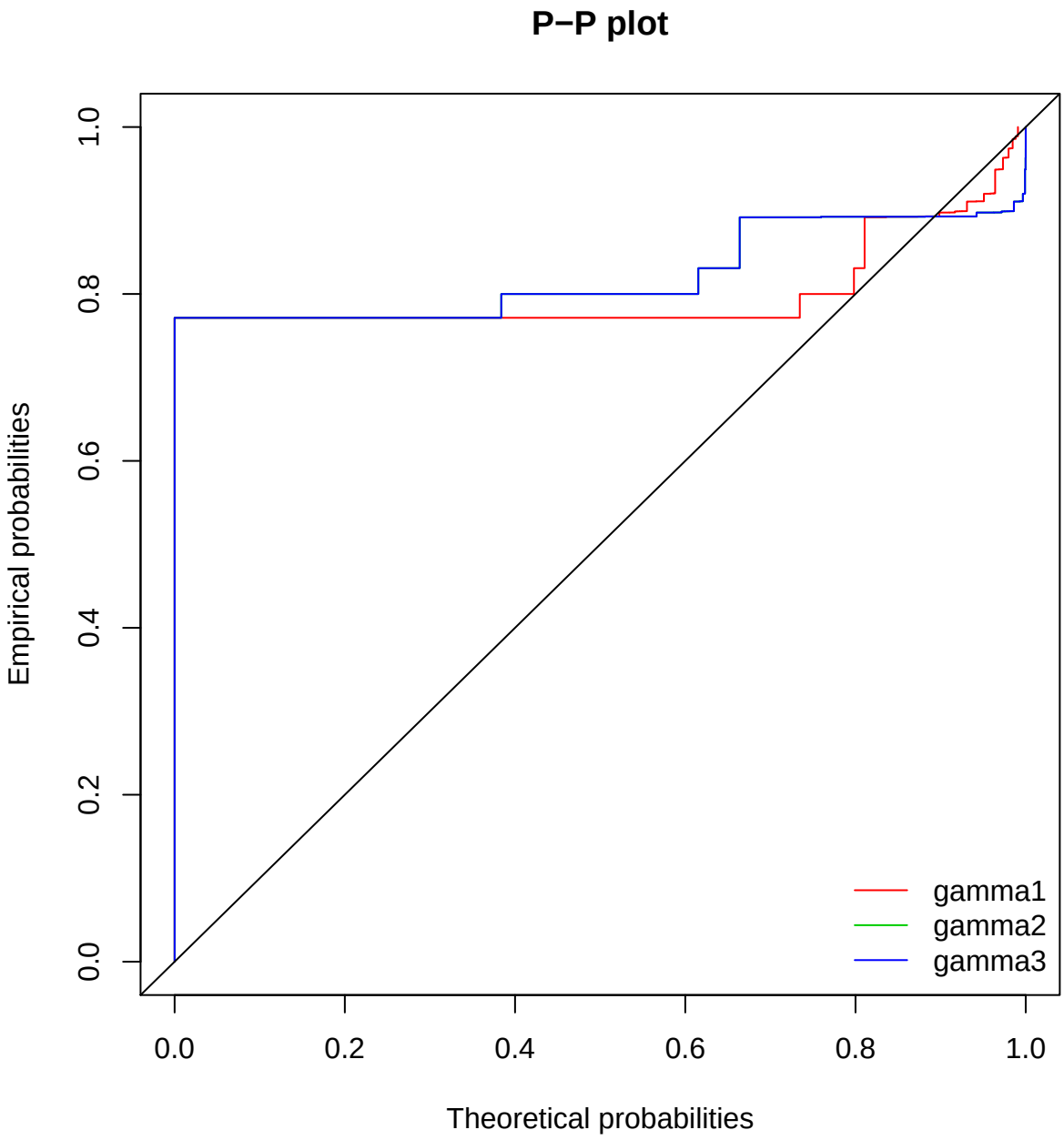
The gamma distributions do appear to be appropriate approximations.

```
qqcomp(list(HANDPERgamma1, HANDPERgamma2, HANDPERgamma3), legendtext = plot.legend)
```



The Q-Q plot appears to contradict our previous conclusion about the gamma distribution being a good fit.

```
ppcomp(list(HANDPERgamma1, HANDPERgamma2, HANDPERgamma3), legendtext = plot.legend)
```



The P-P plot supports our conclusion from the Q-Q plot, and contradicts our conclusion from the density and cumulative distribution function plots.

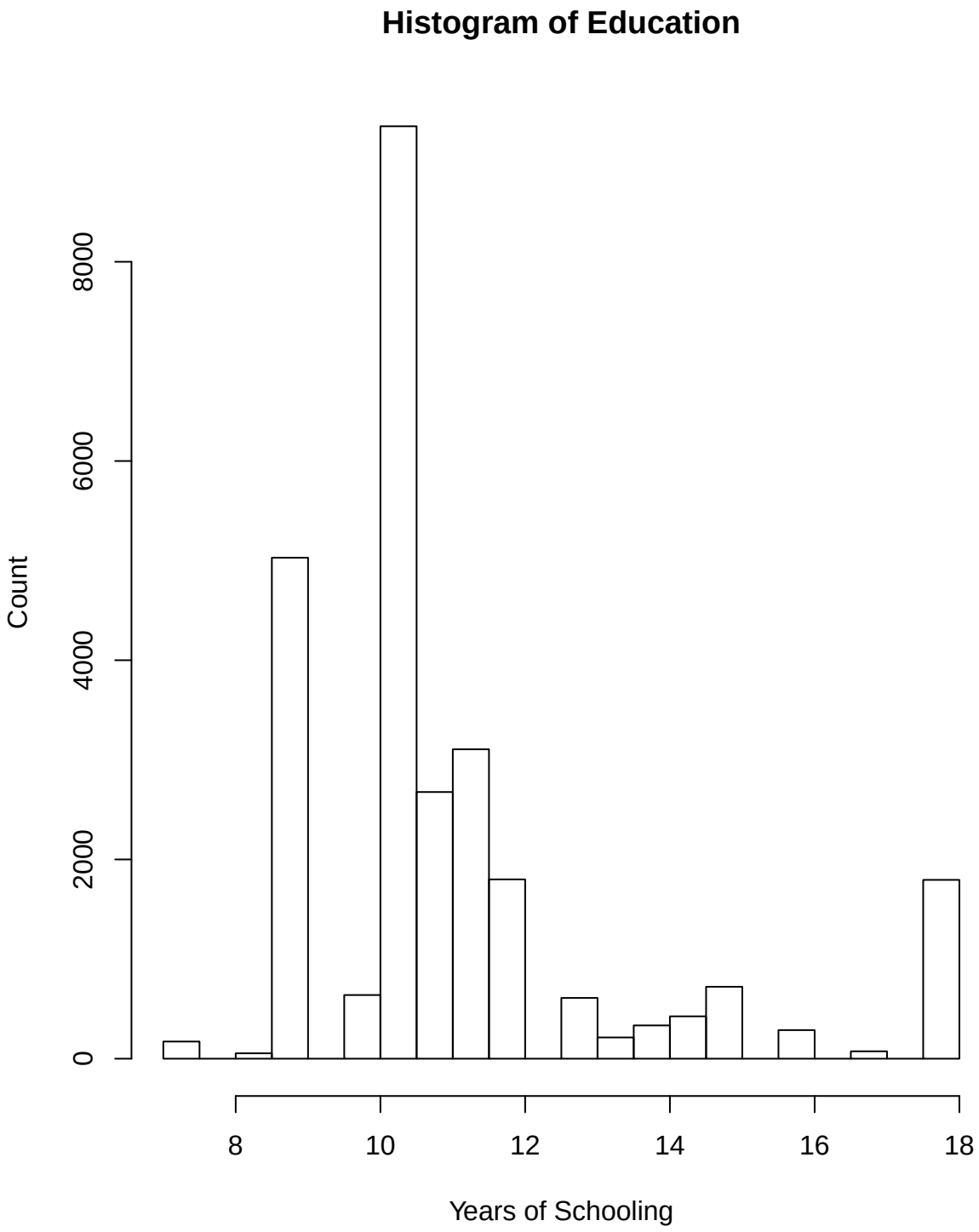
Conclusion about HANDPER

We conclude that the HANDPER variable is best approximated by a gamma distribution, given that basically all other distributions we would attempt based on the Cullen and Frey graph do not appear to be good fits and the empirical vs. theoretical CDF plot as well as the Q-Q plot indicate that the two distributions are quite similar.

Histogram and Density Curve for EDUC

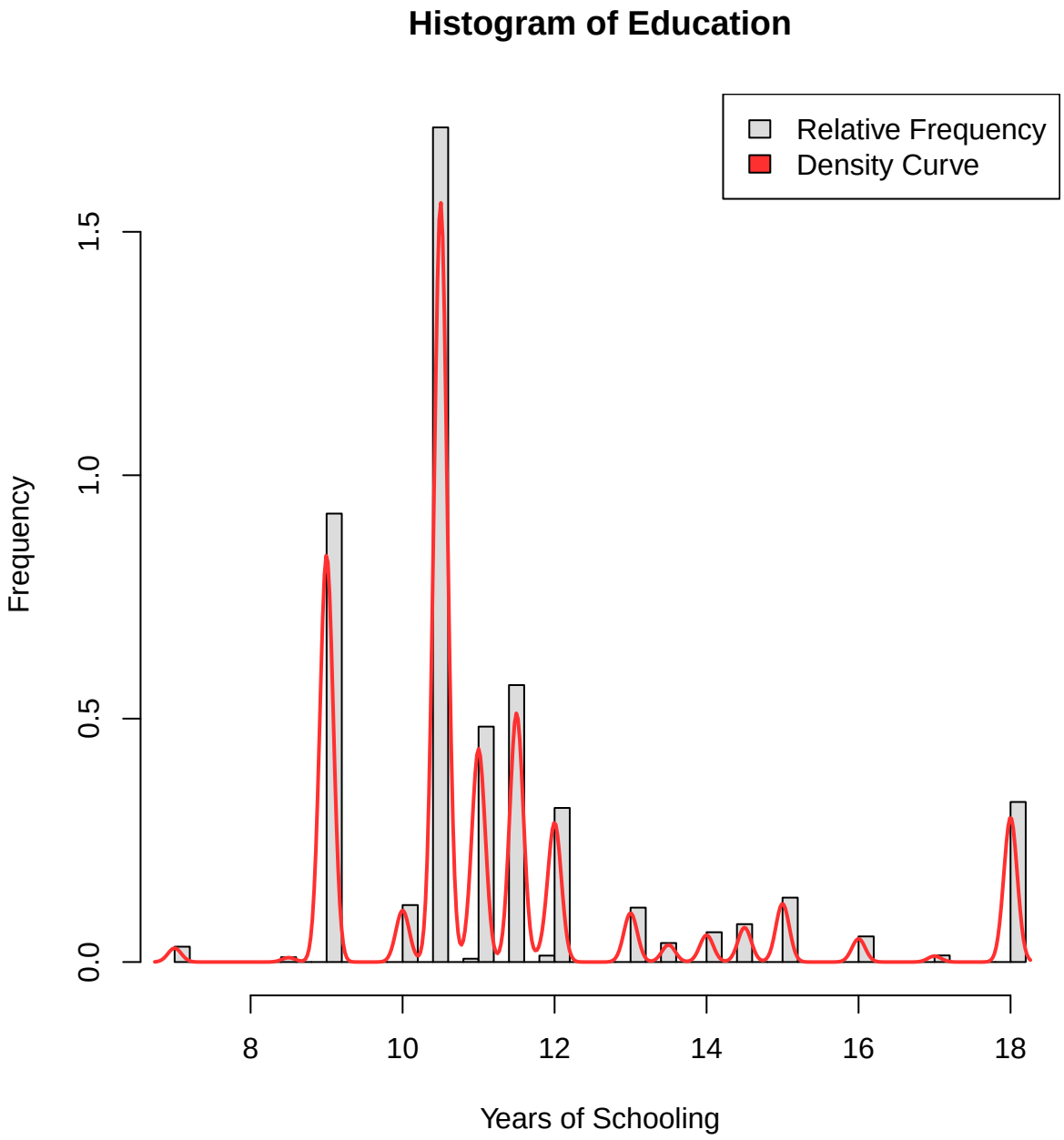
Histogram

```
hist(Pr3c$EDUC, xlab= "Years of Schooling", ylab= "Count",
     main= "Histogram of Education")
```



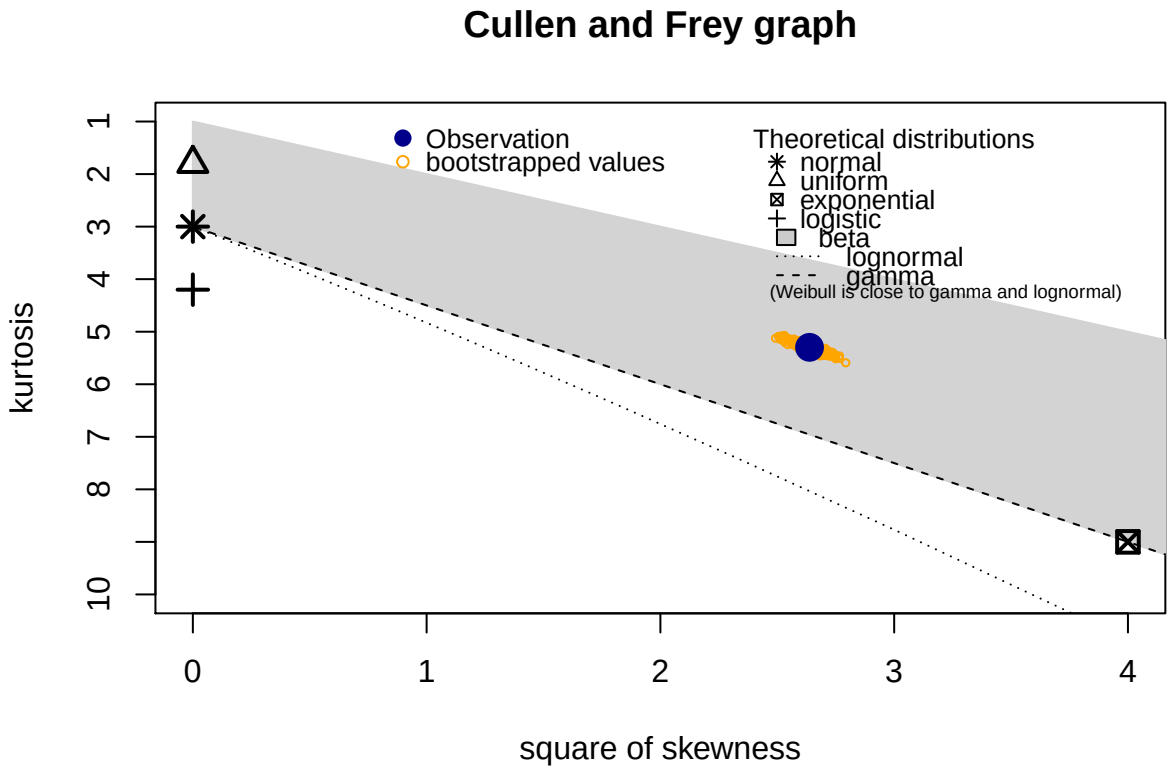
Histogram and Density Curve

```
truehist(Pr3c$EDUC,col="gainsboro", ylab="Frequency",
         xlab= "Years of Schooling", main= "Histogram of Education")
lines(density((Pr3c$EDUC)), lwd=2,col="firebrick1")
legend("topright", c("Relative Frequency", "Density Curve"),
       fill=c("gainsboro", "firebrick1"))
```



We plot the Cullen and Frey graph to give us an idea of what distributions to try out:

```
descdist(Pr3c$EDUC, boot = 1000)
```




```
## summary statistics
## ------
## min:  7   max:  18
## median:  10.5
## mean:  11.31889
## estimated sd:  2.322938
## estimated skewness:  1.62419
## estimated kurtosis:  5.297453
```

Observe that gamma, lognormal, and weibull distributions are possibilities.

Note that normal, uniform, and logistic distributions have to have the square of skewness equal to 0, so EDUC is unlikely to fit any of those distributions.

Also, observe that the exponential distribution has to have a square of skewness value equal to 4.

Since the data for EDUC has a square of skewness less than 3, the best distribution is very unlikely to be exponential, but we will attempt to test that fit as well.

We will attempt to fit various gamma distributions, a lognormal distribution, a weibull distribution, and an exponential distribution.

Testing fits for distributions

Testing fit for gamma distribution

```
EDUCgamma1 <- fitdist(Pr3c$EDUC, distr = "gamma", method = "mle",
                      lower = c(0, 0), start = list(scale = 1, shape = 1))
```

Testing fit for gamma distribution with different parameters

```
EDUCgamma2 <- fitdist(Pr3c$EDUC, distr = "gamma", method = "mle",
                      lower = c(0, 0), start = list(scale = 9, shape = 2))
```

Testing fit for gamma distribution with different parameters

```
EDUCgamma3 <- fitdist(Pr3c$EDUC, distr = "gamma", method = "mle",
                      lower = c(0, 0), start = list(scale = 0.5, shape = 1))
```

Testing fit for a lognormal distribution

```
EDUClnorm <- fitdist(Pr3c$EDUC, "lnorm")
```

Testing fit for a weibull distribution

```
EDUCweibull <- fitdist(Pr3c$EDUC, "weibull")
```

Testing for an exponential distribution

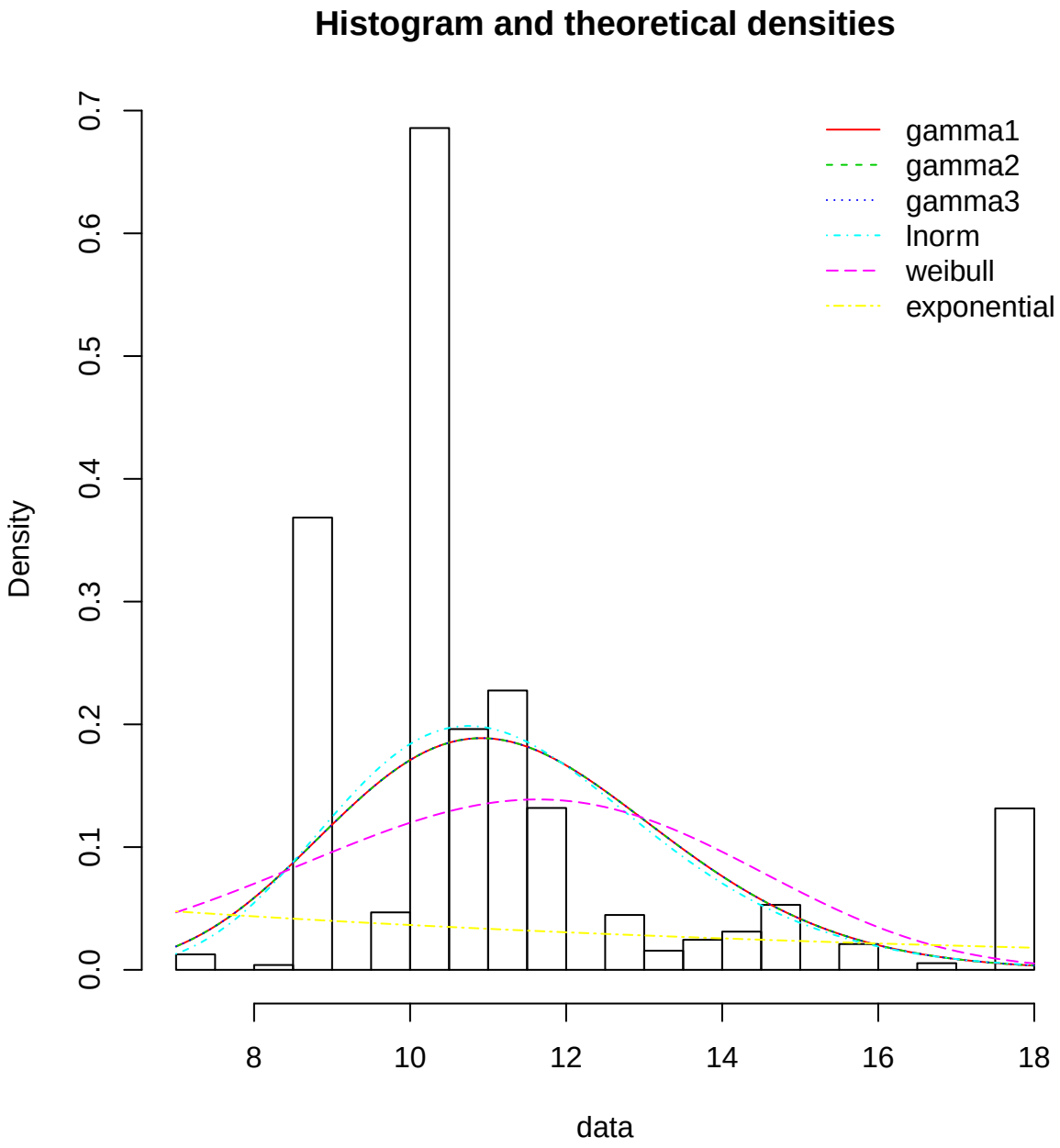
```
EDUCexp <- fitdist(Pr3c$EDUC, "exp")
```

Setting Legend

```
plot.legend <- c("gamma1", "gamma2", "gamma3", "lnorm", "weibull", "exponential")
```

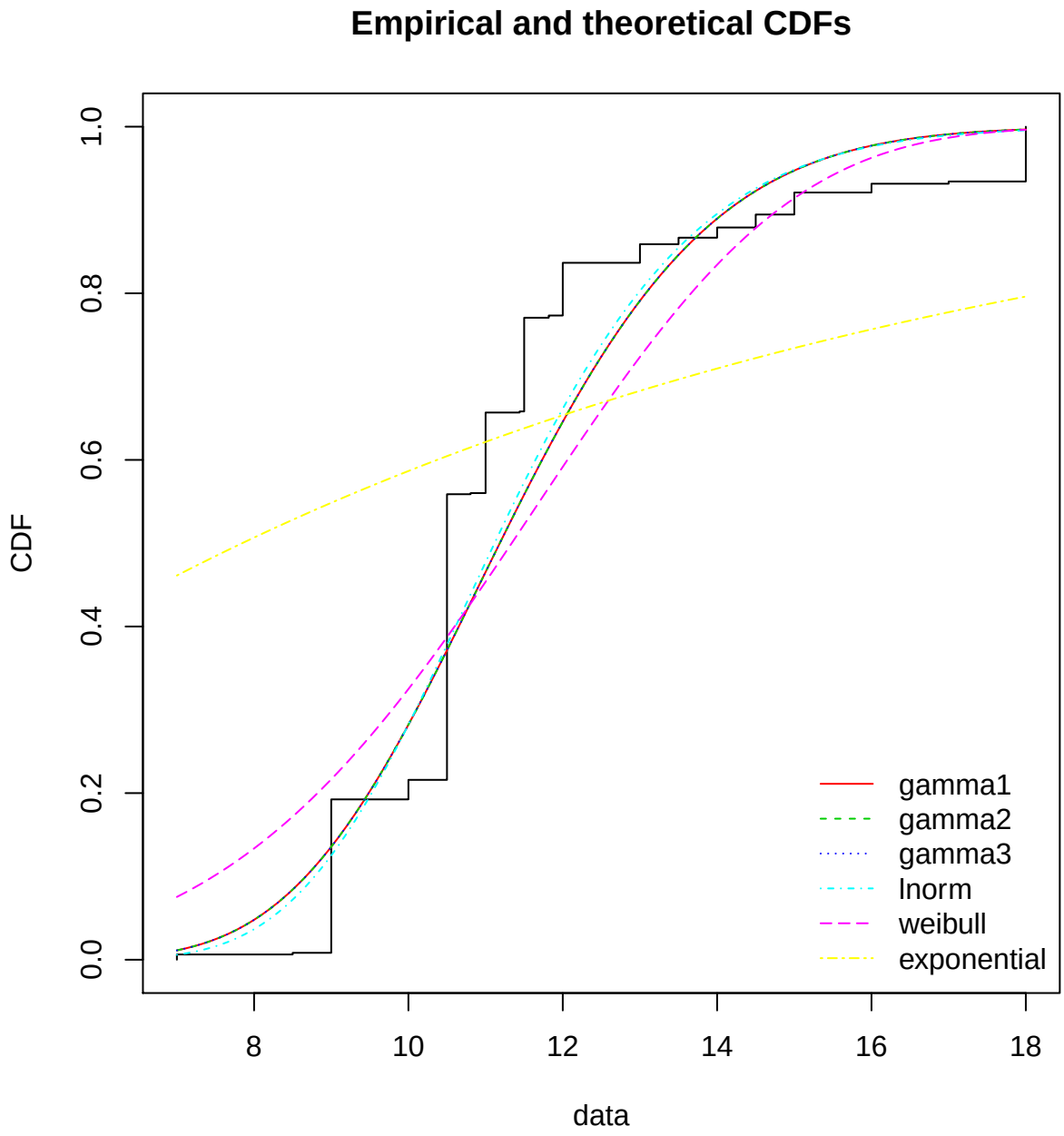
We compare the histogram with the theoretical densities on the following page.

```
denscomp(list(EDUCgamma1, EDUCgamma2, EDUCgamma3, EDUClnorm, EDUCweibull,
              EDUCexp), legendtext = plot.legend)
```



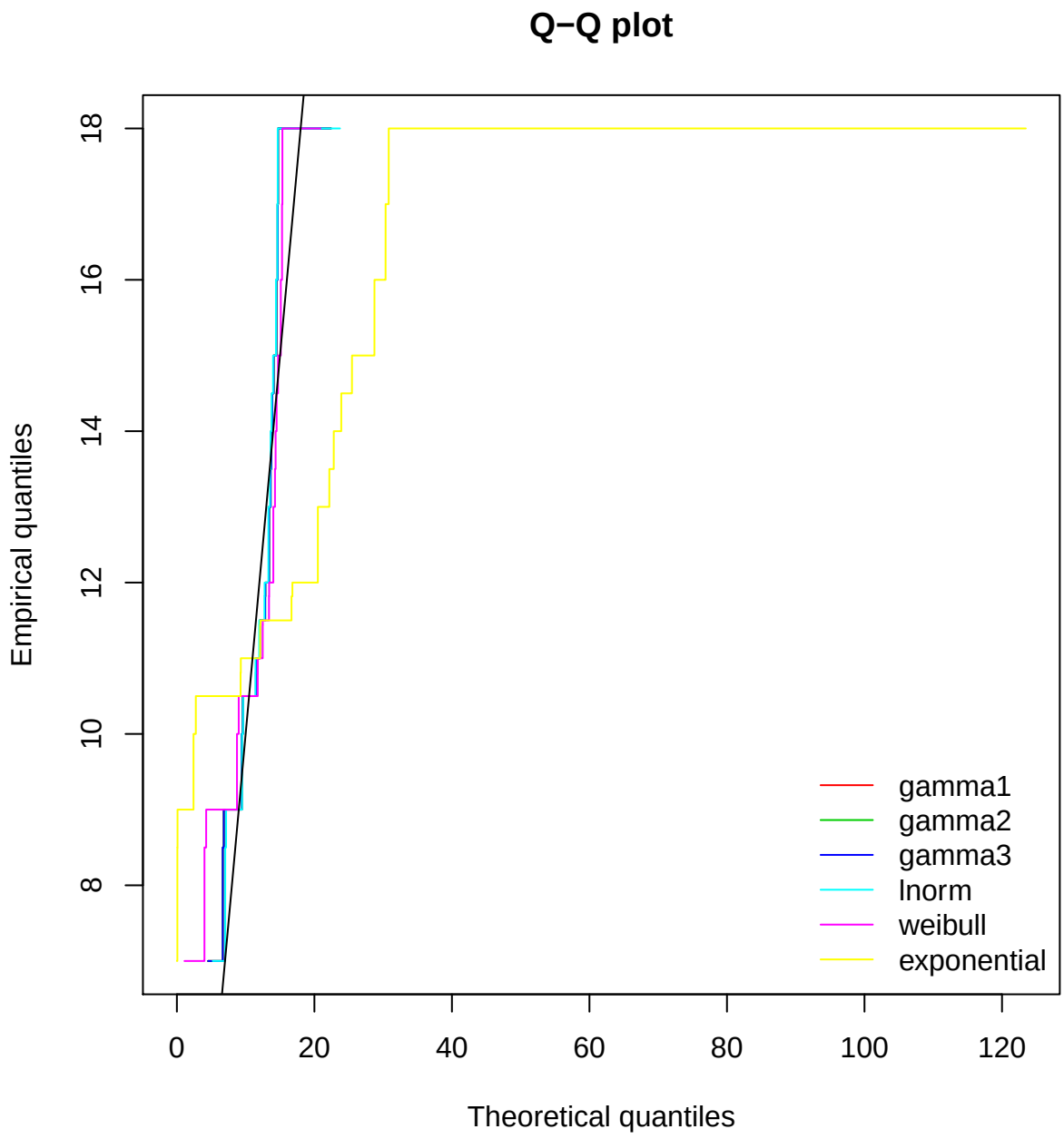
Observe that the lognormal distribution and gamma distribution with scale = 0.5, shape = 1 appear to be the best fits based on the theoretical densities of the distributions.

```
cdfcomp(list(EDUCgamma1, EDUCgamma2, EDUCgamma3, EDUClnorm, EDUCweibull,
             EDUCexp), legendtext = plot.legend)
```



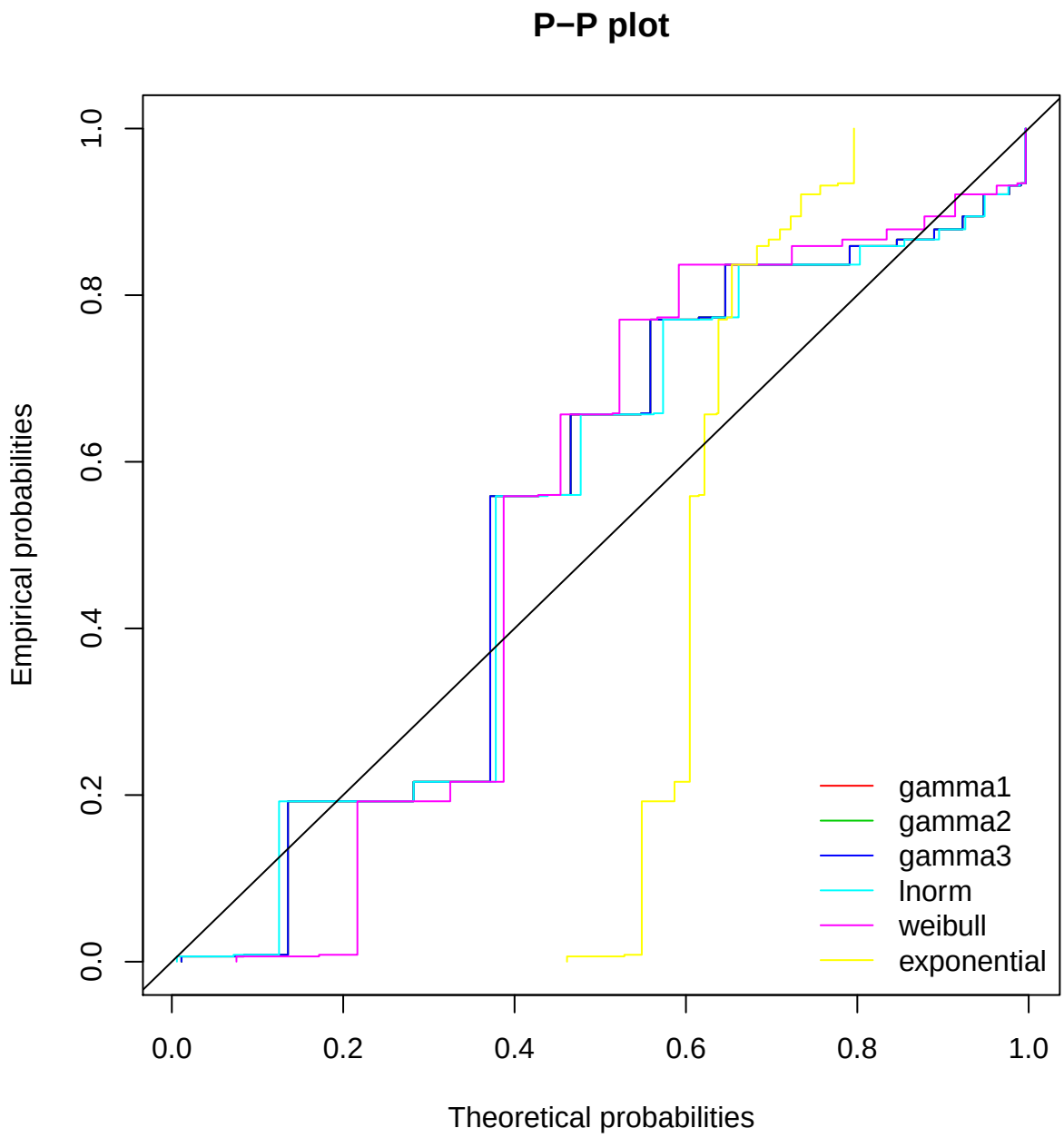
Again, the lognormal distribution and gamma distribution with scale = 0.5, shape = 1 appear to be the best fits.

```
qqcomp(list(EDUCgamma1, EDUCgamma2, EDUCgamma3, EDUClnorm, EDUCweibull,
            EDUCexp), legendtext = plot.legend)
```



The Q-Q plot shows that the exponential distribution is a particularly bad fit. Also, it shows the lognormal distribution, gamma distribution with scale = 0.5, shape = 1, and the weibull distribution being good fits.

```
ppcomp(list(EDUCgamma1, EDUCgamma2, EDUCgamma3, EDUClnorm, EDUCweibull,
            EDUCexp), legendtext = plot.legend)
```



The P-P plot shows the lognormal distribution being a particular good fit compared to the others.

Due to some ambiguity, we analyze goodness-of-fit statistics for each of the fitted distributions

```
gofstat(list(EDUCgamma1, EDUCgamma2, EDUCgamma3, EDUClnorm, EDUCweibull,
             EDUCexp),fitnames=c("gamma1", "gamma2", "gamma3", "lnorm",
                                 "weibull", "exponential"))
```

```
## Goodness-of-fit statistics
##                                    gamma1       gamma2       gamma3
## Kolmogorov-Smirnov statistic    0.2122468    0.2122505    0.2121694
## Cramer-von Mises statistic    307.0849509  307.0946548  306.9032057
## Anderson-Darling statistic   1710.7858201 1710.7968552 1710.2924224
##                                     lnorm      weibull  exponential
## Kolmogorov-Smirnov statistic    0.1973506    0.2482892    0.5401988
## Cramer-von Mises statistic    270.7879289  465.2048309 1894.2090236
## Anderson-Darling statistic   1515.2187874 2521.2188269 8793.7217421
##
## Goodness-of-fit criteria
##                                  gamma1   gamma2   gamma3    lnorm
## Akaike's Information Criterion 118486.5 118486.5 118486.5 116436.5
## Bayesian Information Criterion 118502.9 118502.9 118502.9 116452.9
##                                 weibull exponential
## Akaike's Information Criterion 128620.1    187066.8
## Bayesian Information Criterion 128636.6    187075.1
```

The goodness-of-fit statistics overwhelmingly support the lognormal distribution as being the best fitted distribution.

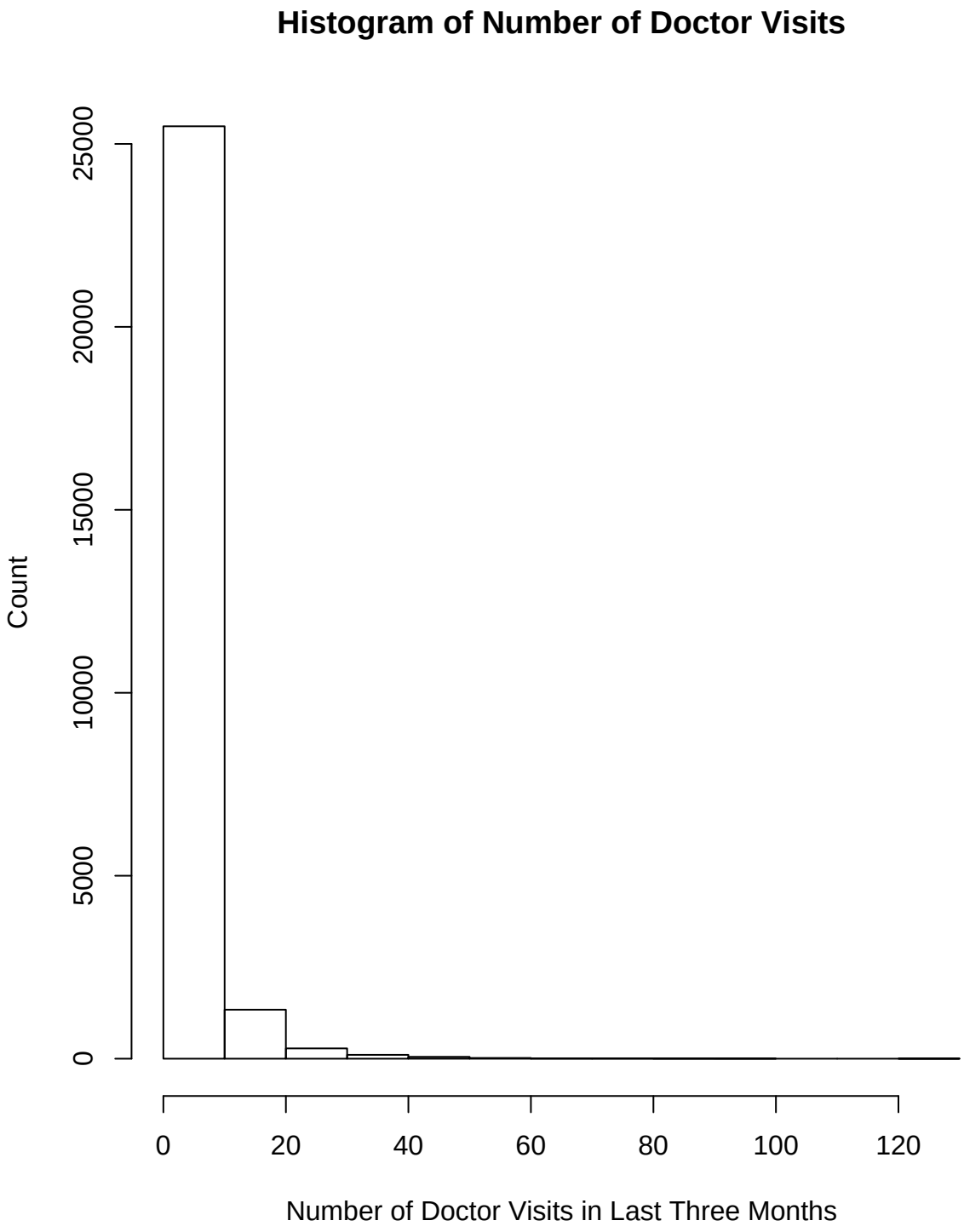
Conclusion about EDUC

We conclude that the EDUC variable is best approximated by a lognormal distribution.

Histogram and Density Curve for DOCVIS

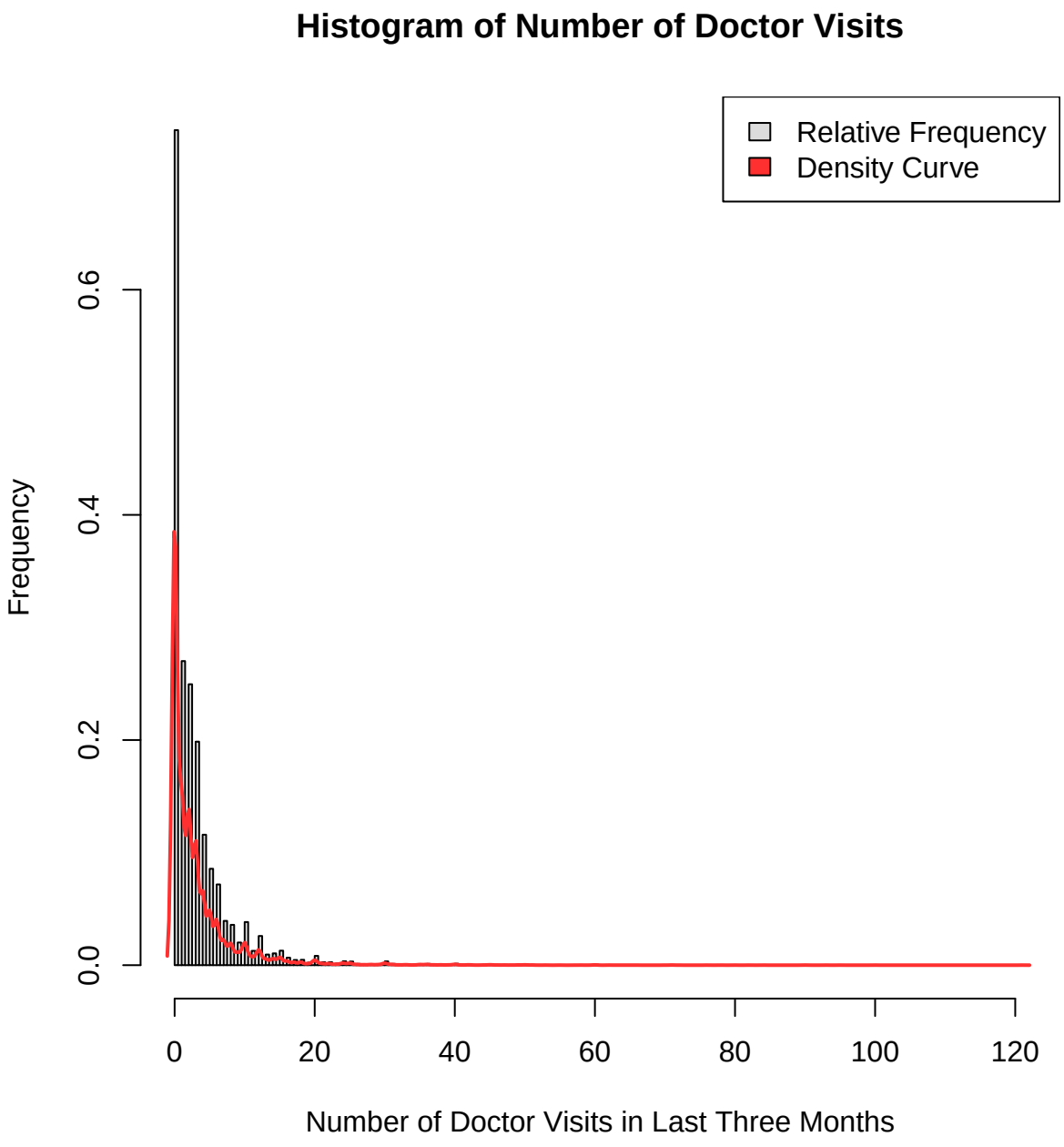
Histogram

```
hist(Pr3c$DOCVIS, xlab= "Number of Doctor Visits in Last Three Months",
     ylab= "Count", main= "Histogram of Number of Doctor Visits")
```



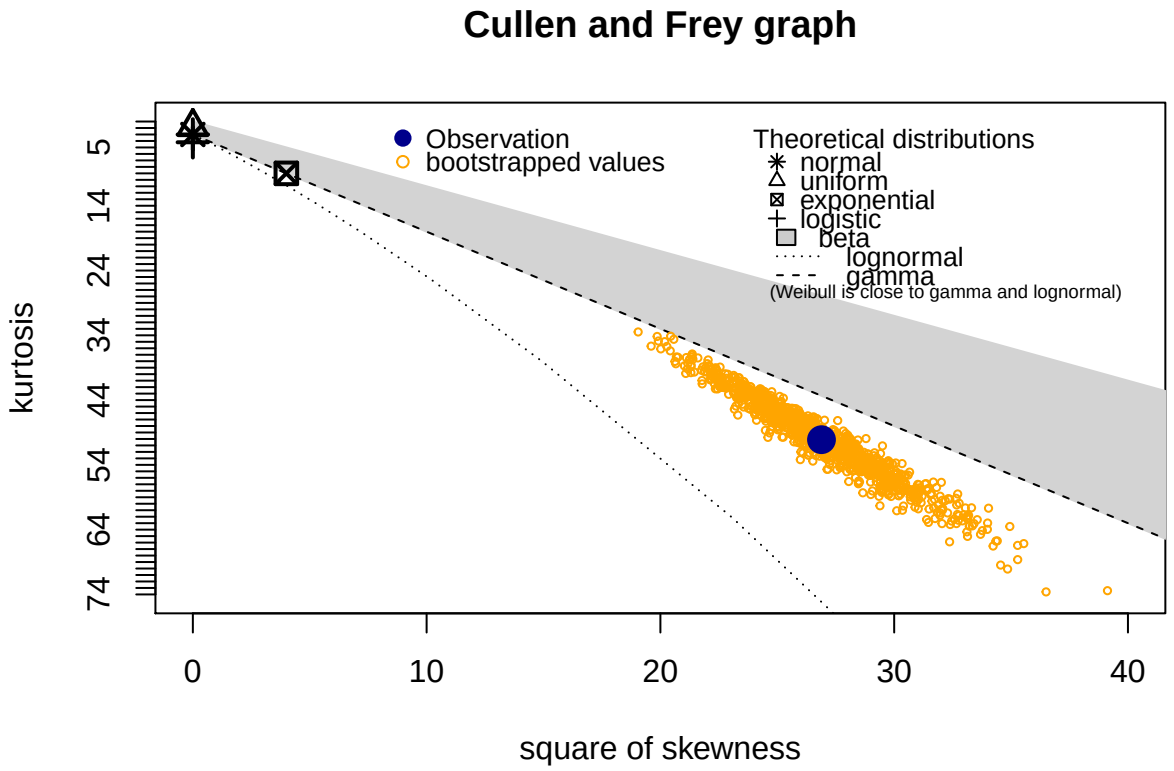
Histogram and Density Curve

```
truehist(Pr3c$DOCVIS,col="gainsboro", ylab="Frequency",
         xlab= "Number of Doctor Visits in Last Three Months",
         main= "Histogram of Number of Doctor Visits")
lines(density((Pr3c$DOCVIS)), lwd=2,col="firebrick1")
legend("topright", c("Relative Frequency", "Density Curve"),
       fill=c("gainsboro", "firebrick1"))
```



We plot the Cullen and Frey graph to give us an idea of what distributions to try out:

```
descdist(Pr3c$DOCVIS, boot = 1000)
```




```
## summary statistics
## ------
## min:  0   max:  121
## median:  1
## mean:  3.184233
## estimated sd:  5.69098
## estimated skewness:  5.185623
## estimated kurtosis:  50.11398
```

Observe that gamma distributions are possibilities. Due to values of zero, a lognormal distribution is not possible.

We will attempt to fit various gamma distributions.

Testing fits for distributions

Testing fit for gamma distribution

```
DOCVISgamma <- fitdist(Pr3c$DOCVIS, distr = "gamma", method = "mme")
```

Testing fit for gamma distribution with different parameters

```
DOCVISgamma1 <- fitdist(Pr3c$DOCVIS, distr = "gamma", method = "mle",
                        lower = c(0, 0), start = list(scale = 1, shape = 1))
```

Testing fit for gamma distribution with different parameters

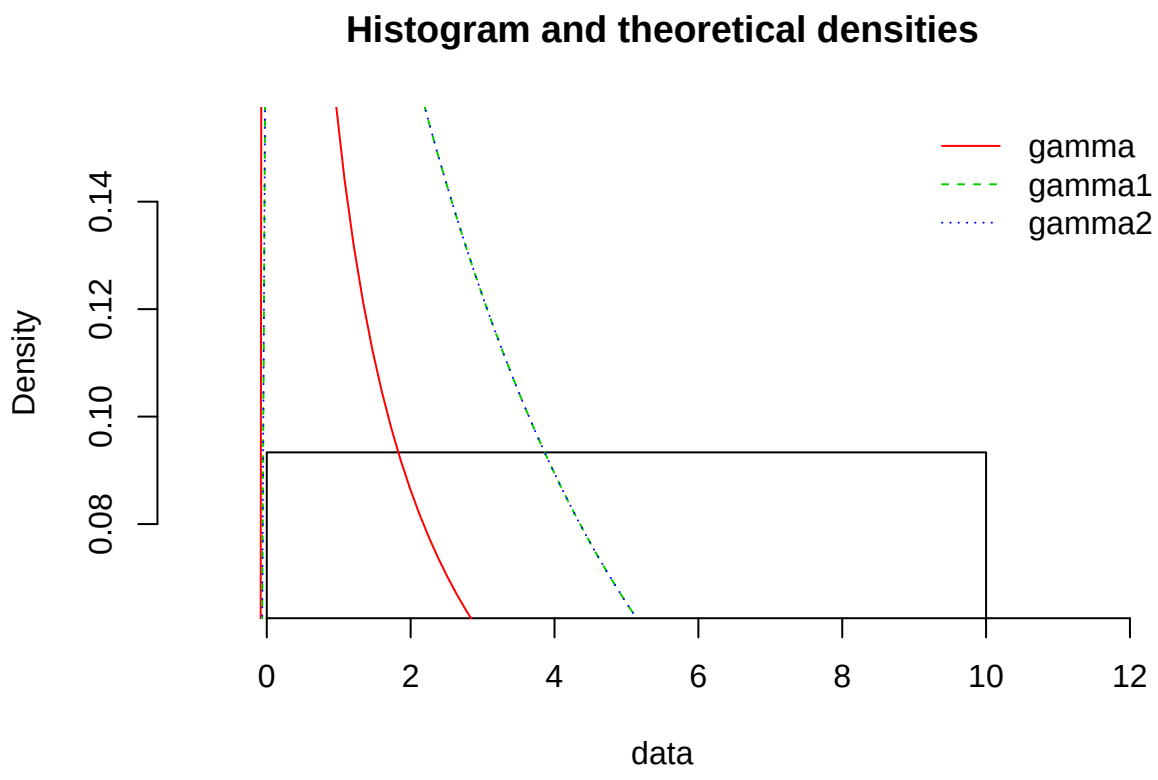
```
DOCVISgamma2 <- fitdist(Pr3c$DOCVIS, distr = "gamma", method = "mle",
                        lower = c(0, 0), start = list(scale = 0.5, shape = 1))
```

Setting Legend

```
plot.legend <- c("gamma", "gamma1", "gamma2")
```

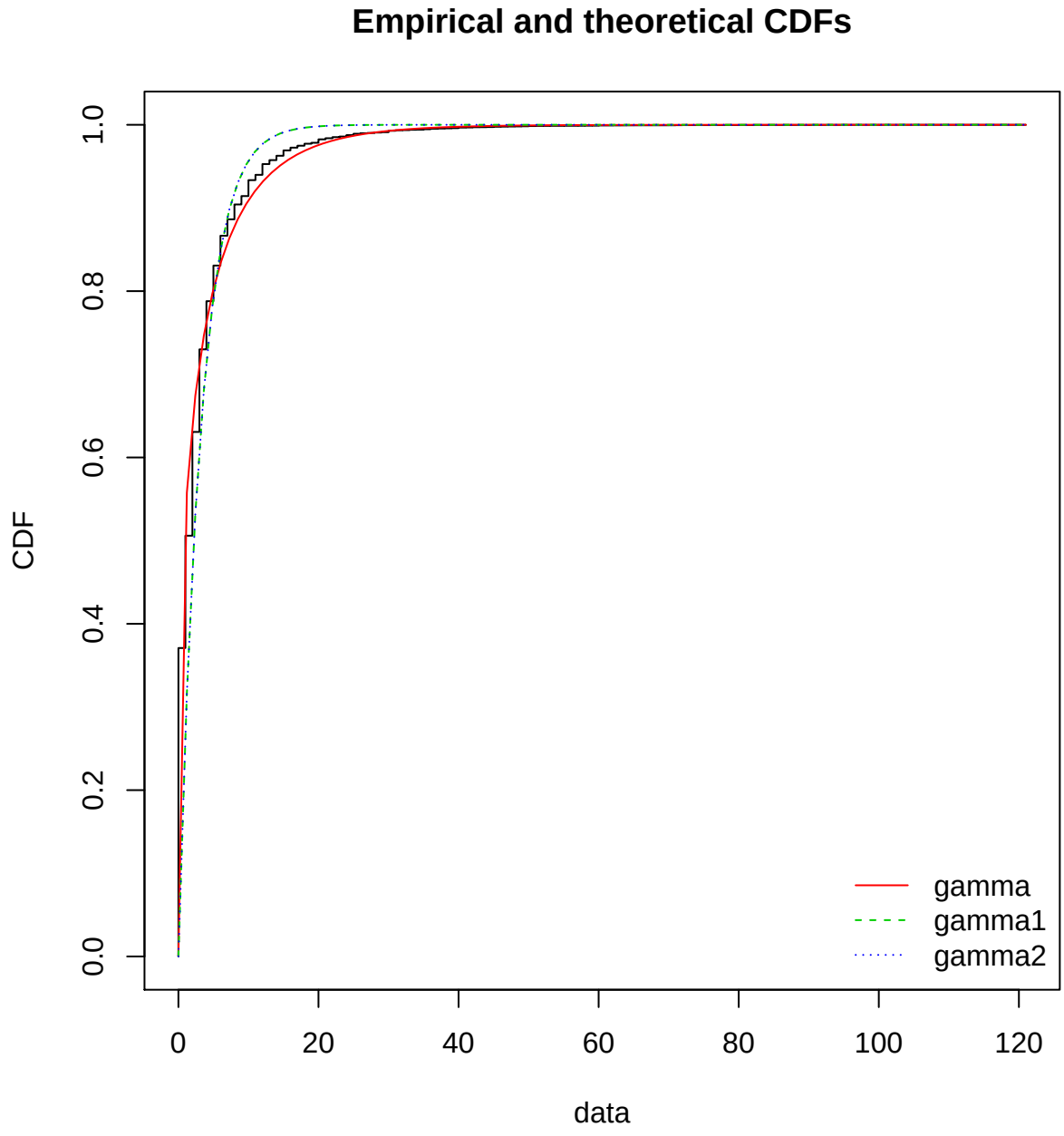
Comparing Histogram and Theoretical Densities

```
denscomp(list(DOCVISgamma, DOCVISgamma1, DOCVISgamma2),
         legendtext = plot.legend, ylim = 0.11, xlim = c(-1, 12))
```



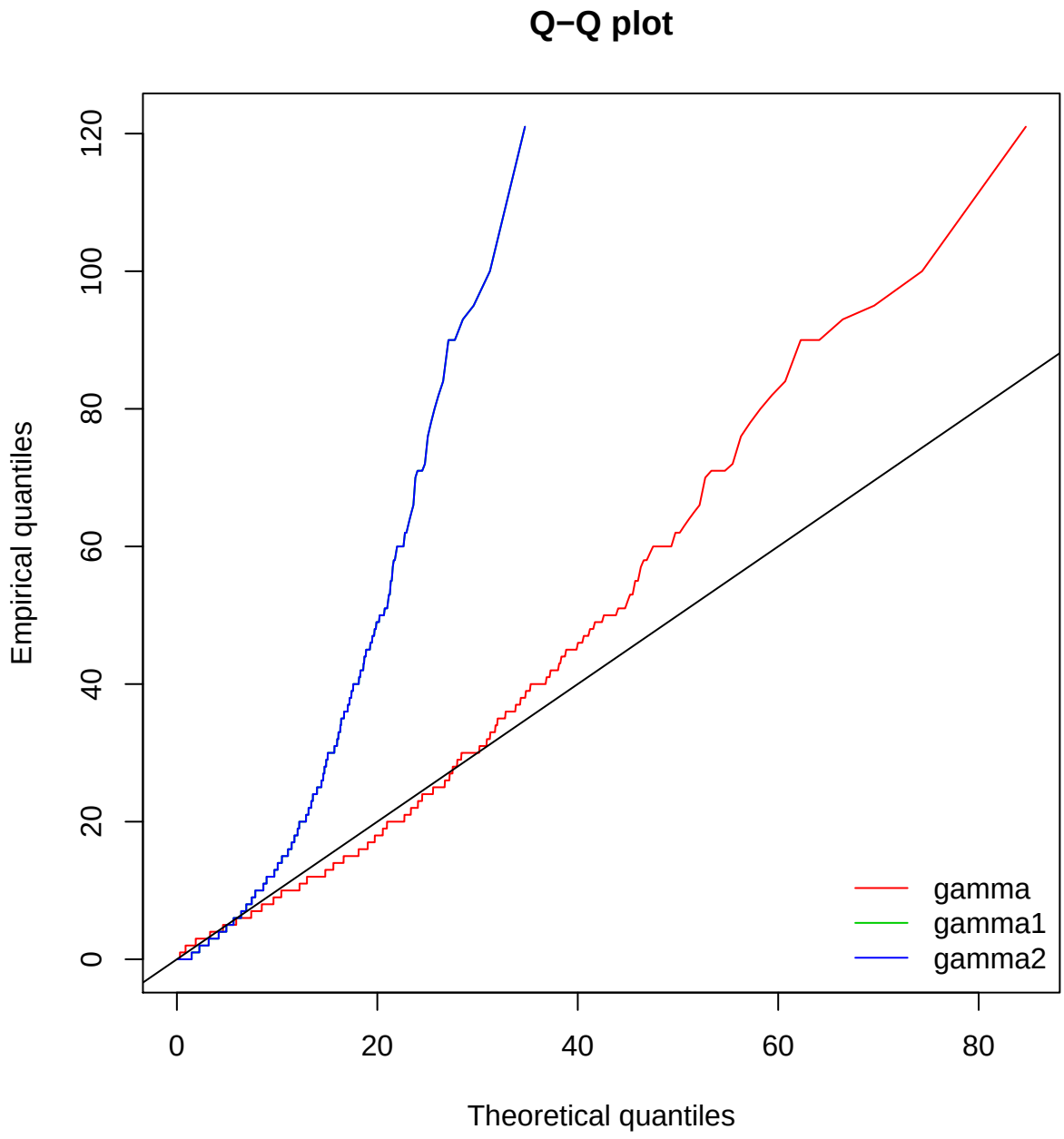
Observe that it is difficult to reach a conclusion from this plot.

```
cdfcomp(list(DOCVISgamma, DOCVISgamma1, DOCVISgamma2), legendtext = plot.legend)
```



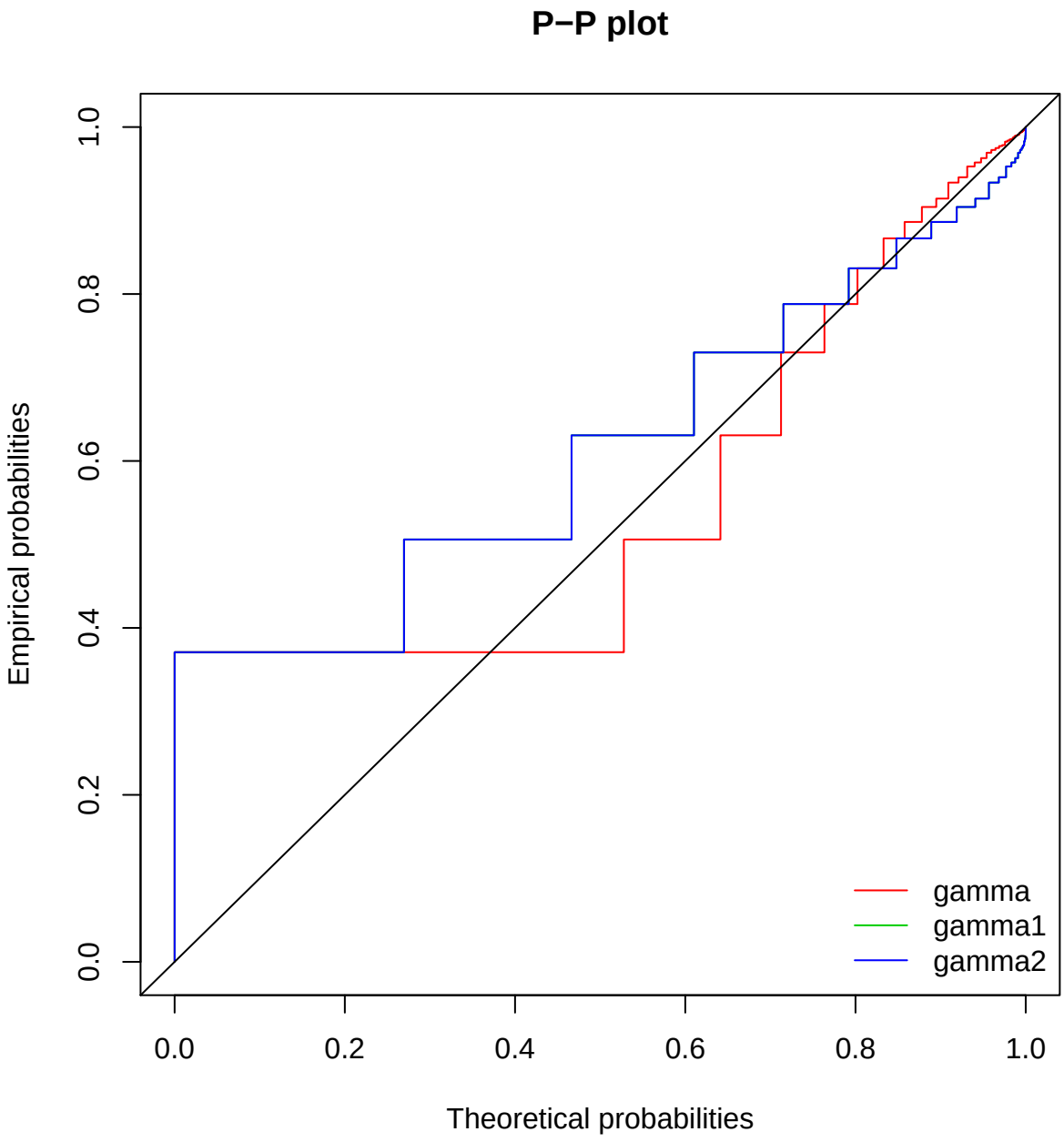
The gamma distributions appear to be good fits.

```
qqcomp(list(DOCVISgamma, DOCVISgamma1, DOCVISgamma2), legendtext = plot.legend)
```



The Q-Q plot shows that one of the gamma distributions is a better fit than the others.

```
ppcomp(list(DOCVISgamma, DOCVISgamma1, DOCVISgamma2), legendtext = plot.legend)
```



The P-P plot shows that again one of the gamma distributions is a better fit than the others.

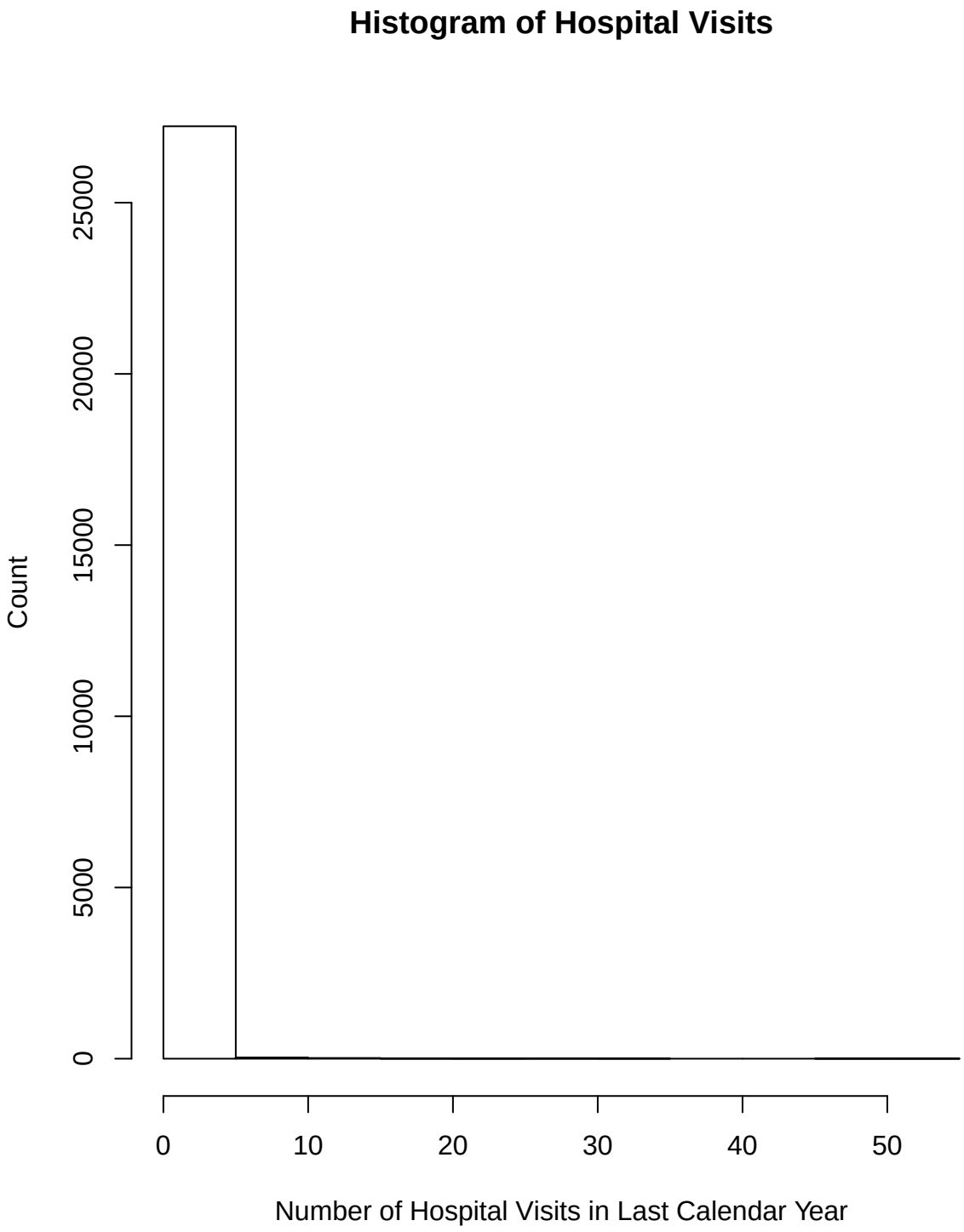
Conclusion about DOCVIS

We conclude that the DOCVIS variable is best approximated by a gamma distribution.

Histogram and Density Curve for HOSPVIS

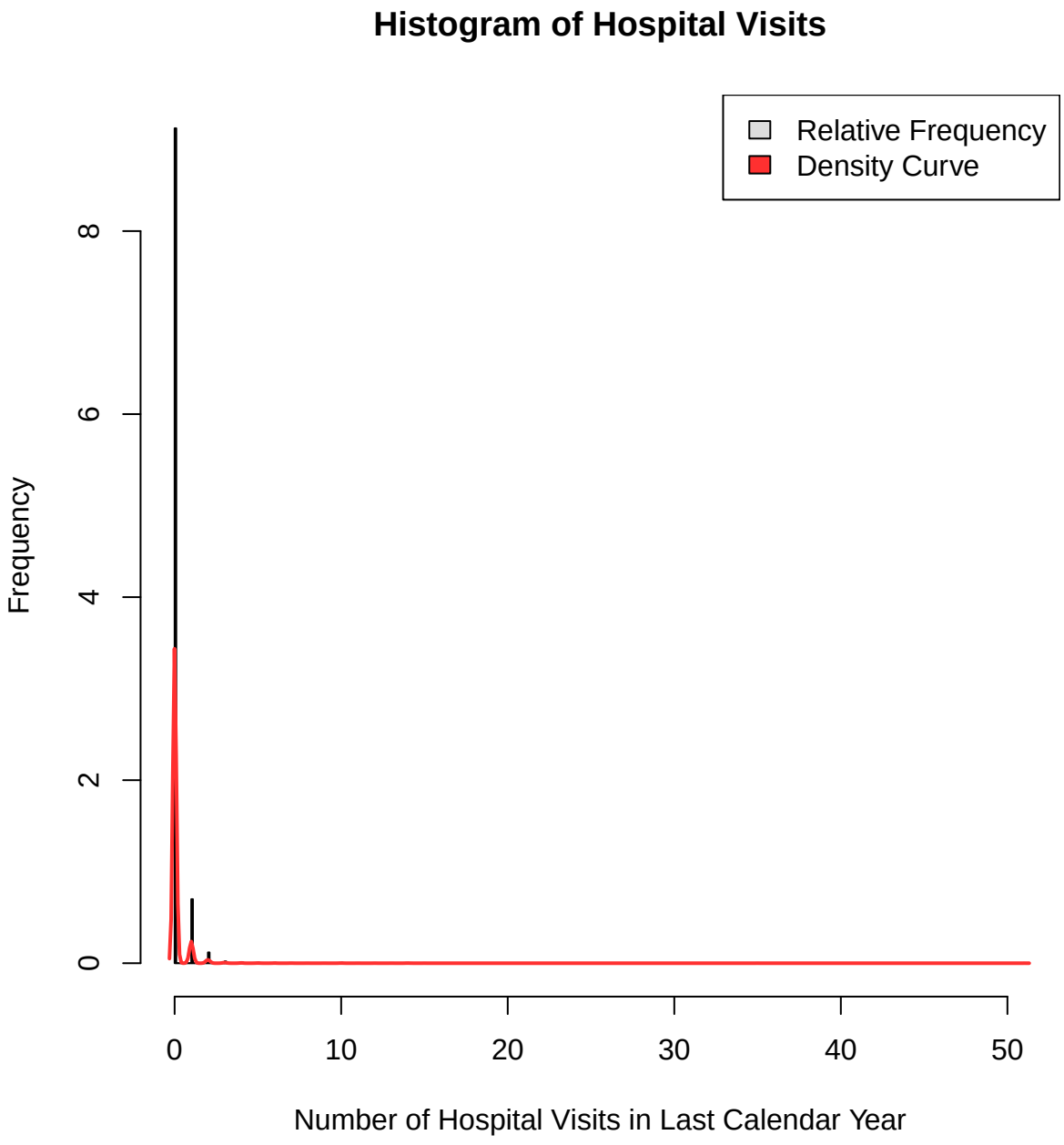
Histogram

```
hist(Pr3c$HOSPVIS, xlab= "Number of Hospital Visits in Last Calendar Year",
     ylab= "Count", main= "Histogram of Hospital Visits")
```



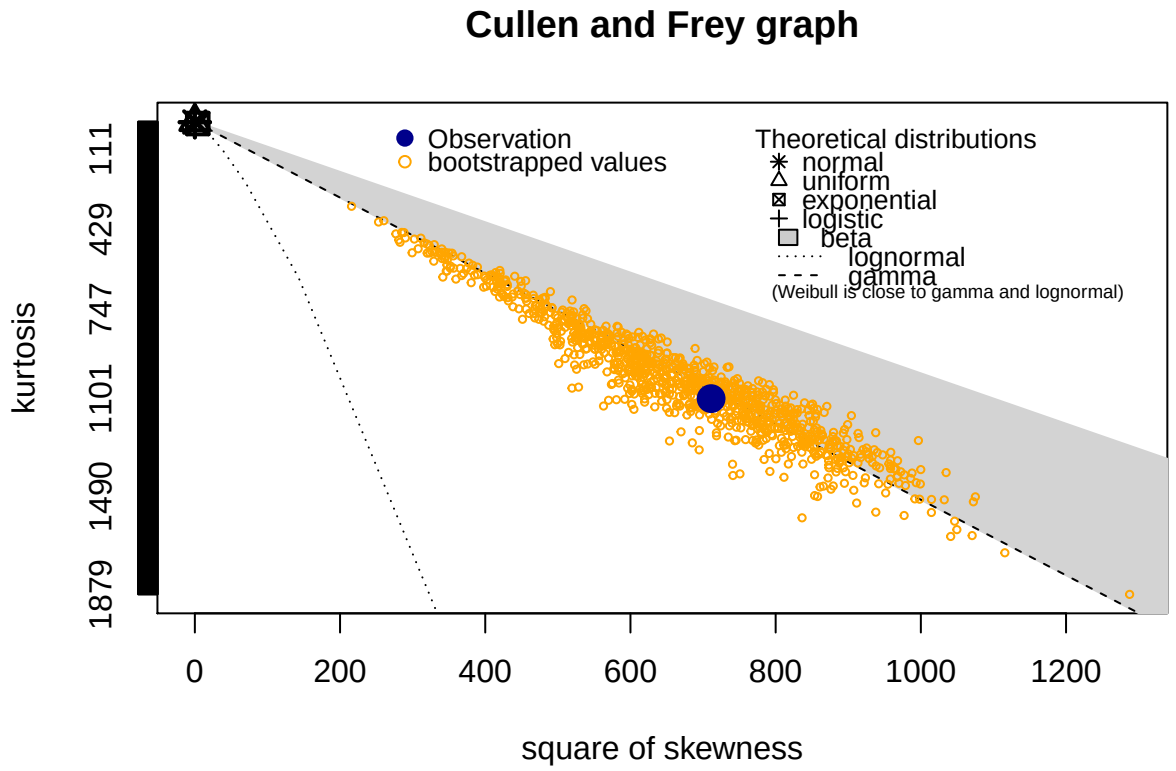
Histogram and Density Curve

```
truehist(Pr3c$HOSPVIS,col="gainsboro", ylab="Frequency",
         xlab= "Number of Hospital Visits in Last Calendar Year",
         main= "Histogram of Hospital Visits")
lines(density((Pr3c$HOSPVIS)), lwd=2,col="firebrick1")
legend("topright", c("Relative Frequency", "Density Curve"),
       fill=c("gainsboro", "firebrick1"))
```



We plot the Cullen and Frey graph to give us an idea of what distributions to try out:

```
descdist(Pr3c$HOSPVIS, boot = 1000)
```




```
## summary statistics
## ------
## min:  0   max:  51
## median:  0
## mean:  0.1382936
## estimated sd:  0.8847108
## estimated skewness:  26.66635
## estimated kurtosis:  1101.437
```

Observe that gamma distributions are the likeliest possibilities.

Note that due to so many values being equal to zero based on the histogram, there is a possibility that a uniform distribution may be a good fit.

Note that a beta distribution does not make sense as the range of the values in the data is [0,51] and the data is based on the number of hospital visits and not on a probability.

We will attempt to fit various gamma distributions and a uniform distribution.

Testing fits for distributions

Testing fit for gamma distribution

```
HOSPVISgamma <- fitdist(Pr3c$HOSPVIS, distr = "gamma", method = "mme")
```

Testing fit for gamma distribution with different parameters

```
HOSPVISgamma1 <- fitdist(Pr3c$HOSPVIS, distr = "gamma", method = "mle",
                         lower = c(0, 0), start = list(scale = 1, shape = 1))
```

Testing fit for gamma distribution with different parameters

```
HOSPVISgamma2 <- fitdist(Pr3c$HOSPVIS, distr = "gamma", method = "mle",
                         lower = c(0, 0), start = list(scale = 0.1, shape = 1))
```

Testing fit for gamma distribution with different parameters

```
HOSPVISgamma3 <- fitdist(Pr3c$HOSPVIS, distr = "gamma", method = "mle",
                         lower = c(0, 0), start = list(scale = 0.5, shape = 1))
```

Testing fit for uniform distribution

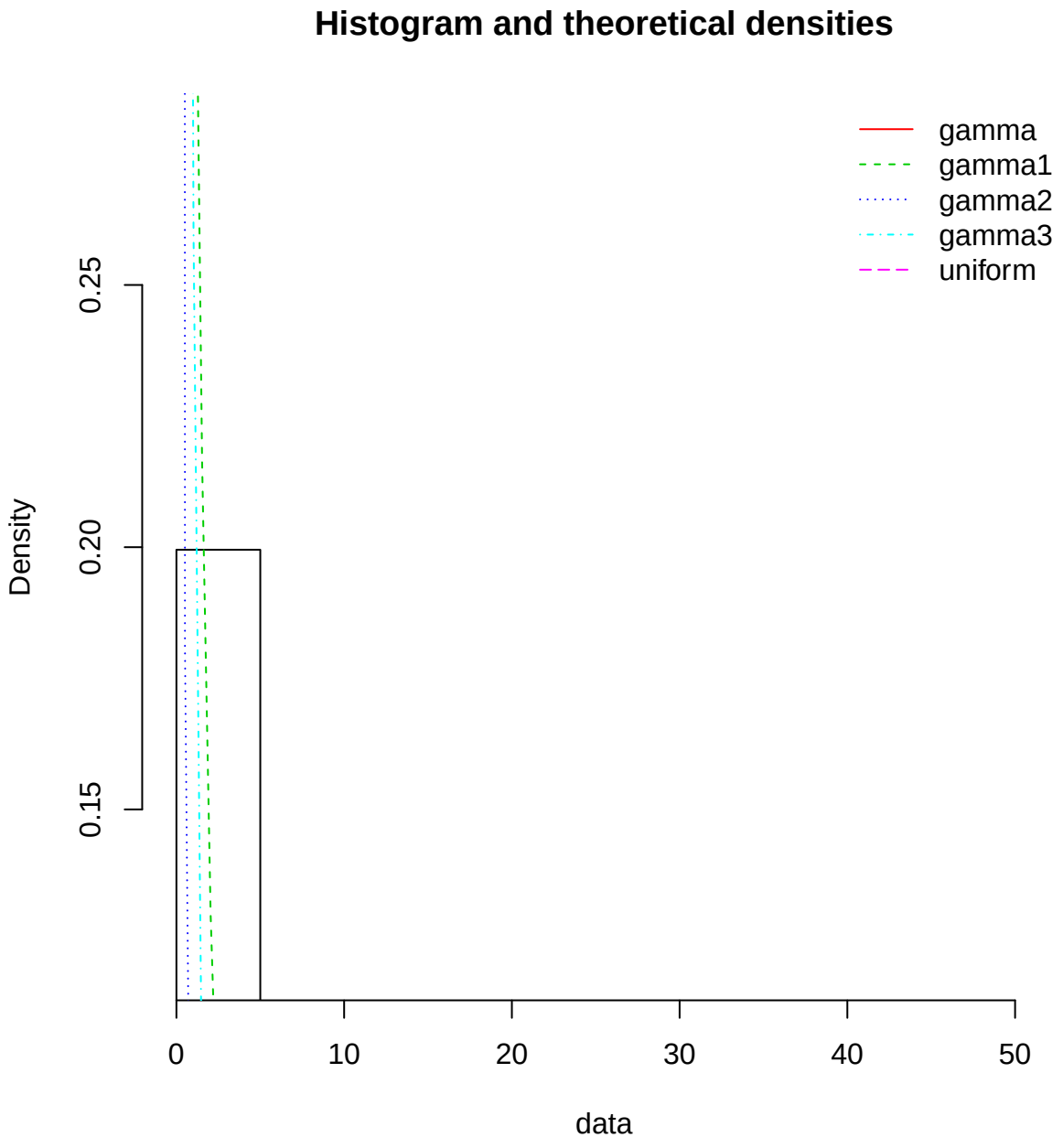
```
HOSPVISunif <- fitdist(Pr3c$HOSPVIS, distr = "unif")
```

Setting Legend

```
plot.legend <- c("gamma", "gamma1", "gamma2", "gamma3", "uniform")
```

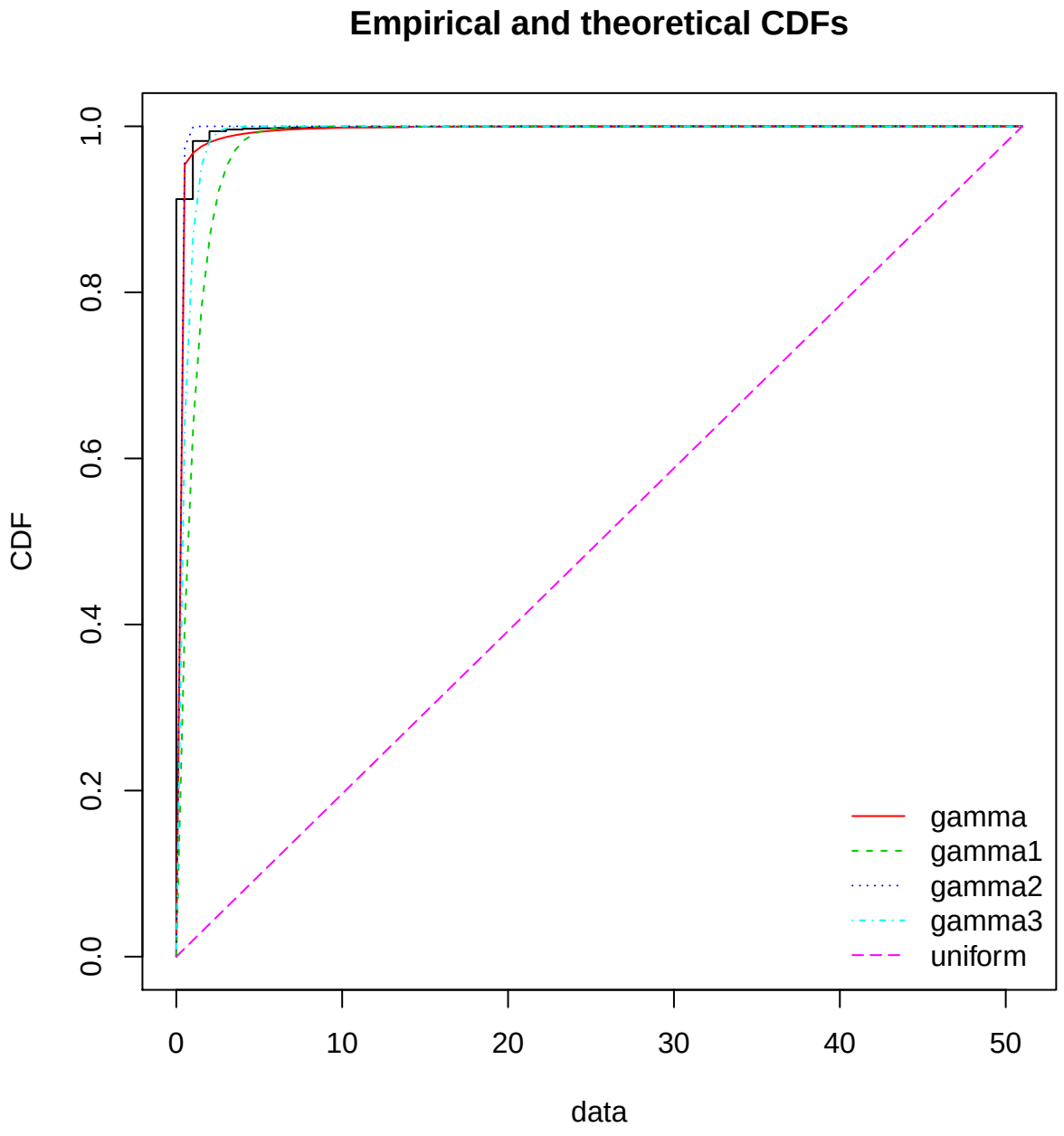
We compare the histogram with the theoretical densities on the following page.

```
denscomp(list(HOSPVISgamma, HOSPVISgamma1, HOSPVISgamma2, HOSPVISgamma3,
              HOSPVISunif), legendtext = plot.legend, ylim = 0.2)
```



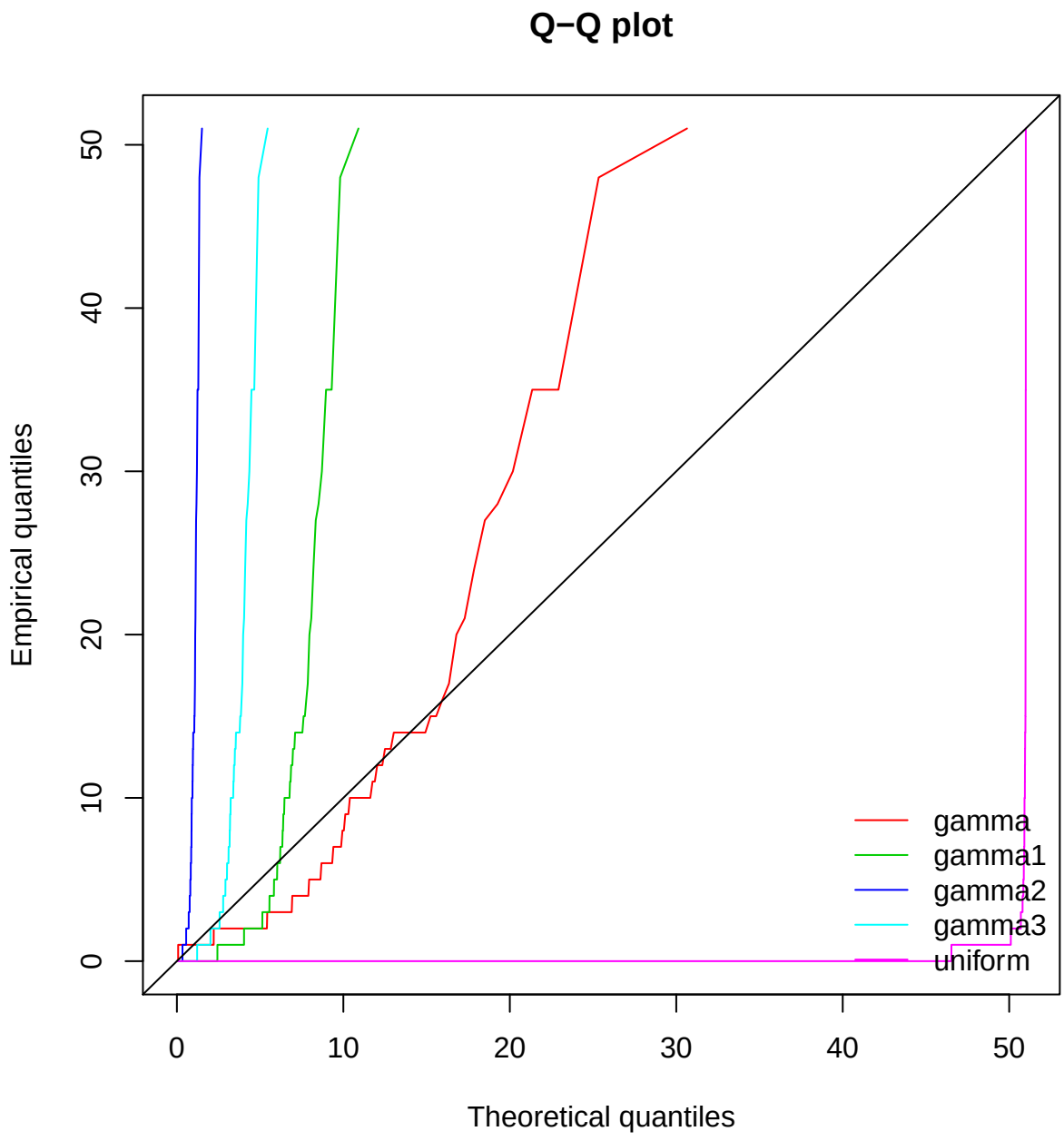
Observe that the graph does not tell us much.

```
cdfcomp(list(HOSPVISgamma, HOSPVISgamma1, HOSPVISgamma2,
             HOSPVISgamma3, HOSPVISunif), legendtext = plot.legend)
```



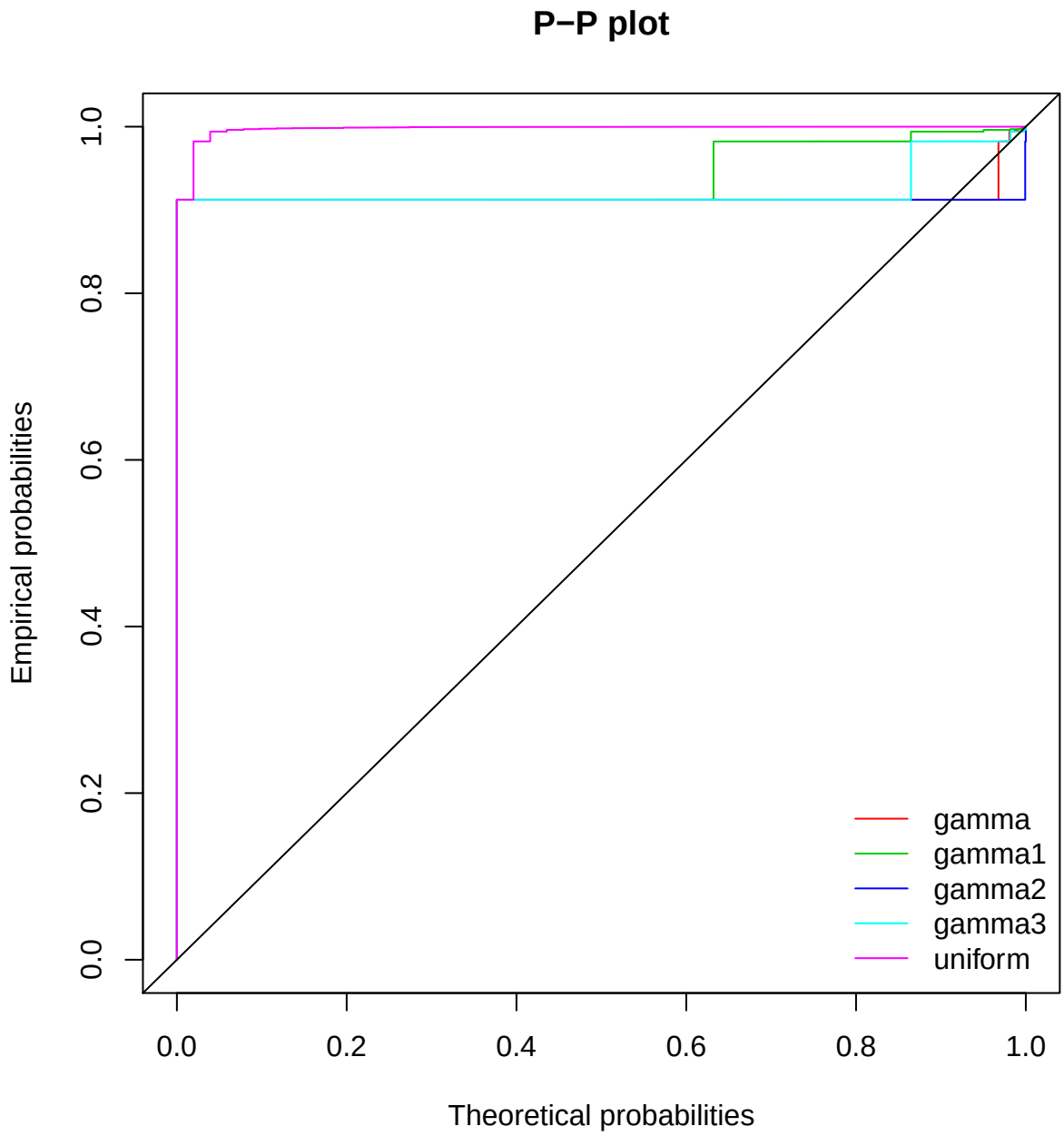
Observe that the gamma distributions do indeed appear to be good fits. The uniform distribution looks like a terrible fit.

```
qqcomp(list(HOSPVISgamma, HOSPVISgamma1, HOSPVISgamma2, HOSPVISgamma3,
            HOSPVISunif), legendtext = plot.legend)
```



The Q-Q plot contradicts our previous conclusion and shows that three of the four gamma distributions are not good fits.

```
ppcomp(list(HOSPVISgamma, HOSPVISgamma1, HOSPVISgamma2, HOSPVISgamma3,
            HOSPVISunif), legendtext = plot.legend)
```



The P-P plot shows that none of the gamma distributions are good fits.

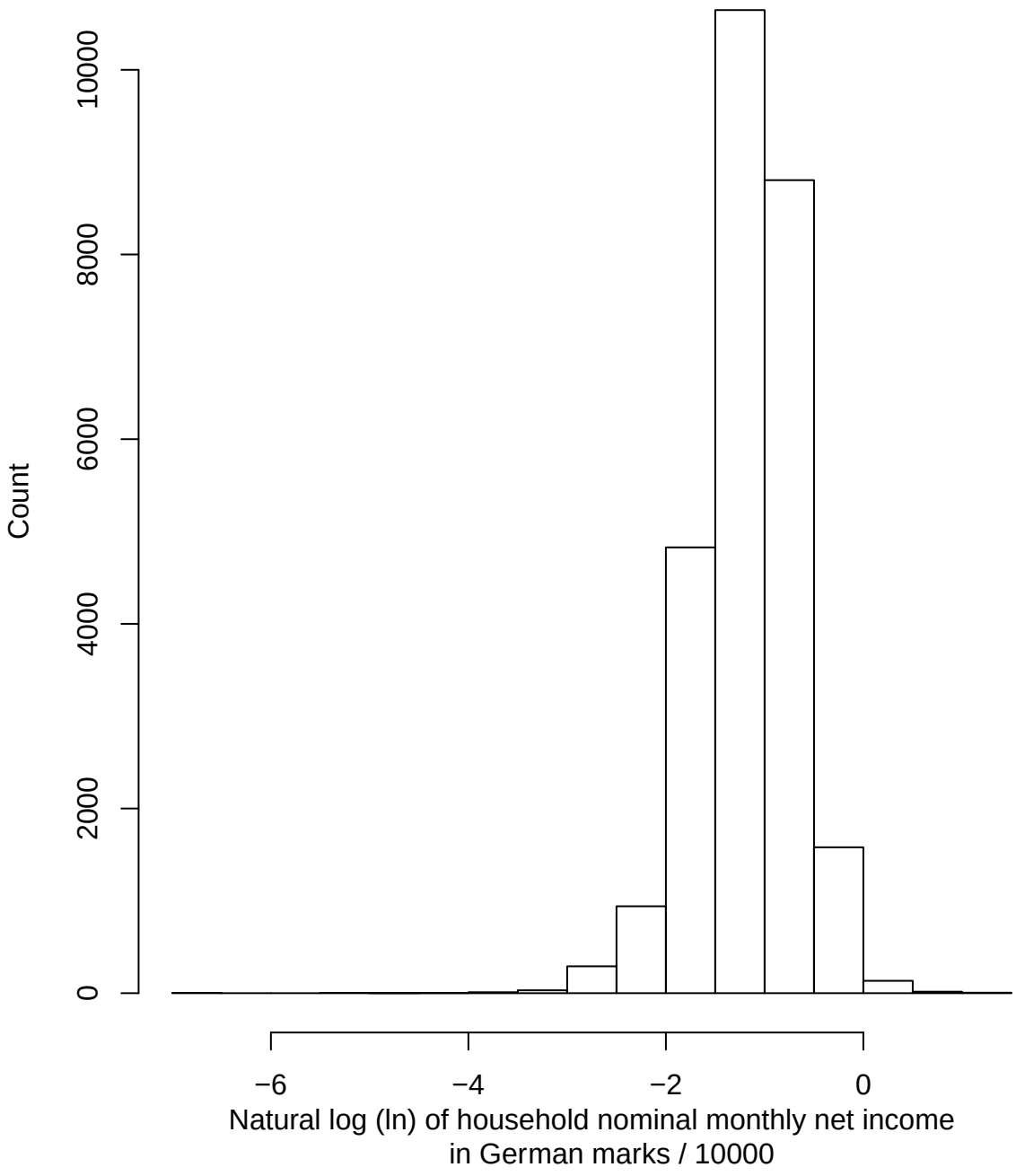
Conclusion about HOSPVIS

We conclude that the HOSPVIS variable is best approximated by a gamma distribution, based mainly on the CUllen and Frey graph and the comparison of empirical vs. theoretical CDFs.

Histogram and Density Curve for LOGINC

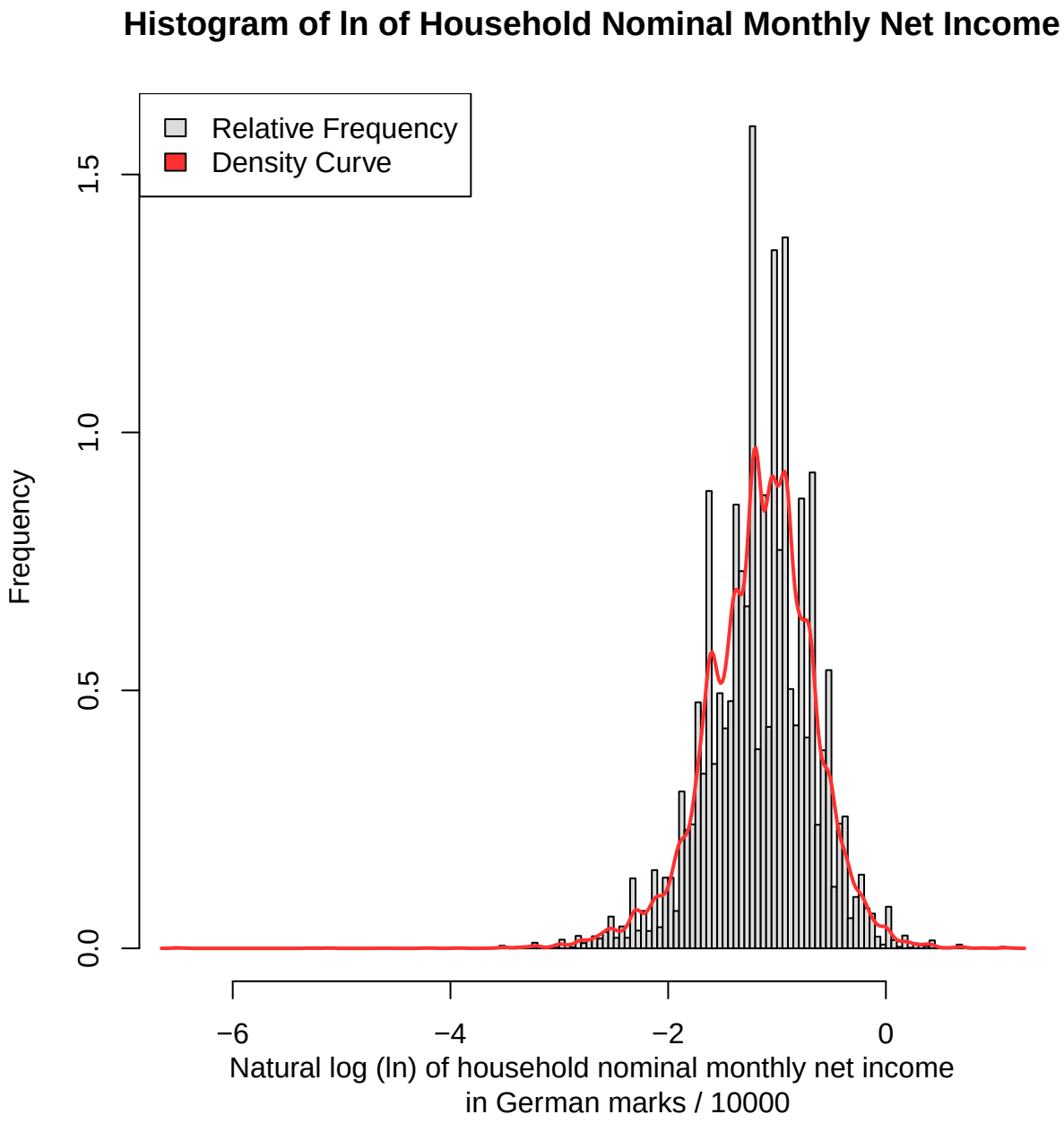
Histogram

```
hist(Pr3c$LOGINC,
     xlab= "Natural log (ln) of household nominal monthly net income
     in German marks / 10000", ylab= "Count",
     main= "Histogram of ln of Household Nominal Monthly Net Income")
```


Histogram of ln of Household Nominal Monthly Net Income
Count
0
2000
4000
6000
8000
10000
−6
−4
−2
0
Natural log (ln) of household nominal monthly net income
in German marks / 10000

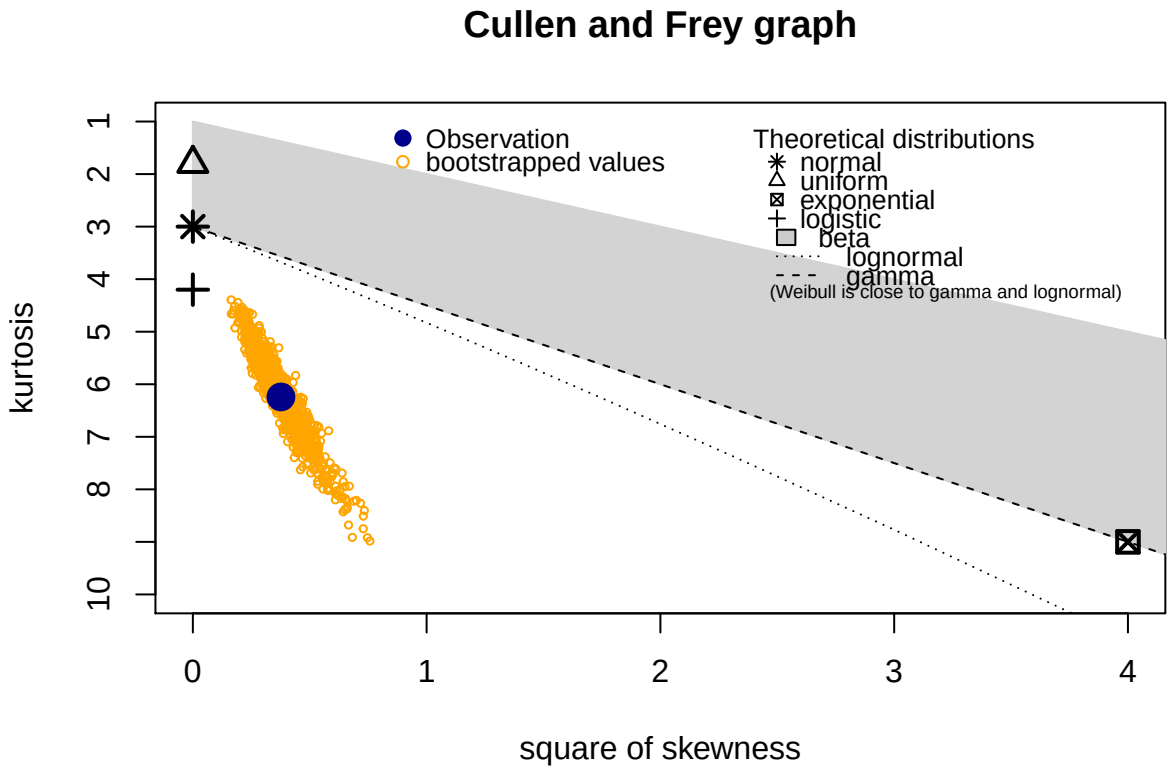
Histogram and Density Curve

```
truehist(Pr3c$LOGINC,col="gainsboro", ylab="Frequency",
         xlab= "Natural log (ln) of household nominal monthly net income
         in German marks / 10000",
         main= "Histogram of ln of Household Nominal Monthly Net Income")
lines(density((Pr3c$LOGINC)), lwd=2,col="firebrick1")
legend("topleft", c("Relative Frequency", "Density Curve"),
       fill=c("gainsboro", "firebrick1"))
```



We plot the Cullen and Frey graph to give us an idea of what distributions to try out:

```
descdist(Pr3c$LOGINC, boot = 1000)
```




```
## summary statistics
## ------
## min:  -6.50229   max:  1.120732
## median:  -1.139434
## mean:  -1.157179
## estimated sd:  0.4911825
## estimated skewness:  -0.6138659
## estimated kurtosis:  6.241342
```

Observe that a logistic distribution may be a good fit. We will also attempt to fit a normal distribution, based on the shape of the density curve. A lognormal distribution does not make sense as the data has negative values.

We will attempt to fit a logistic distribution, a lognormal distribution, and a normal distribution.

Testing fits for distributions

Testing fit for a logistic distribution

```
LOGINClogis <- fitdist(Pr3c$LOGINC, "logis")
```

Testing for a normal distribution

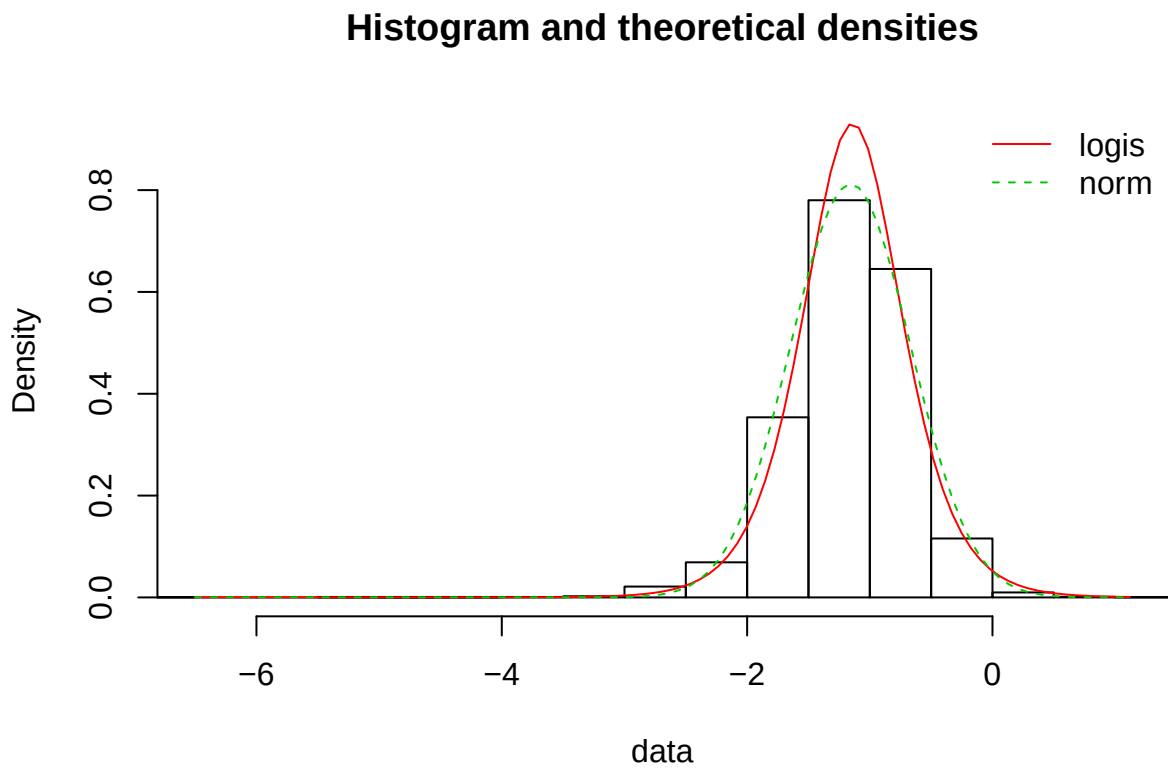
```
LOGINCnorm <- fitdist(Pr3c$LOGINC, "norm")
```

Setting Legend

```
plot.legend <- c("logis", "norm")
```

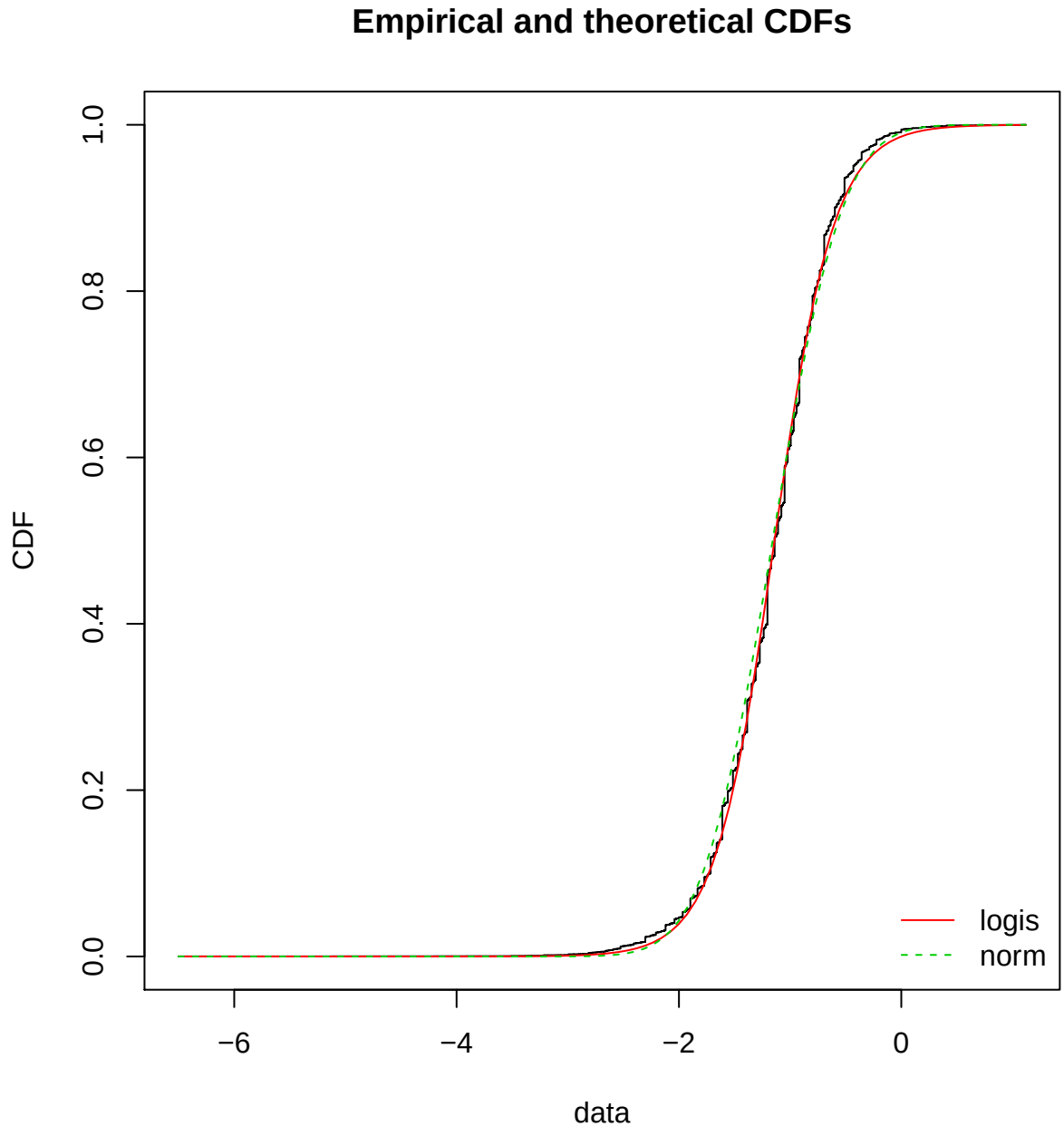
Comparing Histogram and Theoretical Densities

```
denscomp(list(LOGINClogis, LOGINCnorm), legendtext = plot.legend)
```



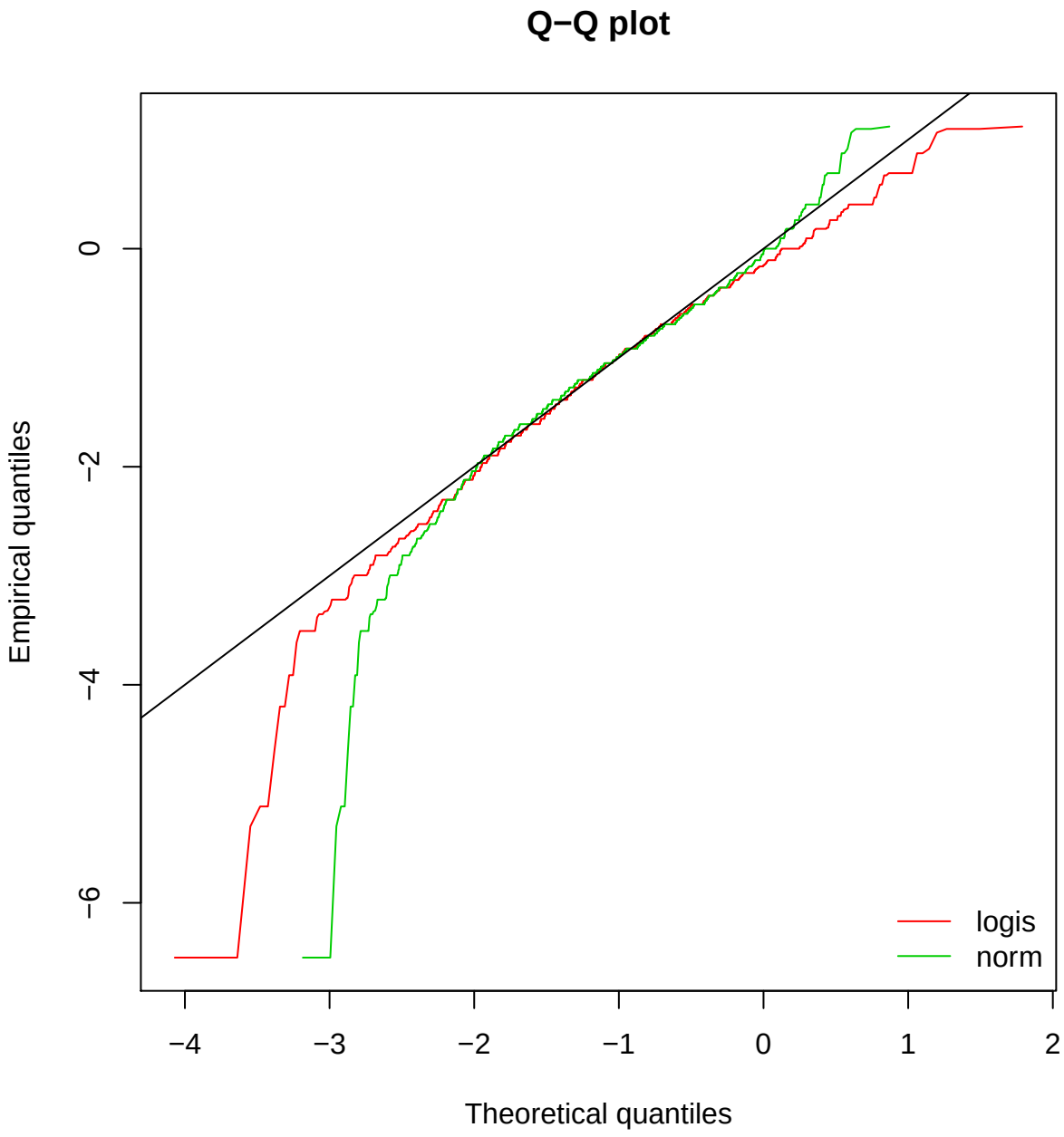
Observe that the normal distribution seems to be the best fit.

```
cdfcomp(list(LOGINClogis, LOGINCnorm), legendtext = plot.legend)
```



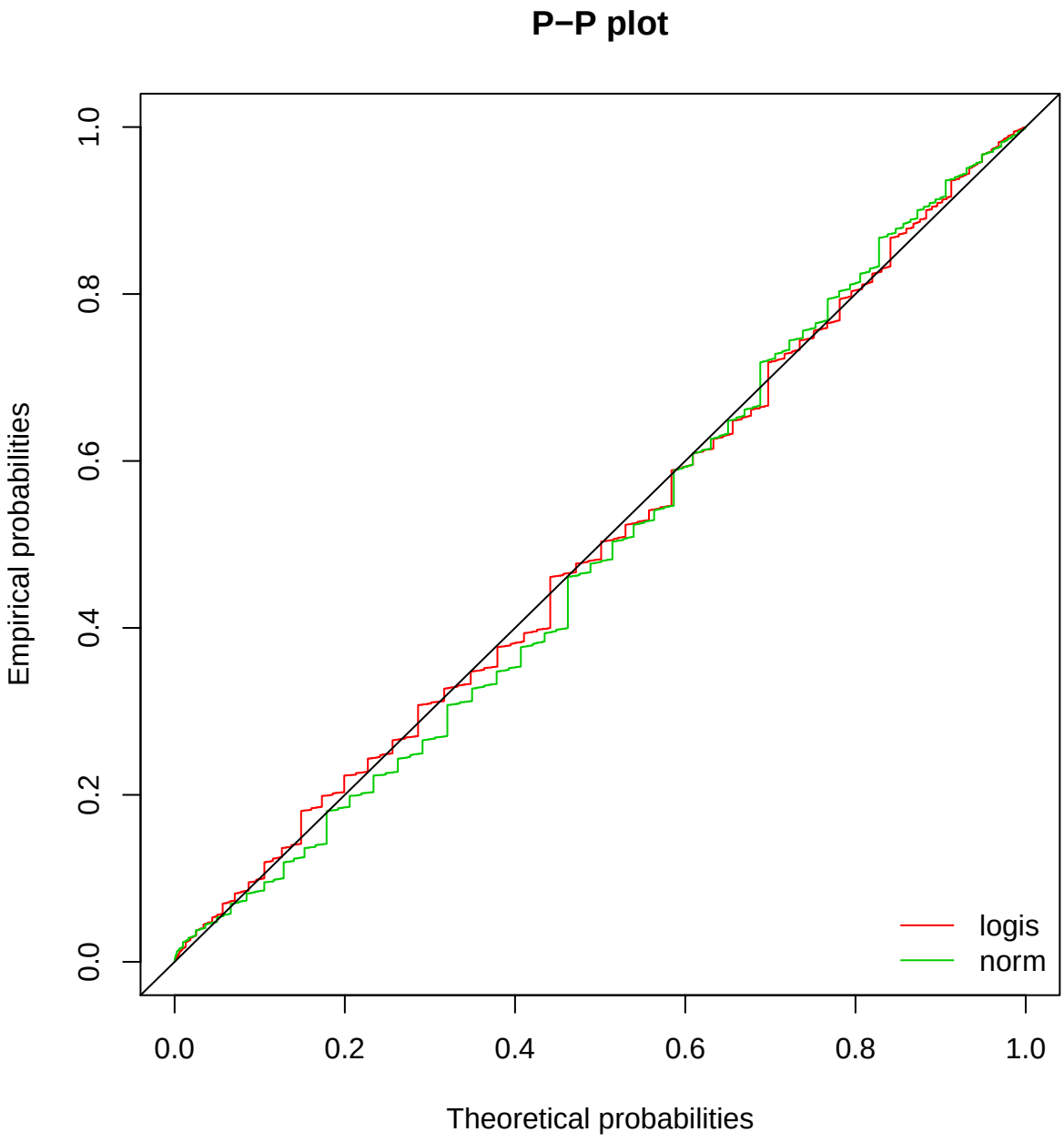
It is difficult to interpret the empirical vs. theoretical CDFs.

```
qqcomp(list(LOGINClogis, LOGINCnorm), legendtext = plot.legend)
```



The Q-Q plot shows that different distributions may be better fits for different sections of the dataset.

```
ppcomp(list(LOGINClogis, LOGINCnorm), legendtext = plot.legend)
```



The P-P plot shows that both distributions are relatively good fits.

Due to some ambiguity, we analyze goodness-of-fit statistics for each of the fitted distributions

```
gofstat(list(LOGINClogis, LOGINCnorm),fitnames=c("logis", "norm"))
```

```
## Goodness-of-fit statistics
##                                  logis        norm
## Kolmogorov-Smirnov statistic  0.04157616  0.06229993
## Cramer-von Mises statistic    5.05494916 15.23905630
## Anderson-Darling statistic   37.08574387 93.23049118
##
## Goodness-of-fit criteria
##                                 logis     norm
## Akaike's Information Criterion 37433.50 38655.50
## Bayesian Information Criterion 37449.93 38671.93
```

The goodness-of-fit statistics support the logistic distribution as being the best fitted distribution.

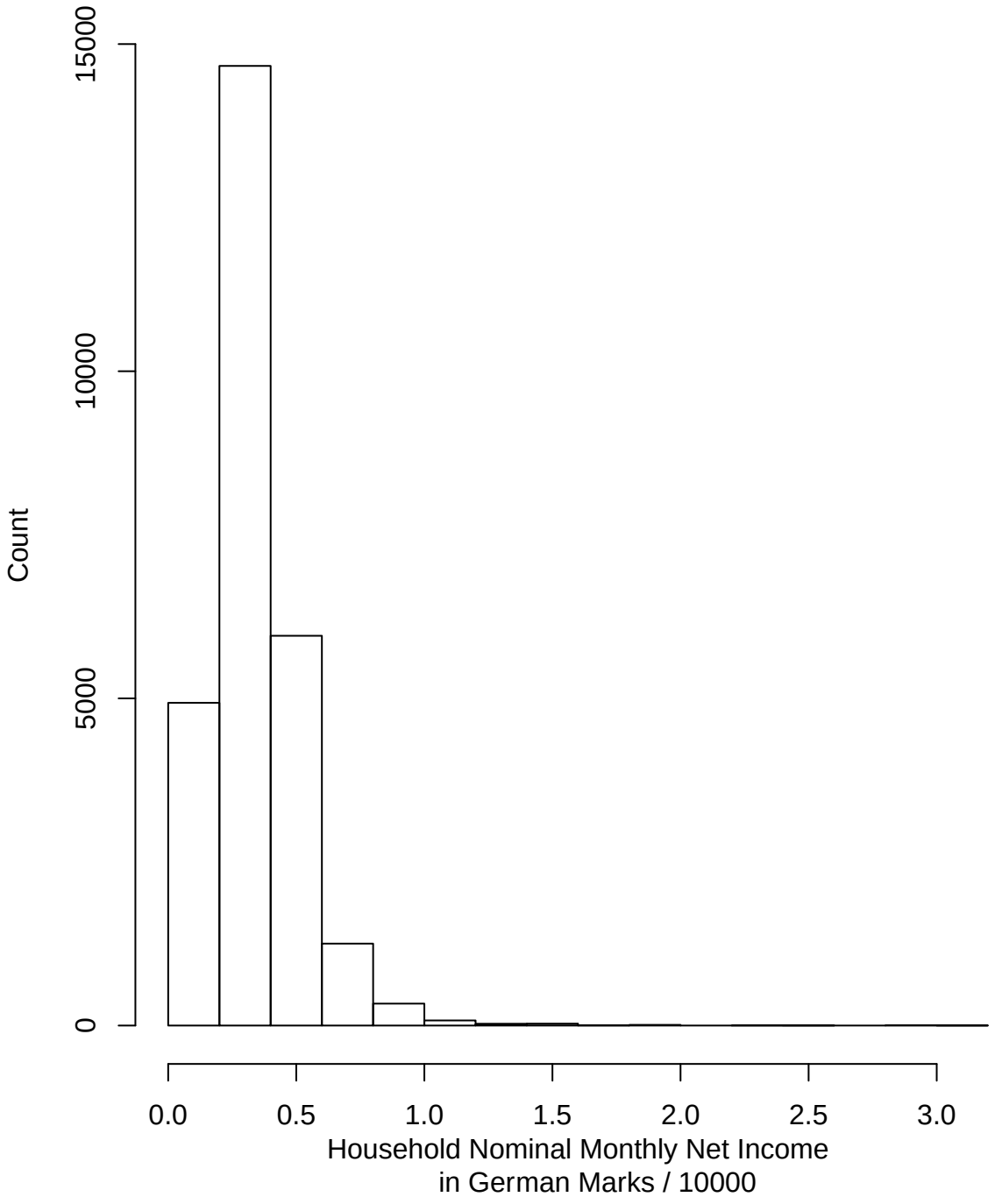
Conclusion about LOGINC

We conclude that the LOGINC variable is best approximated by a lognormal distribution.

Histogram and Density Curve for HHNINC

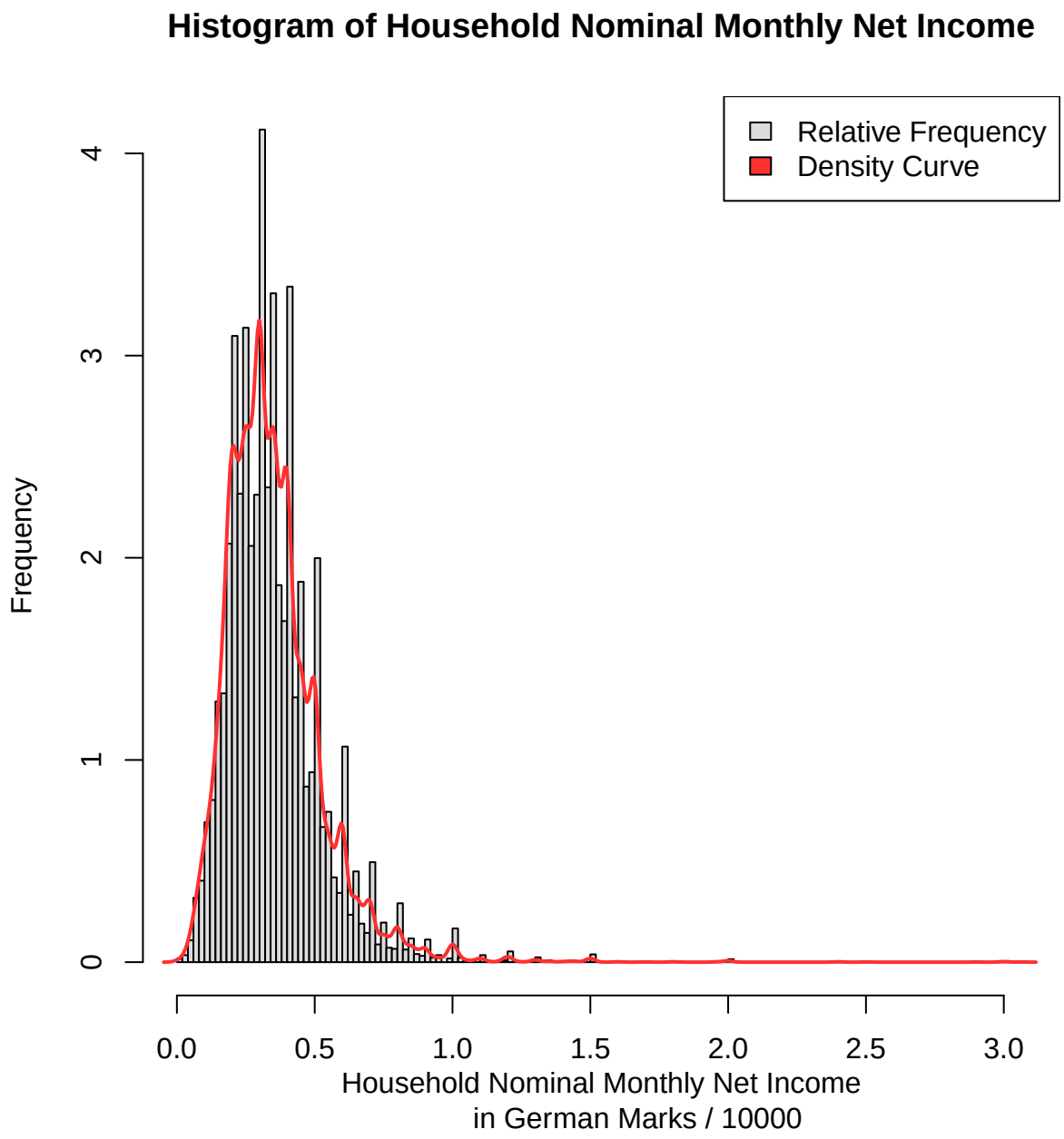
Histogram

```
hist(Pr3c$HHNINC, xlab= "Household Nominal Monthly Net Income
     in German Marks / 10000", ylab= "Count",
     main= "Histogram of Household Nominal Monthly Net Income")
```


Histogram of Household Nominal Monthly Net Income
Count
0
5000
10000
15000
0.0
0.5
1.0
1.5
2.0
2.5
3.0
Household Nominal Monthly Net Income
in German Marks / 10000

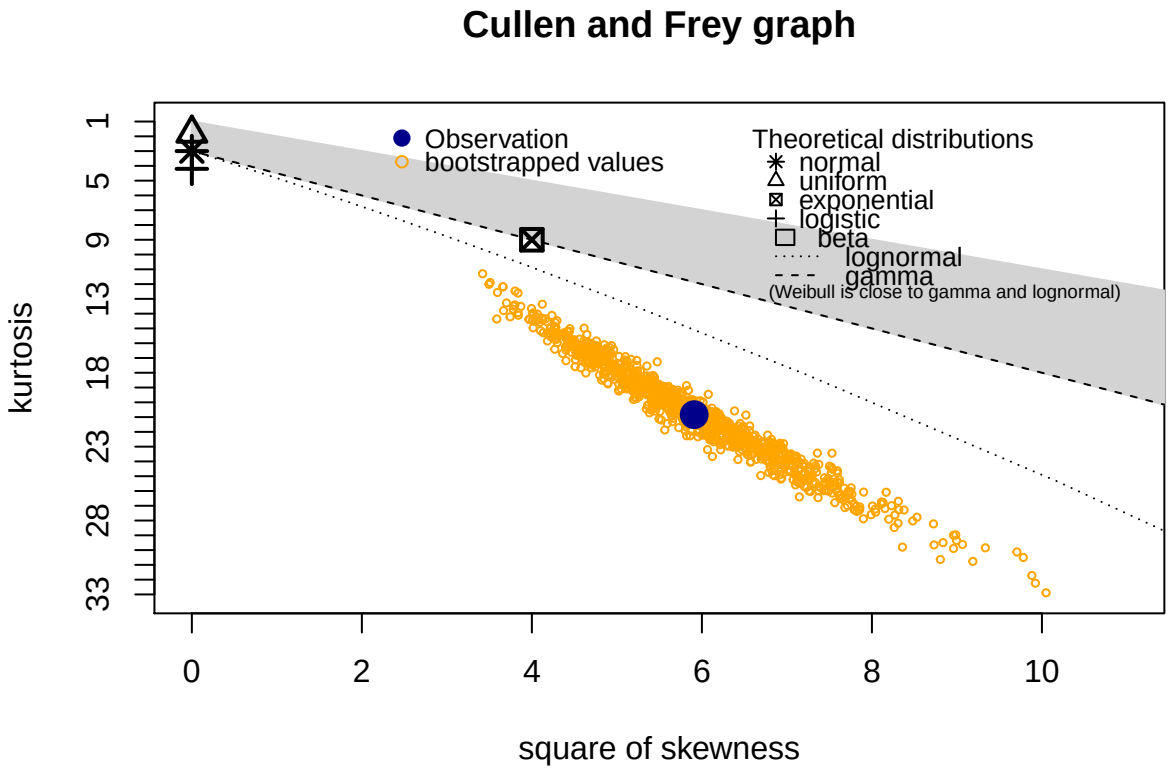
Histogram and Density Curve

```
truehist(Pr3c$HHNINC,col="gainsboro", ylab="Frequency",
         xlab= "Household Nominal Monthly Net Income
         in German Marks / 10000",
         main= "Histogram of Household Nominal Monthly Net Income")
lines(density((Pr3c$HHNINC)), lwd=2,col="firebrick1")
legend("topright", c("Relative Frequency", "Density Curve"),
       fill=c("gainsboro", "firebrick1"))
```



We plot the Cullen and Frey graph to give us an idea of what distributions to try out:

```
descdist(Pr3c$HHNINC, boot = 1000)
```




```
## summary statistics
## ------
## min:  0.0015   max:  3.0671
## median:  0.32
## mean:  0.3521926
## estimated sd:  0.1768513
## estimated skewness:  2.430908
## estimated kurtosis:  20.83494
```

Observe that gamma, lognormal, and weibull distributions are possibilities.

Also, observe that the exponential distribution has to have a square of skewness value equal to 4, so the best distribution is unlikely to be exponential, but we will attempt to test that fit as well.

We will attempt to fit various gamma distributions, a lognormal distribution, a weibull distribution, and an exponential distribution.

Testing fits for distributions

Testing fit for gamma distribution

```
HHNINCgamma <- fitdist(Pr3c$HHNINC, distr = "gamma", method = "mme")
```

Testing fit for gamma distribution with different parameters

```
HHNINCgamma1 <- fitdist(Pr3c$HHNINC, distr = "gamma", method = "mle",
                        lower = c(0, 0), start = list(scale = 1, shape = 1))
```

Testing fit for gamma distribution with different parameters

```
HHNINCgamma2 <- fitdist(Pr3c$HHNINC, distr = "gamma", method = "mle",
                        lower = c(0, 0), start = list(scale = 9, shape = 2))
```

Testing fit for gamma distribution with different parameters

```
HHNINCgamma3 <- fitdist(Pr3c$HHNINC, distr = "gamma", method = "mle",
                        lower = c(0, 0), start = list(scale = 0.5, shape = 1))
```

Testing fit for a lognormal distribution

```
HHNINClnorm <- fitdist(Pr3c$HHNINC, "lnorm")
```

Testing fit for a weibull distribution

```
HHNINCweibull <- fitdist(Pr3c$HHNINC, "weibull")
```

Testing for an exponential distribution

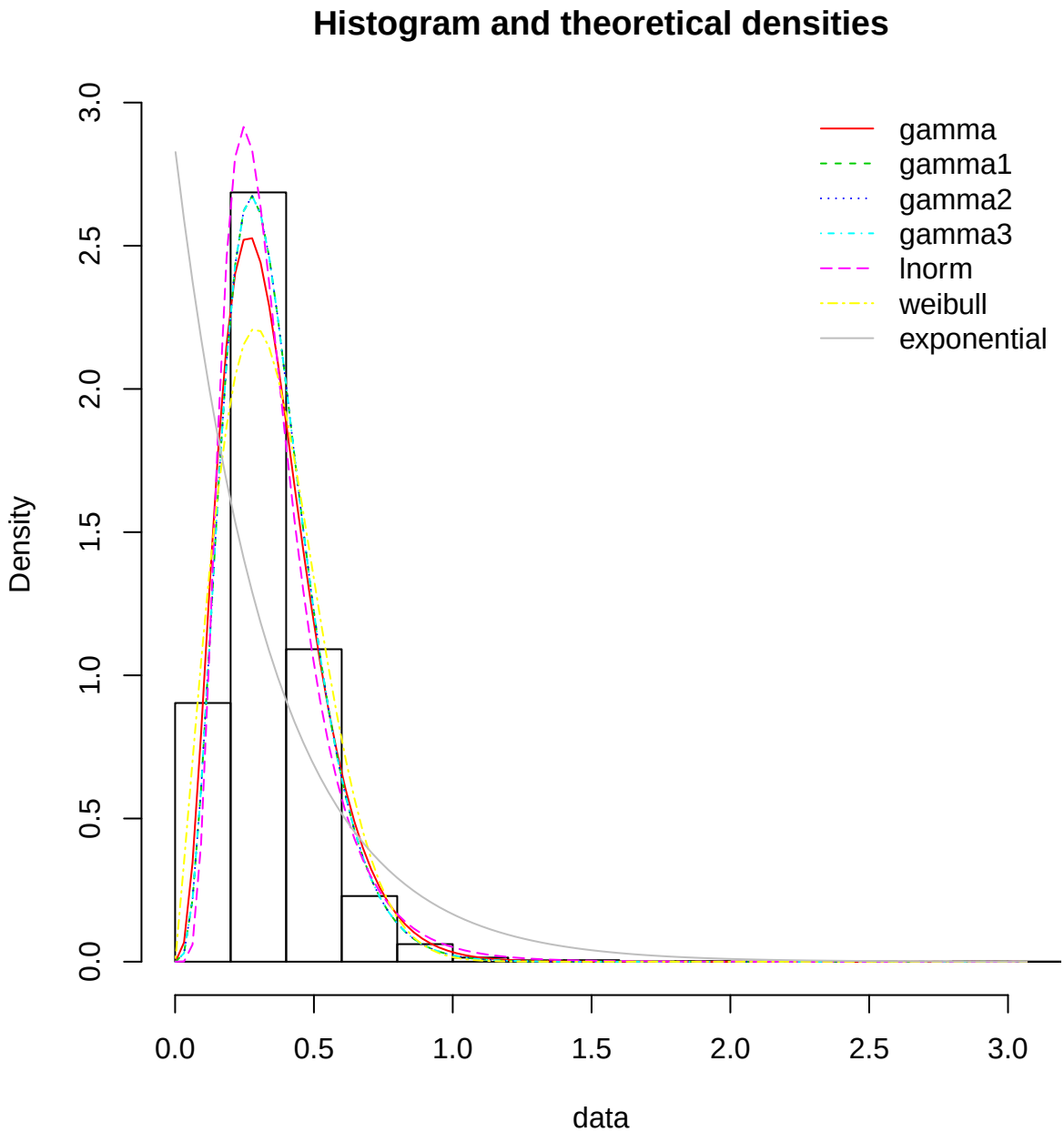
```
HHNINCexp <- fitdist(Pr3c$HHNINC, "exp")
```

Setting Legend

```
plot.legend <- c("gamma", "gamma1", "gamma2", "gamma3", "lnorm", "weibull", "exponential")
```

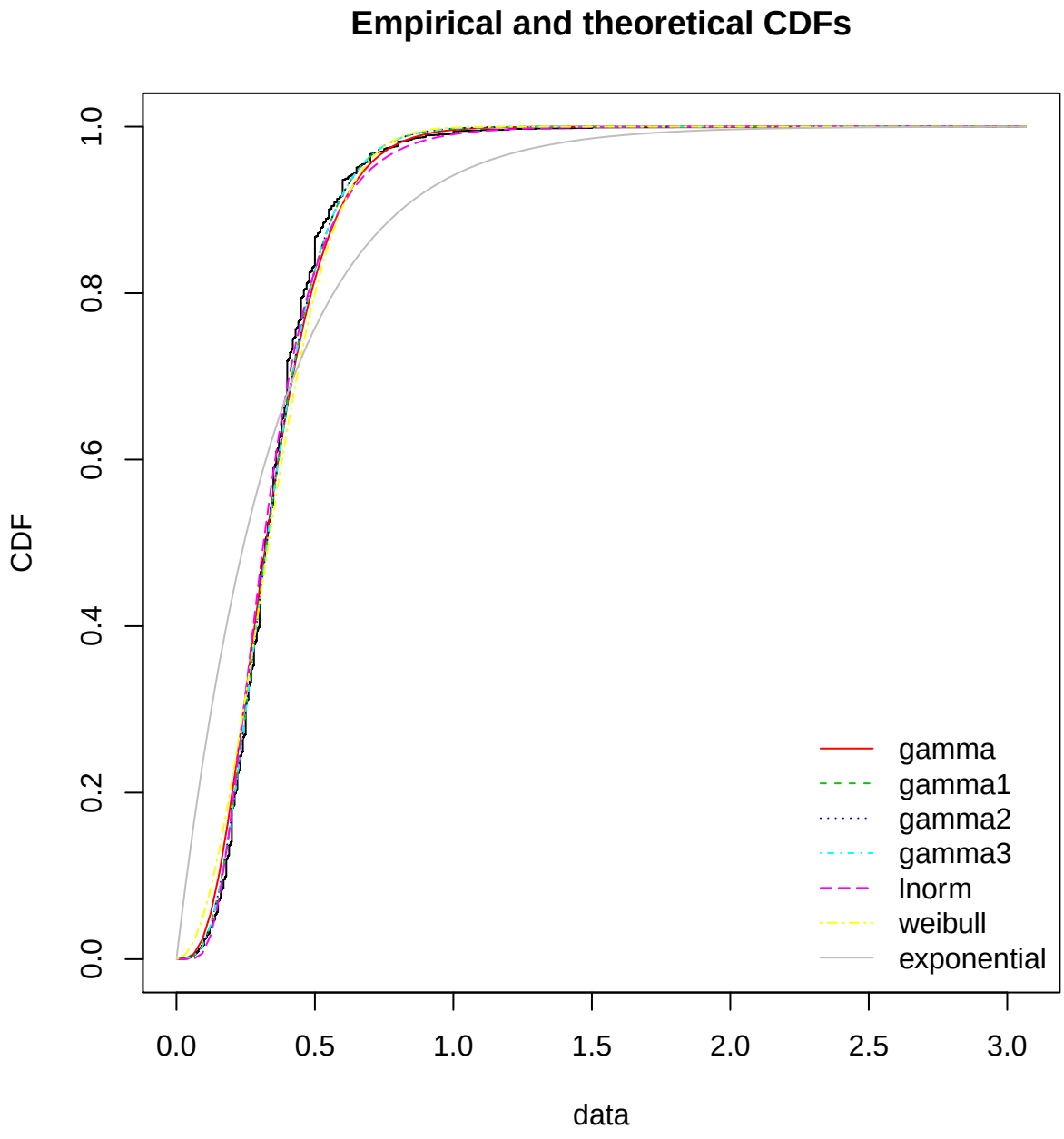
We compare the histogram with the theoretical densities on the following page.

```
denscomp(list(HHNINCgamma, HHNINCgamma1, HHNINCgamma2, HHNINCgamma3,
              HHNINClnorm, HHNINCweibull, HHNINCexp), legendtext = plot.legend)
```



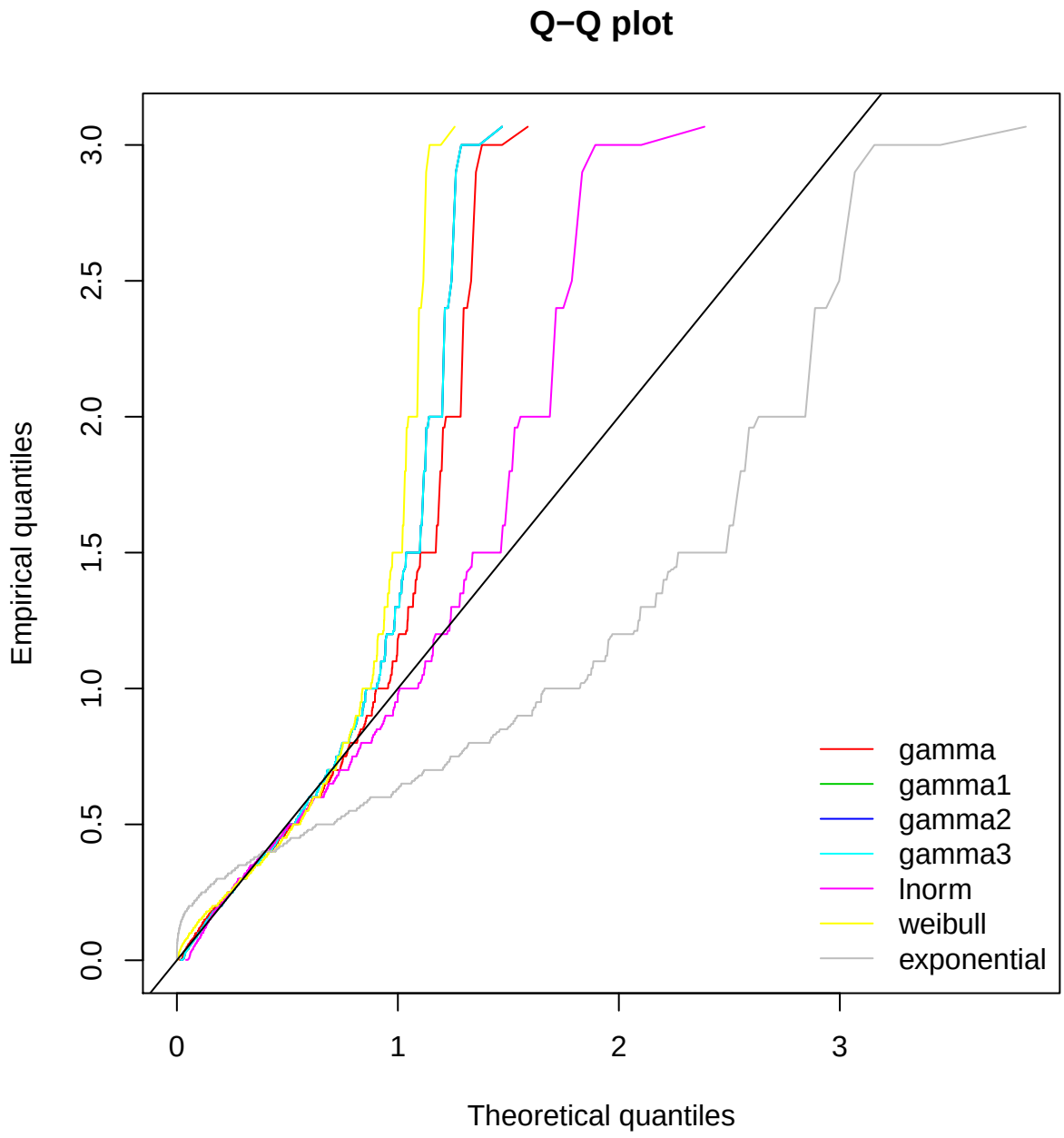
Observe that the lognormal and gamma distributions appear to be the best fits based on the theoretical densities of the distributions.

```
cdfcomp(list(HHNINCgamma, HHNINCgamma1, HHNINCgamma2, HHNINCgamma3,
             HHNINClnorm, HHNINCweibull, HHNINCexp), legendtext = plot.legend)
```



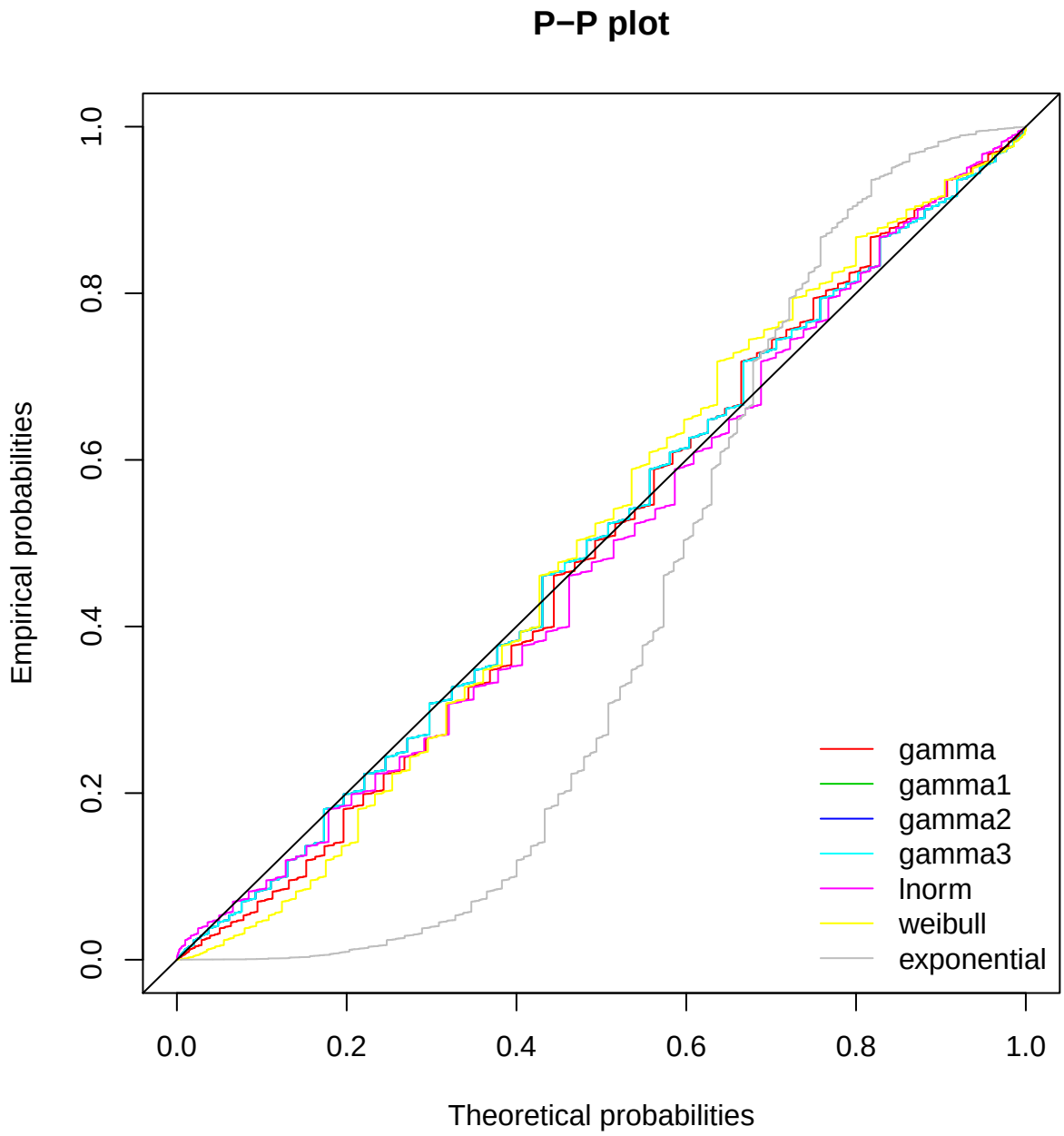
Most of the distributions appear to be relatively good fits, except for the exponential distribution.

```
qqcomp(list(HHNINCgamma, HHNINCgamma1, HHNINCgamma2, HHNINCgamma3,
            HHNINClnorm, HHNINCweibull, HHNINCexp), legendtext = plot.legend)
```



The Q-Q plot shows that the exponential distribution is a particularly bad fit. Also, it shows the lognormal distribution is the best fit of the distributions attempted.

```
ppcomp(list(HHNINCgamma, HHNINCgamma1, HHNINCgamma2, HHNINCgamma3,
            HHNINClnorm, HHNINCweibull, HHNINCexp), legendtext = plot.legend)
```



Most of the distributions appear to be relatively good fits, except for the exponential distribution.

Due to some ambiguity, we analyze goodness-of-fit statistics for each of the fitted distributions

```
gofstat(list(HHNINCgamma, HHNINCgamma1, HHNINCgamma2, HHNINCgamma3,
             HHNINClnorm, HHNINCweibull, HHNINCexp),
             fitnames=c("gamma", "gamma1", "gamma2", "gamma3",
                        "lnorm", "weibull", "exponential"))
```

```
## Goodness-of-fit statistics
##                                    gamma      gamma1      gamma2
## Kolmogorov-Smirnov statistic   0.05500451  0.05049242  0.05049819
## Cramer-von Mises statistic    20.35538793  8.34834434  8.35110330
## Anderson-Darling statistic   135.64326756 52.87808061 52.89351919
##                                   gamma3       lnorm     weibull
## Kolmogorov-Smirnov statistic  0.05050486  0.06229992  0.08137379
## Cramer-von Mises statistic    8.34974981 15.23905799 48.33527488
## Anderson-Darling statistic   52.87856884 93.23049685         Inf
##                                exponential
## Kolmogorov-Smirnov statistic    0.3002539
## Cramer-von Mises statistic    764.0237060
## Anderson-Darling statistic   3897.2529926
##
## Goodness-of-fit criteria
##                                     gamma    gamma1    gamma2    gamma3
## Akaike's Information Criterion -24869.35 -25143.60 -25143.60 -25143.60
## Bayesian Information Criterion -24852.92 -25127.17 -25127.17 -25127.17
##                                    lnorm   weibull exponential
## Akaike's Information Criterion -24519.56 -21819.82   -2377.048
## Bayesian Information Criterion -24503.13 -21803.39   -2368.833
```

The goodness-of-fit statistics suggest that the gamma distribution with scale = 1 and shape = 1 is the best fitted distribution.

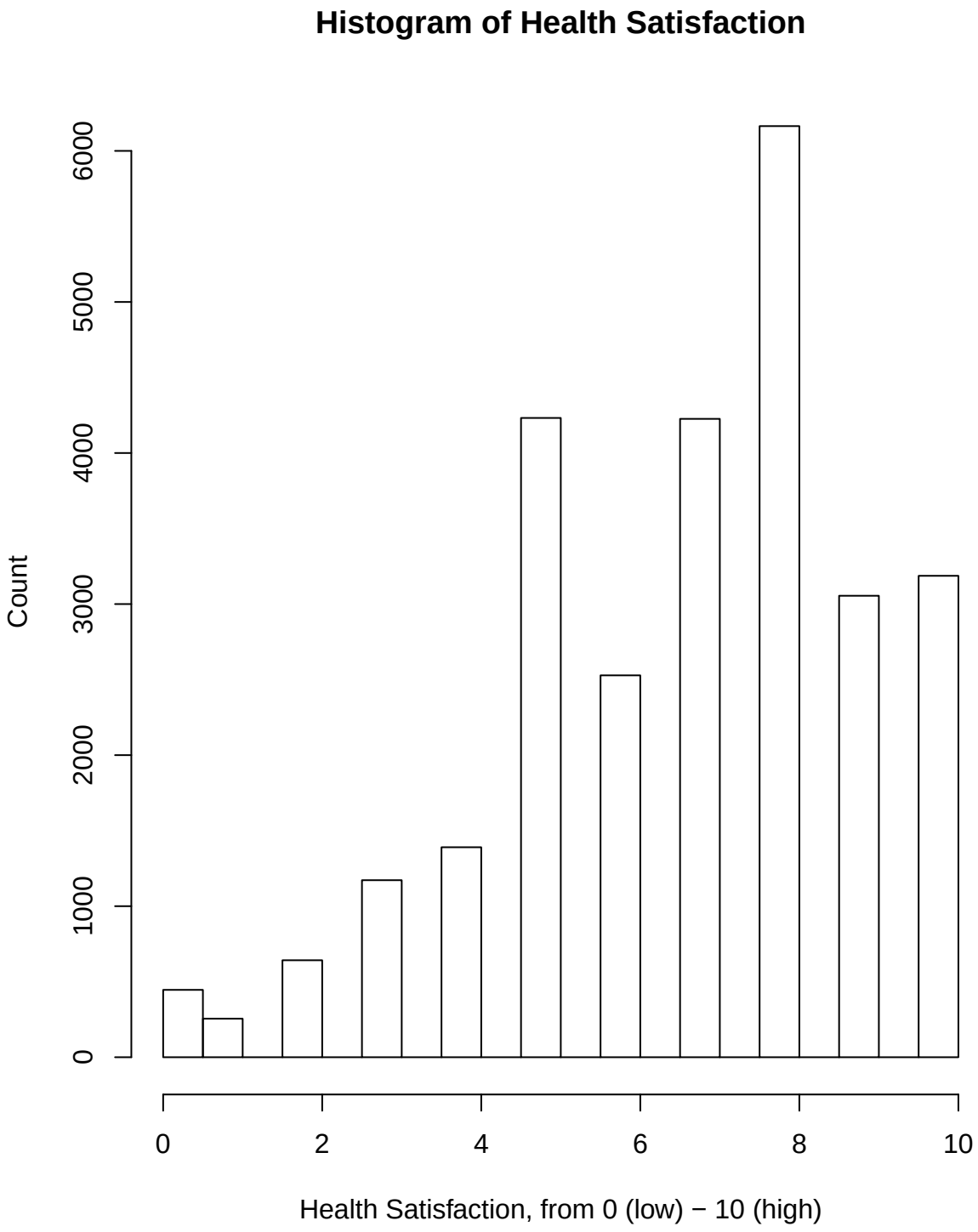
Conclusion about HHNINC

We conclude that the HHNINC variable is best approximated by a gamma distribution, particularly one with scale = 1 and shape = 1.

Histogram and Density Curve for NEWHSAT

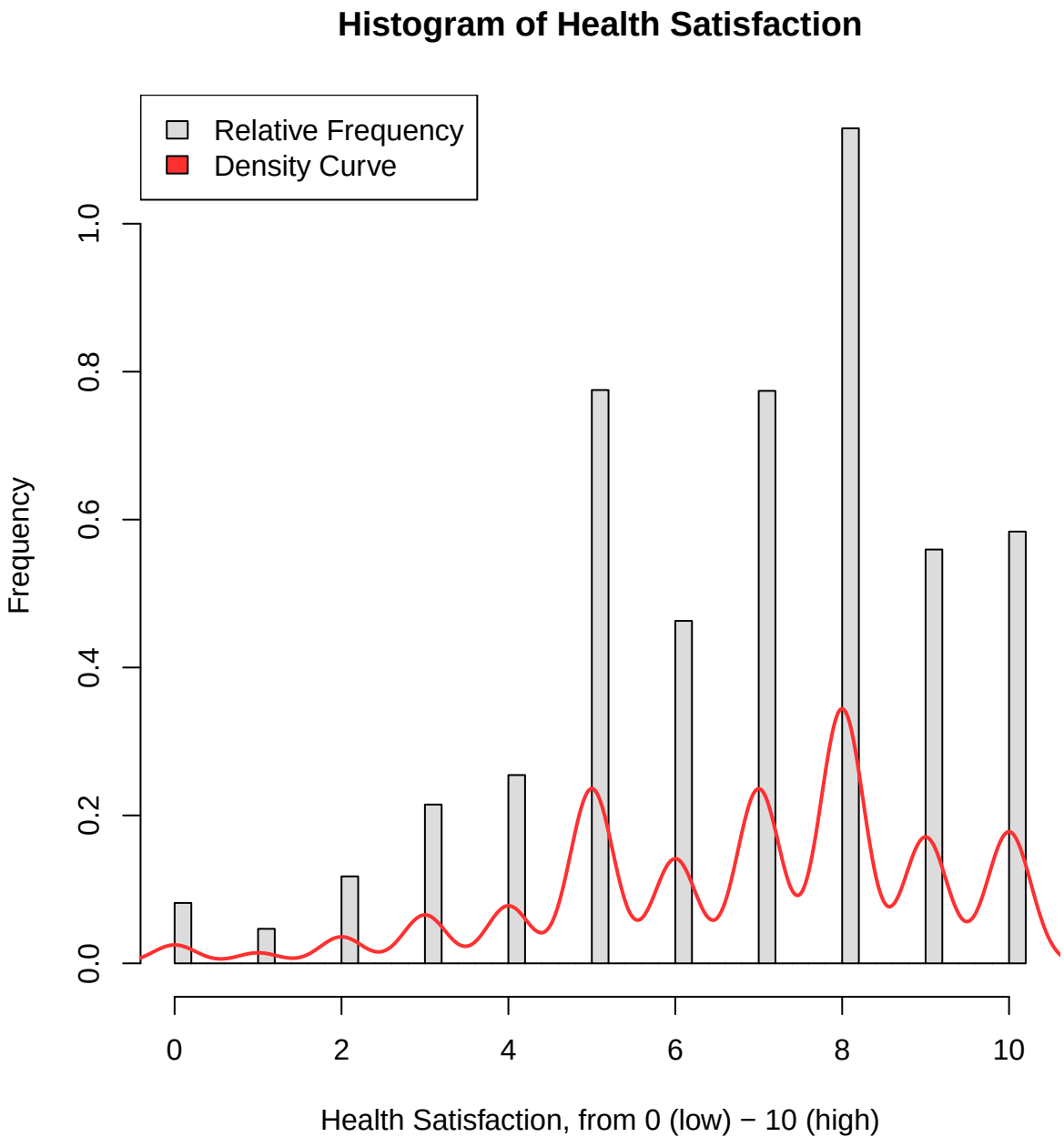
Histogram

```
hist(Pr3c$NEWHSAT, xlab= "Health Satisfaction, from 0 (low) - 10 (high)",
     ylab= "Count", main= "Histogram of Health Satisfaction")
```



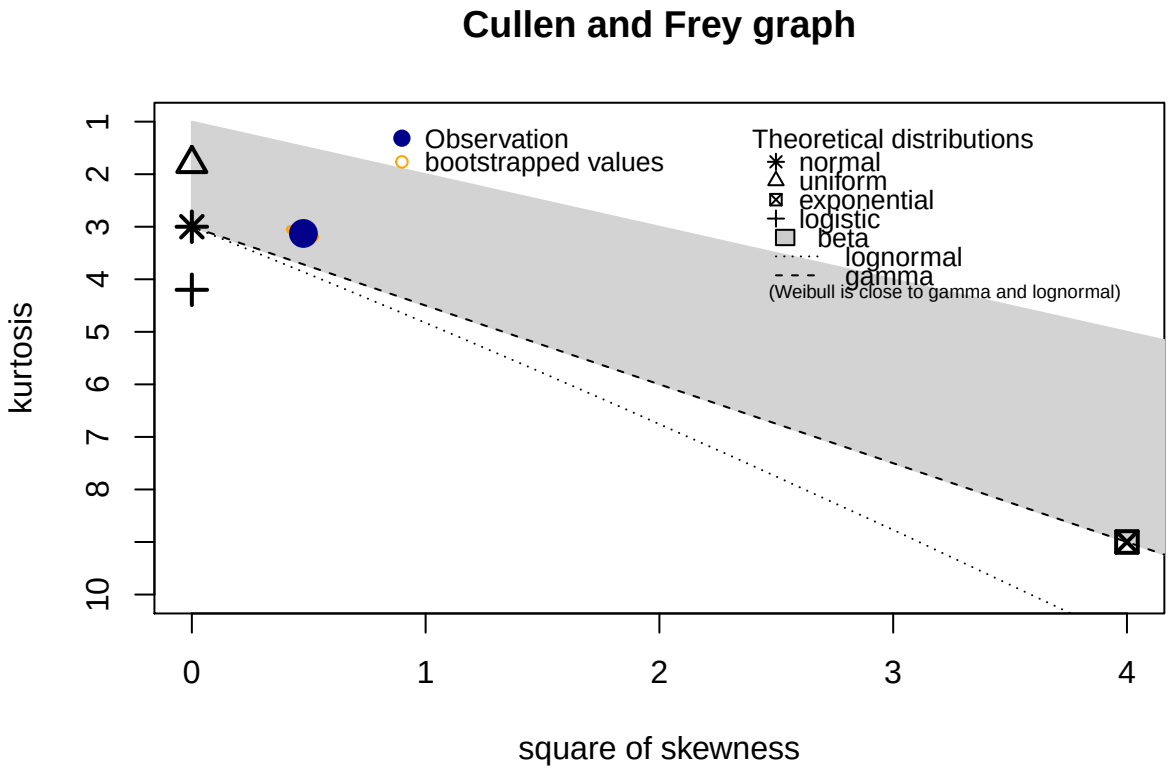
Histogram and Density Curve

```
truehist(Pr3c$NEWHSAT,col="gainsboro", ylab="Frequency",
         xlab= "Health Satisfaction, from 0 (low) - 10 (high)",
         main= "Histogram of Health Satisfaction")
lines(density((Pr3c$NEWHSAT)), lwd=2,col="firebrick1")
legend("topleft", c("Relative Frequency", "Density Curve"),
       fill=c("gainsboro", "firebrick1"))
```



We plot the Cullen and Frey graph to give us an idea of what distributions to try out:

```
descdist(Pr3c$NEWHSAT, boot = 1000)
```




```
## summary statistics
## ------
## min:  0   max:  10
## median:  7
## mean:  6.784702
## estimated sd:  2.293682
## estimated skewness:  -0.6908964
## estimated kurtosis:  3.127661
```

Observe that gamma distributions are possibilities.

Note that normal distribution is also a possibility as the sample itself and most of the bootstrapped samples have a kurtosis very close to that of a normal distribution.

Note that a beta distribution has to have a value between [0,1] whereas our data has values between [0,10].

Also, the lognormal distribution does not make sense as we do not have all positive values in our dataset.

We will attempt to fit various gamma distributions and a normal distribution.

Testing fits for distributions

Testing fit for gamma distribution

```
NEWHSATgamma <- fitdist(Pr3c$NEWHSAT, distr = "gamma", method = "mme")
```

Testing fit for gamma distribution with different parameters

```
NEWHSATgamma1 <- fitdist(Pr3c$NEWHSAT, distr = "gamma", method = "mle",
                         lower = c(0, 0), start = list(scale = 1, shape = 1))
```

Testing fit for gamma distribution with different parameters

```
NEWHSATgamma2 <- fitdist(Pr3c$NEWHSAT, distr = "gamma", method = "mle",
                         lower = c(0, 0), start = list(scale = 0.5, shape = 1))
```

Testing for a normal distribution

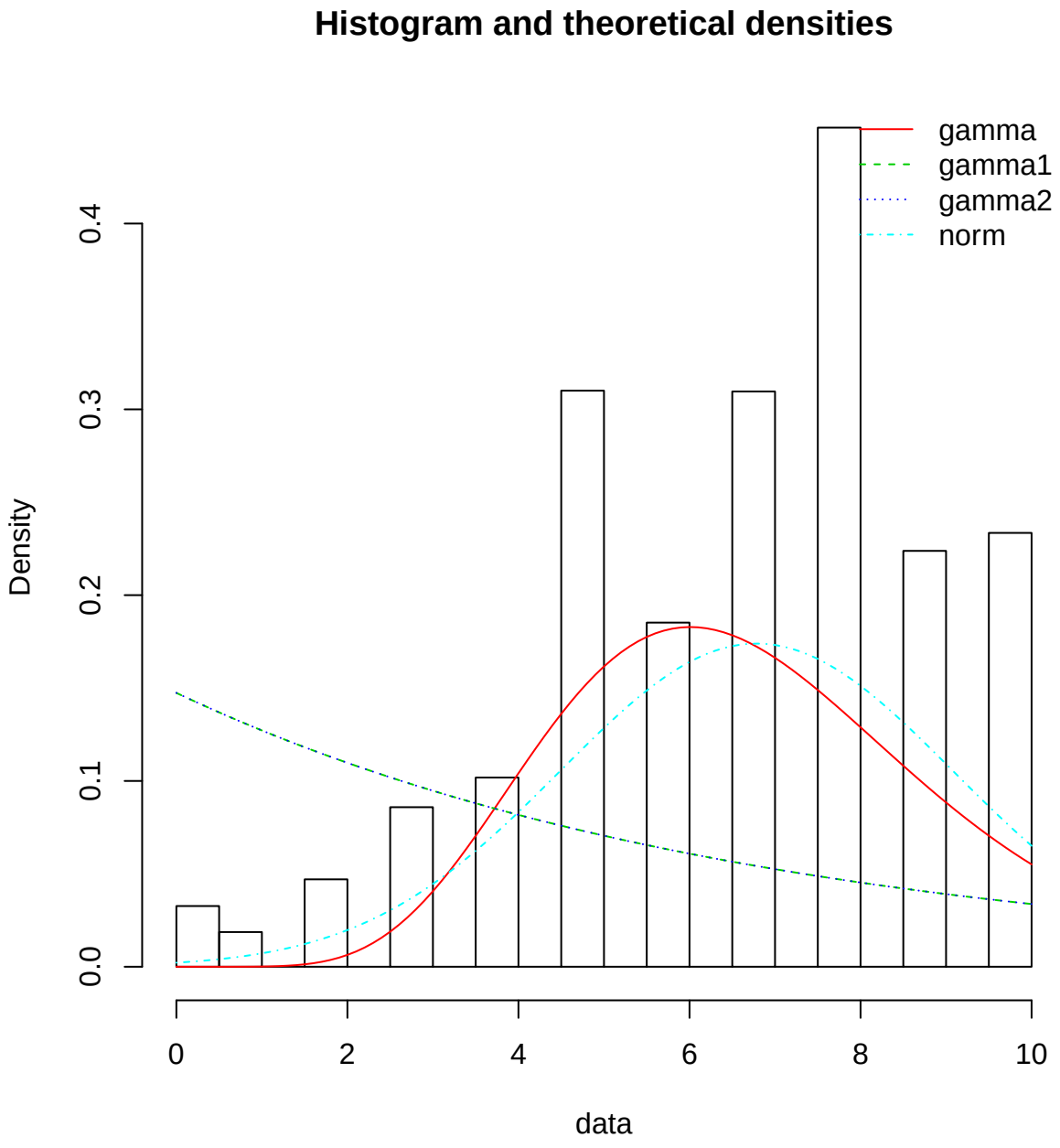
```
NEWHSATnorm <- fitdist(Pr3c$NEWHSAT, "norm")
```

Setting Legend

```
plot.legend <- c("gamma", "gamma1", "gamma2", "norm")
```

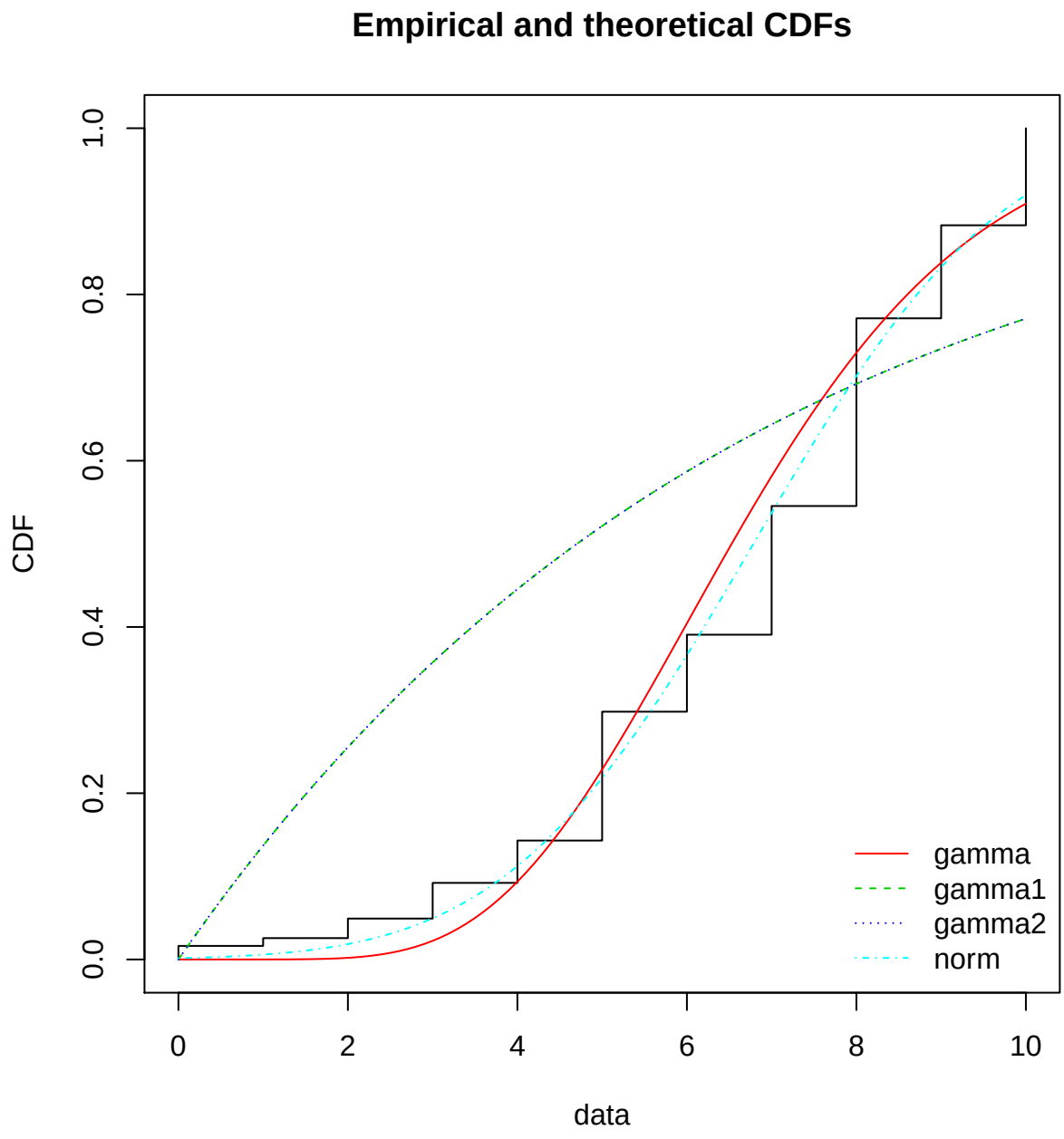
We compare the histogram with the theoretical densities on the following page.

```
denscomp(list(NEWHSATgamma, NEWHSATgamma1, NEWHSATgamma2, NEWHSATnorm),
         legendtext = plot.legend)
```



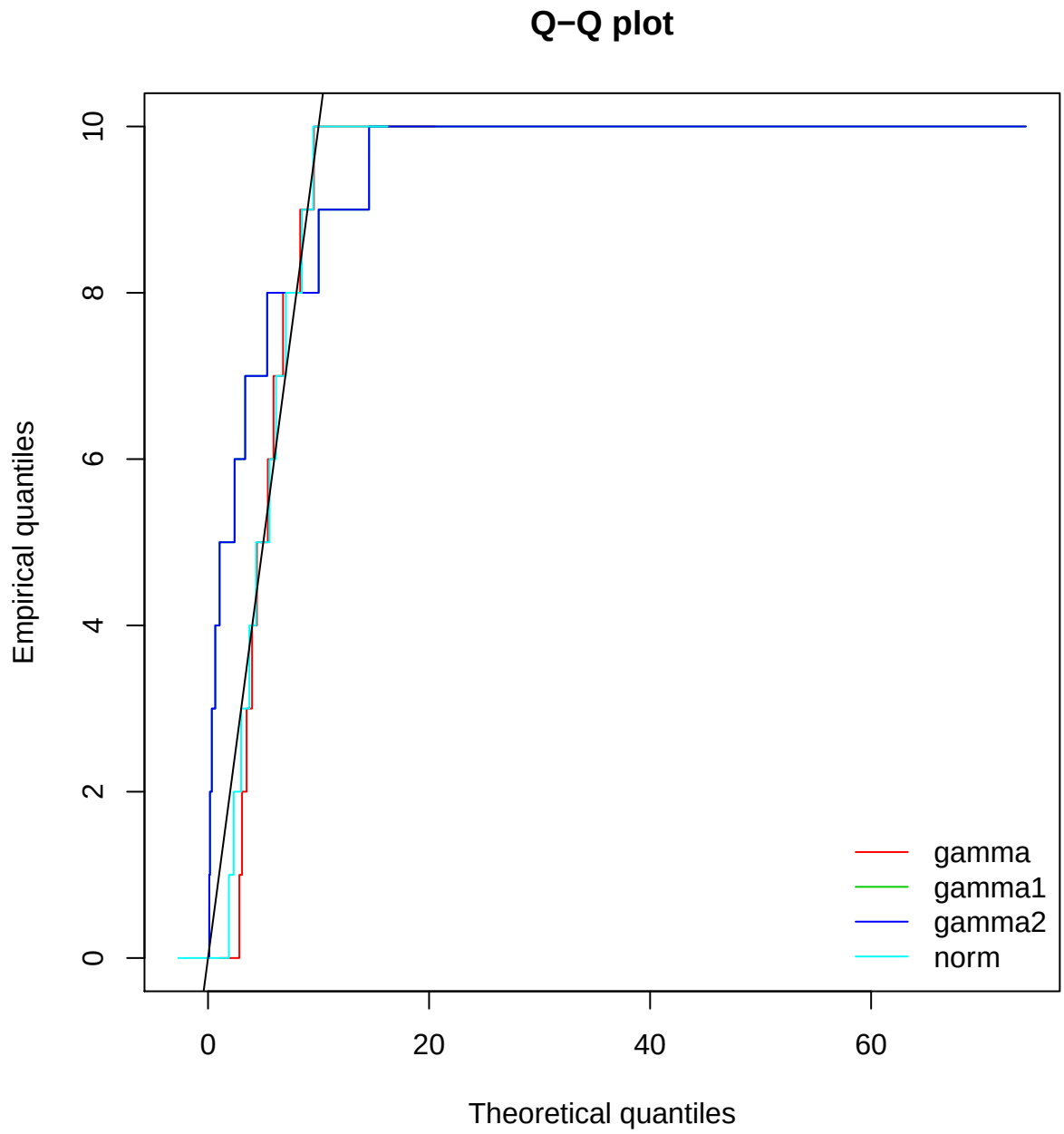
Observe that the normal distribution and gamma distribution fitted by matching moments appear to be the best fits based on the theoretical densities of the distributions.

```
cdfcomp(list(NEWHSATgamma, NEWHSATgamma1, NEWHSATgamma2, NEWHSATnorm),
        legendtext = plot.legend)
```



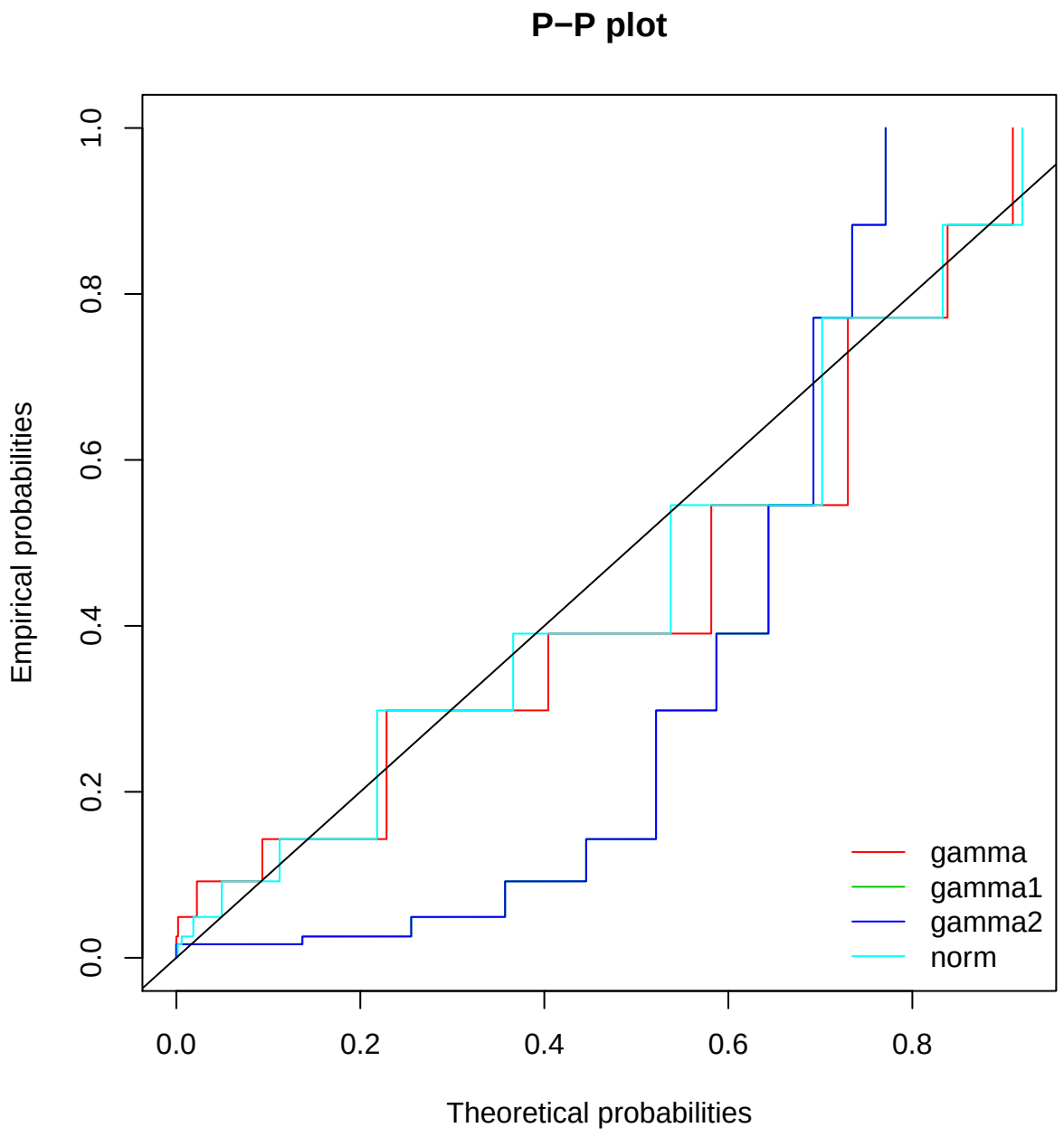
The same distributions as before appear to be the best fits.

```
qqcomp(list(NEWHSATgamma, NEWHSATgamma1, NEWHSATgamma2, NEWHSATnorm),
       legendtext = plot.legend)
```



The Q-Q plot basically agrees with the empirical vs. theoretical CDF plot.

```
ppcomp(list(NEWHSATgamma, NEWHSATgamma1, NEWHSATgamma2, NEWHSATnorm),
       legendtext = plot.legend)
```



The P-P plot shows again that the normal distribution and gamma distribution fitted by matching moments appear to be the best fits.

Due to some ambiguity, we analyze goodness-of-fit statistics for each of the fitted distributions

```
gofstat(list(NEWHSATgamma, NEWHSATgamma1, NEWHSATgamma2,
             NEWHSATnorm),fitnames=c("gamma", "gamma1", "gamma2", "norm"))
```

```
## Goodness-of-fit statistics
##                                  gamma       gamma1       gamma2
## Kolmogorov-Smirnov statistic   0.1907387    0.3783736    0.3783736
## Cramer-von Mises statistic   155.5831480 1120.1279074 1120.1279074
## Anderson-Darling statistic           Inf          Inf          Inf
##                                   norm
## Kolmogorov-Smirnov statistic   0.1563762
## Cramer-von Mises statistic    87.9263118
## Anderson-Darling statistic   512.5909473
##
## Goodness-of-fit criteria
##                                gamma   gamma1   gamma2     norm
## Akaike's Information Criterion   Inf 159127.5 159127.5 122790.2
## Bayesian Information Criterion   Inf 159143.9 159143.9 122806.6
```

The goodness-of-fit statistics suggest that the normal distribution is the best fitted distribution.

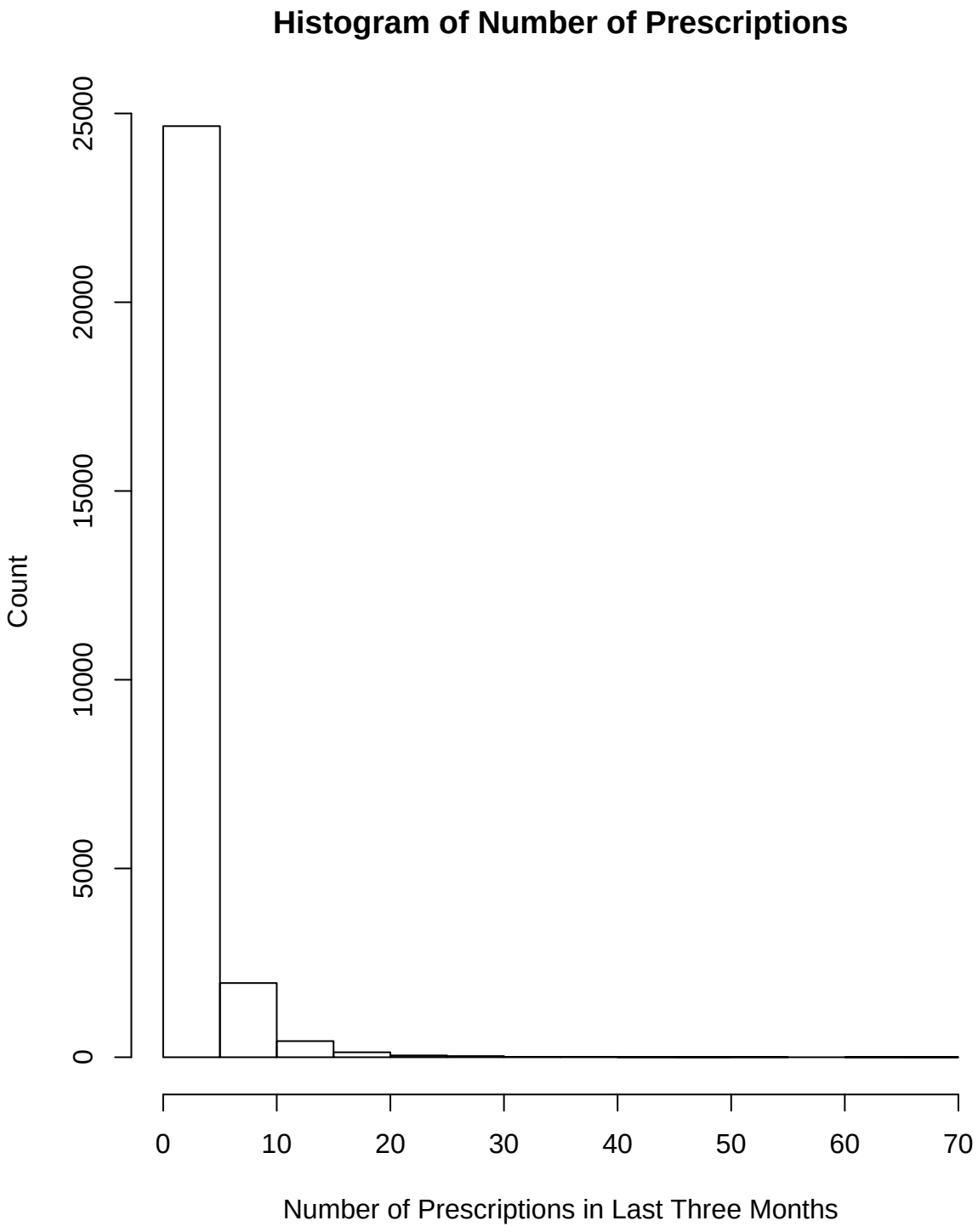
Conclusion about NEWHSAT

We conclude that the NEWHSAT variable is best approximated by a normal distribution.

Histogram and Density Curve for PRESCRIP

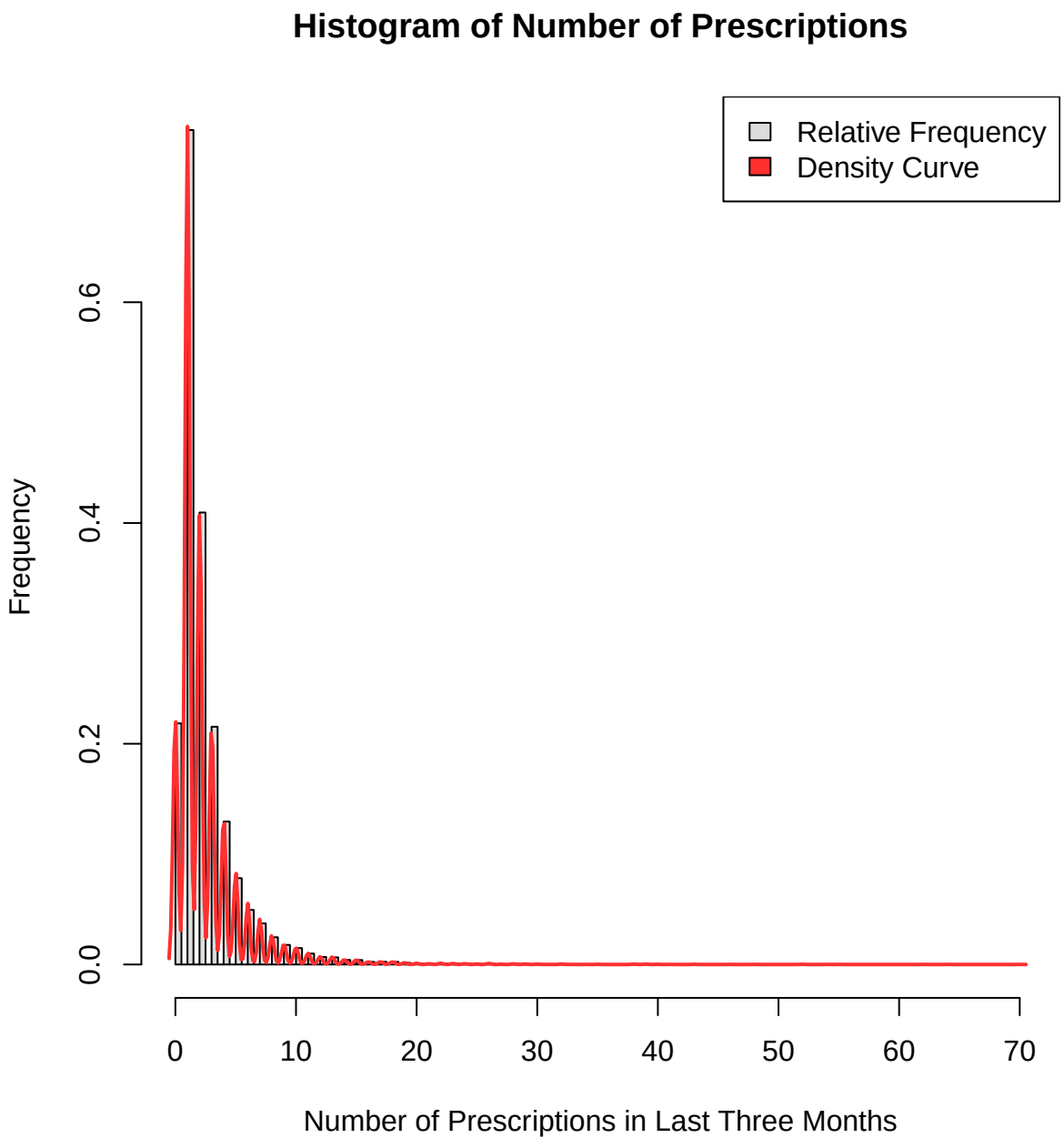
Histogram

```
hist(Pr3c$PRESCRIP, xlab= "Number of Prescriptions in Last Three Months",
     ylab= "Count", main= "Histogram of Number of Prescriptions")
```



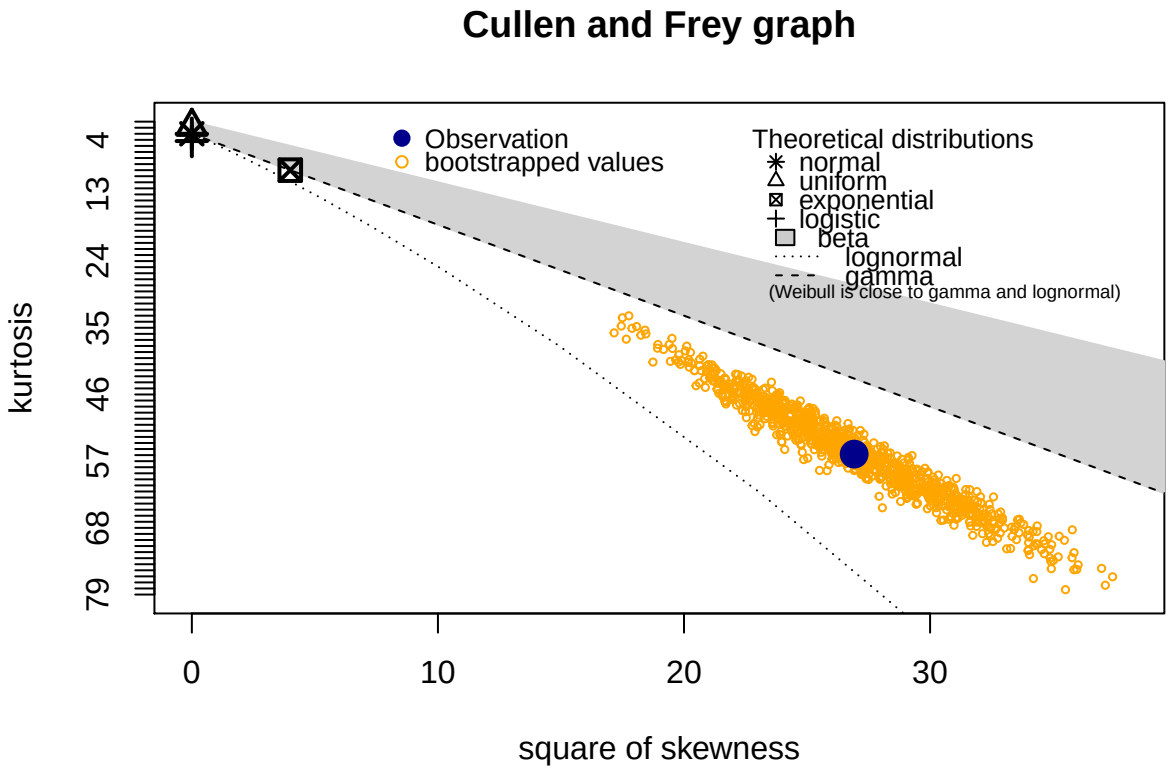
Histogram and Density Curve

```
truehist(Pr3c$PRESCRIP,col="gainsboro", ylab="Frequency",
         xlab= "Number of Prescriptions in Last Three Months",
         main= "Histogram of Number of Prescriptions")
lines(density((Pr3c$PRESCRIP)), lwd=2,col="firebrick1")
legend("topright", c("Relative Frequency", "Density Curve"),
       fill=c("gainsboro", "firebrick1"))
```



We plot the Cullen and Frey graph to give us an idea of what distributions to try out:

```
descdist(Pr3c$PRESCRIP, boot = 1000)
```




```
## summary statistics
## ------
## min:  0   max:  70
## median:  2
## mean:  2.496318
## estimated sd:  3.149182
## estimated skewness:  5.187982
## estimated kurtosis:  55.8363
```

Observe that gamma distributions are possibilities.

Note that because of the range of values in the dataset, a lognormal distribution does not make sense.

We will attempt to fit various gamma distributions.

Testing fits for distributions

Testing fit for a gamma distribution

```
PRESCRIPgamma <- fitdist(Pr3c$PRESCRIP, "gamma", method = "mme")
```

Testing fit for gamma distribution with different parameters

```
PRESCRIPgamma1 <- fitdist(Pr3c$PRESCRIP, distr = "gamma", method = "mle",
                          lower = c(0, 0), start = list(scale = 1, shape = 1))
```

Testing fit for gamma distribution with different parameters

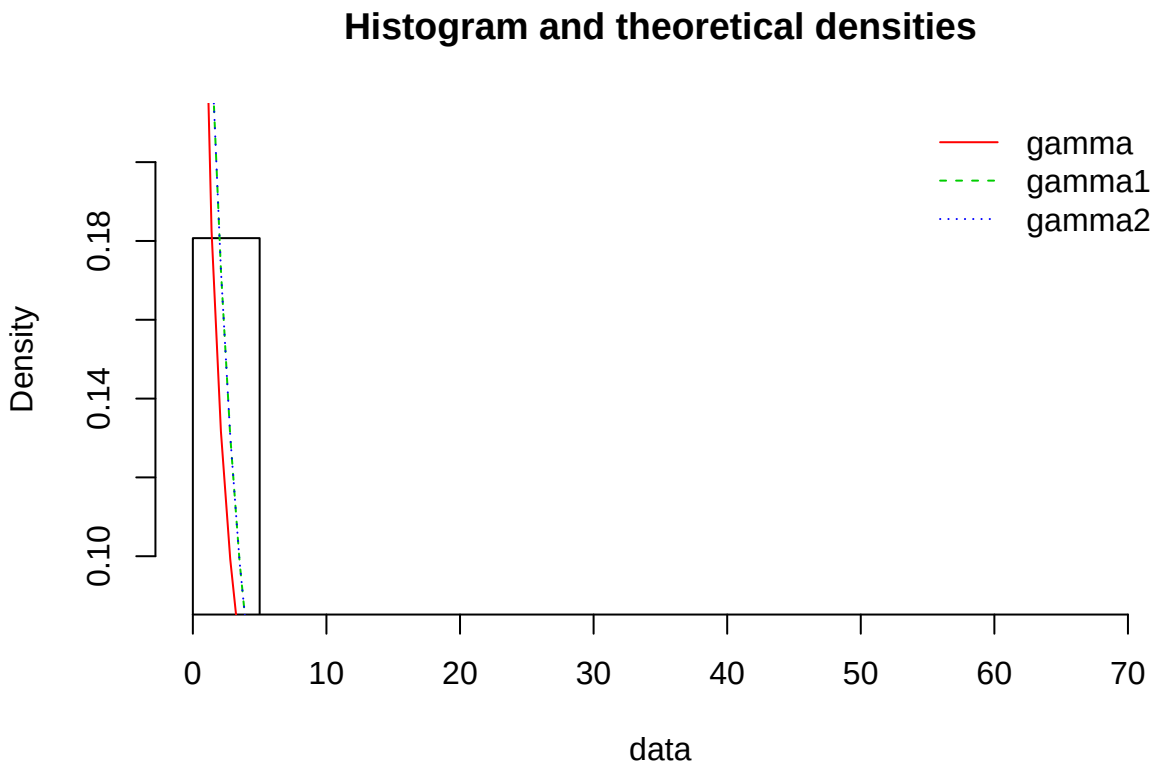
```
PRESCRIPgamma2 <- fitdist(Pr3c$PRESCRIP, distr = "gamma", method = "mle",
                         lower = c(0, 0), start = list(scale = 0.5, shape = 1))
```

Setting Legend

```
plot.legend <- c("gamma", "gamma1", "gamma2")
```

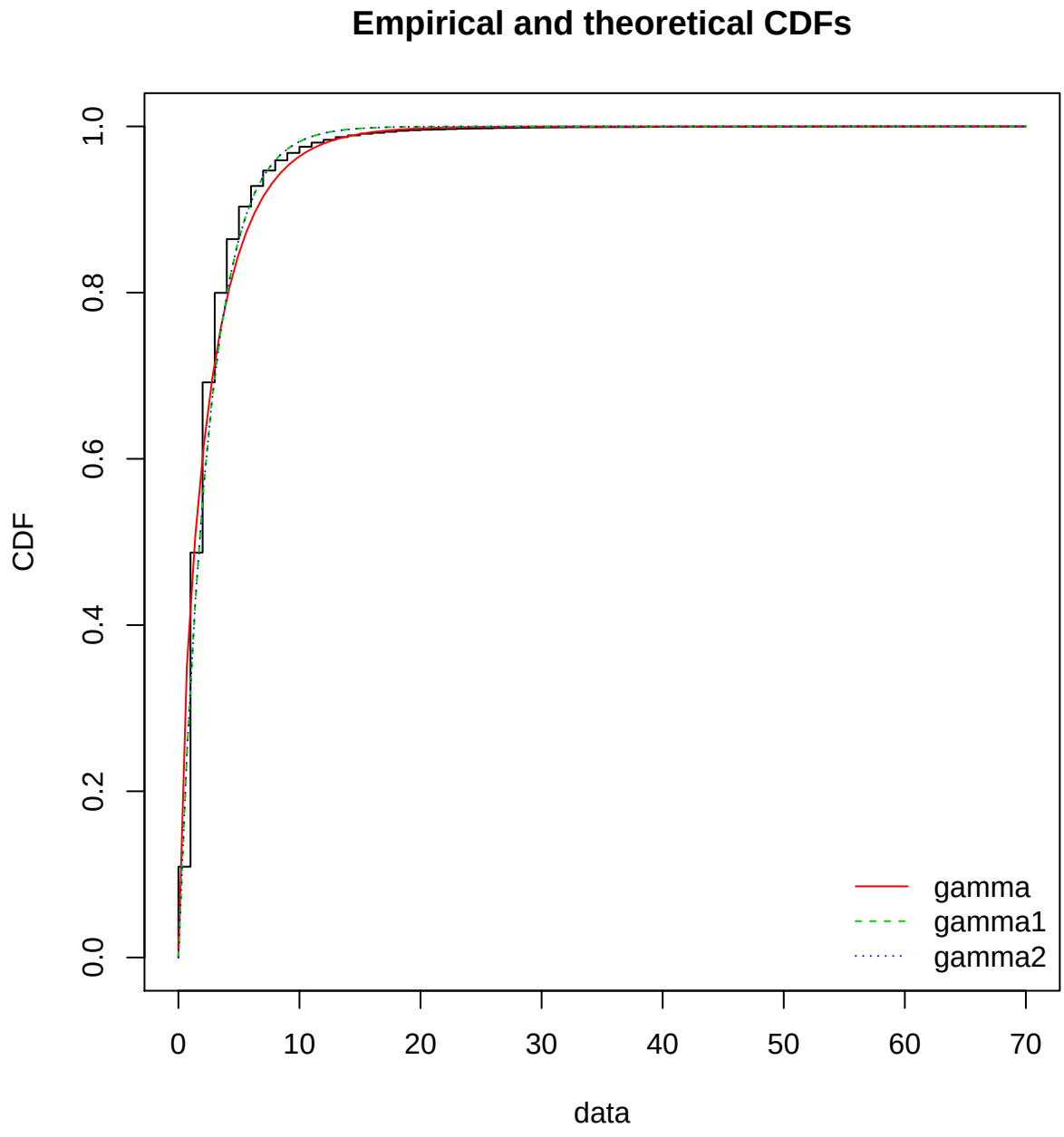
Comparing Histogram and Theoretical Densities

```
denscomp(list(PRESCRIPgamma, PRESCRIPgamma1, PRESCRIPgamma2),
         legendtext = plot.legend, ylim = .15)
```



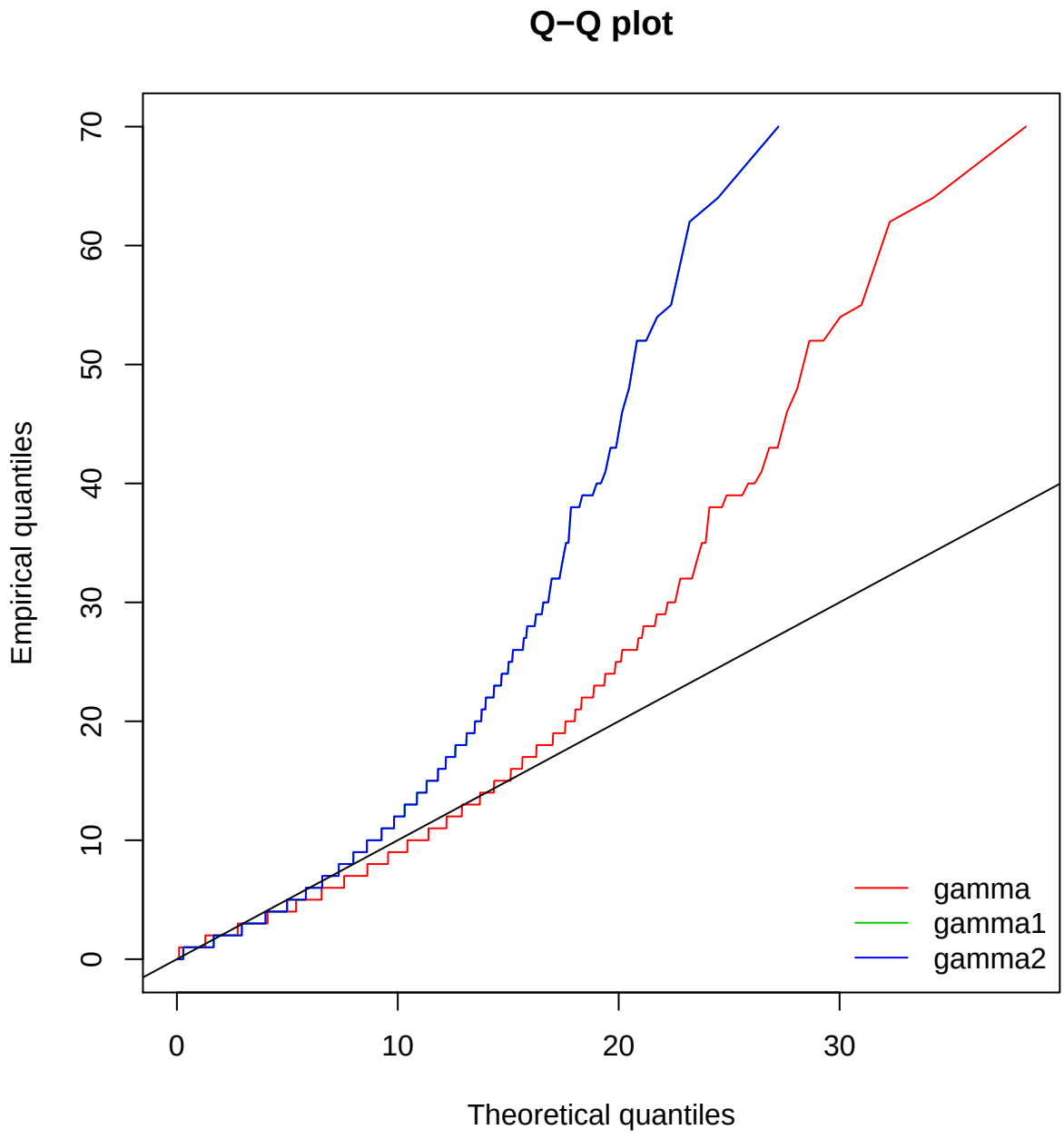
Observe that the plot does not tell us much except that all three distribution are relatively in the correct range.

```
cdfcomp(list(PRESCRIPgamma, PRESCRIPgamma1, PRESCRIPgamma2),
        legendtext = plot.legend)
```



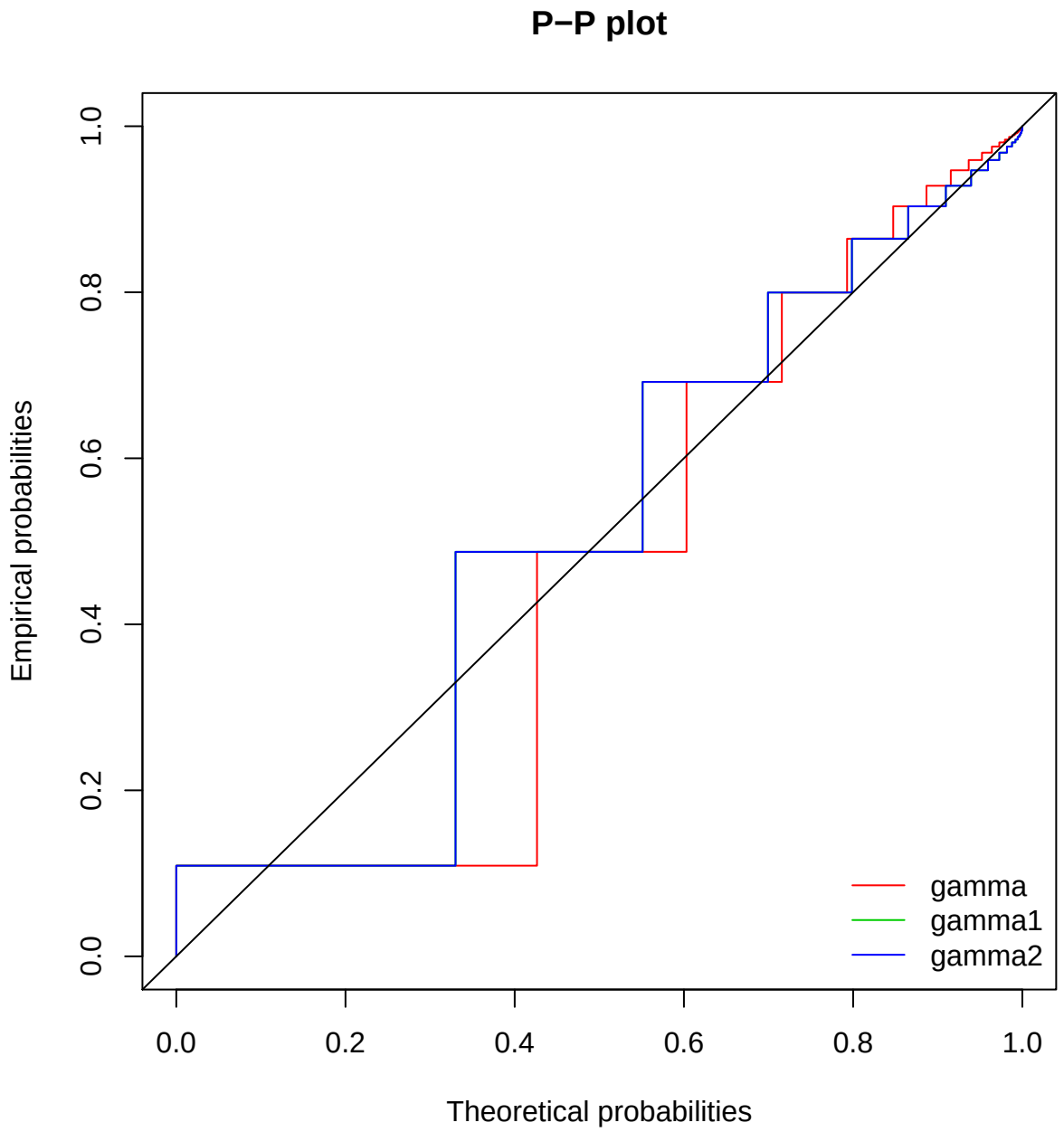
Again, all three distributions appear to be good fits.

```
qqcomp(list(PRESCRIPgamma, PRESCRIPgamma1, PRESCRIPgamma2),
       legendtext = plot.legend)
```



The Q-Q plot shows that the moment matched gamma distribution is the better fit.

```
ppcomp(list(PRESCRIPgamma, PRESCRIPgamma1, PRESCRIPgamma2),
       legendtext = plot.legend)
```



The P-P plot shows that all the distributions appear to be good fits.

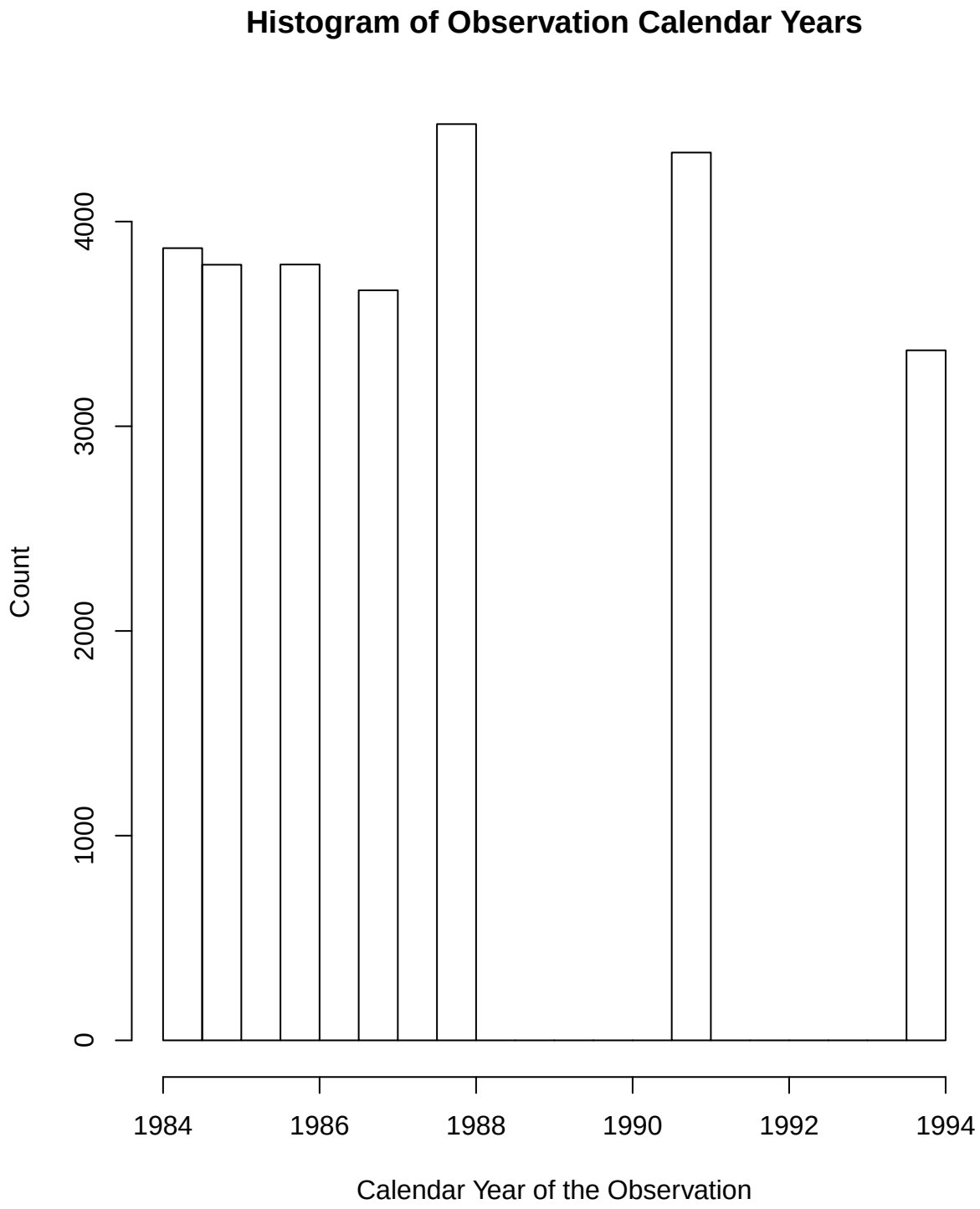
Conclusion about PRESCRIP

We conclude that the PRESCRIP variable is best approximated by a gamma distribution. A primary reason for this is that a lot of other distributions did not make sense based on the Cullen and Frey graph and also based on the range of values of the dataset.

Histogram and Density Curve for YEAR

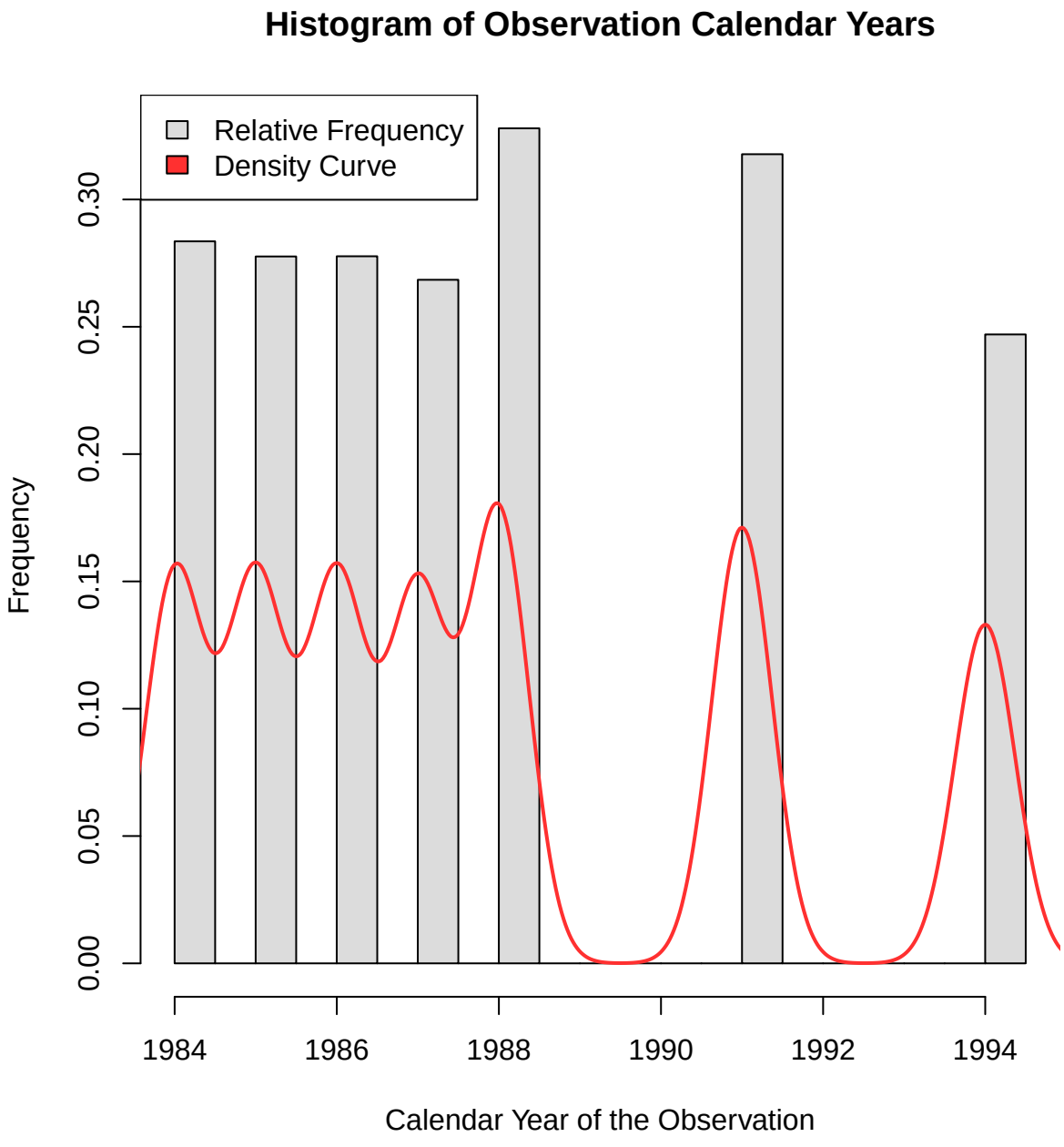
Histogram

```
hist(Pr3c$YEAR, xlab= "Calendar Year of the Observation", ylab= "Count",
     main= "Histogram of Observation Calendar Years")
```



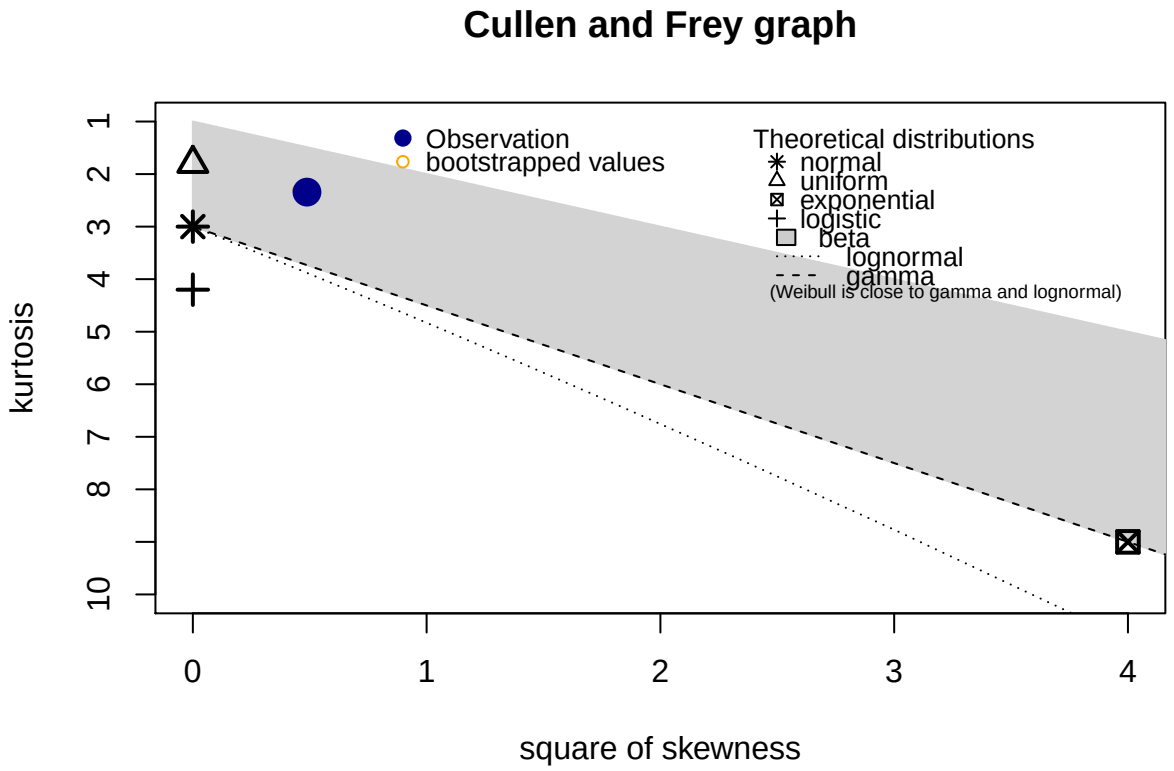
Histogram and Density Curve

```
truehist(Pr3c$YEAR,col="gainsboro", ylab="Frequency",
         xlab= "Calendar Year of the Observation",
         main= "Histogram of Observation Calendar Years")
lines(density((Pr3c$YEAR)), lwd=2,col="firebrick1")
legend("topleft", c("Relative Frequency", "Density Curve"),
       fill=c("gainsboro", "firebrick1"))
```



We plot the Cullen and Frey graph to give us an idea of what distributions to try out:

```
descdist(Pr3c$YEAR, boot = 1000)
```




```
## summary statistics
## ------
## min:  1984   max:  1994
## median:  1987
## mean:  1987.822
## estimated sd:  3.17044
## estimated skewness:  0.6987584
## estimated kurtosis:  2.343381
```

Observe that the normal and uniform distributions are basically the only possibilities, based on both the Cullen and Frey graph and the histogram.

Note that the beta distribution does not make sense due to the range of the values of the dataset not being between [0,1].

We will attempt to fit a normal distribution and a uniform distribution.

Testing fits for distributions

Testing fit for a normal distribution

```
YEARnorm <- fitdist(Pr3c$YEAR, "norm")
```

Testing fit for a uniform distribution

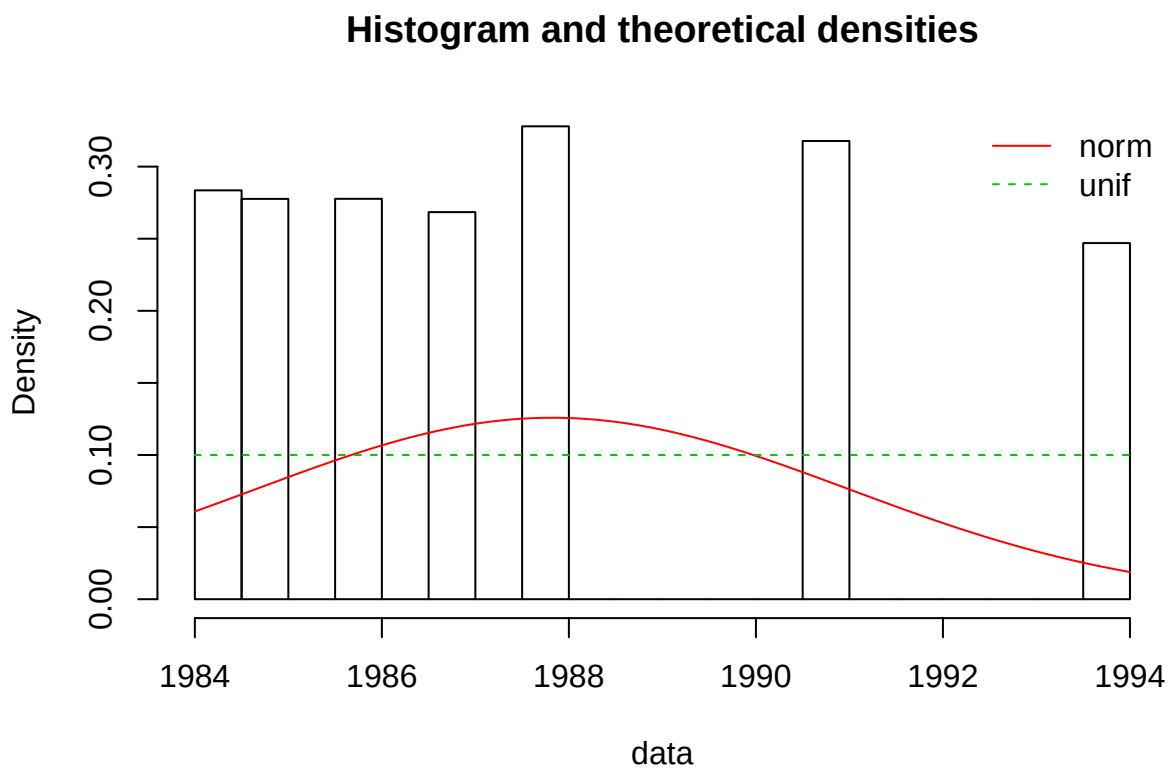
```
YEARunif <- fitdist(Pr3c$YEAR, "unif")
```

Setting Legend

```
plot.legend <- c("norm", "unif")
```

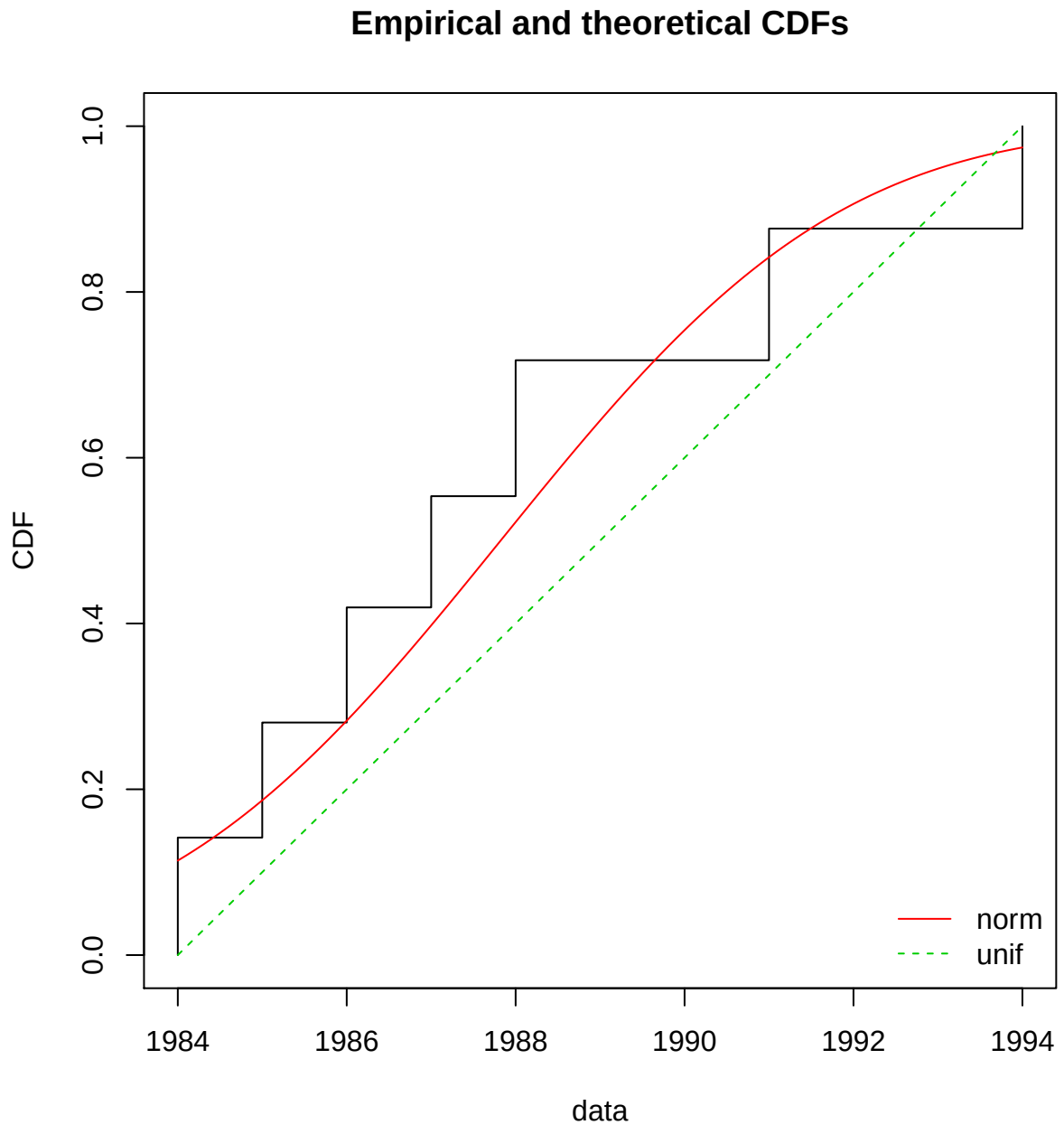
Comparing Histogram and Theoretical Densities

```
denscomp(list(YEARnorm, YEARunif), legendtext = plot.legend)
```



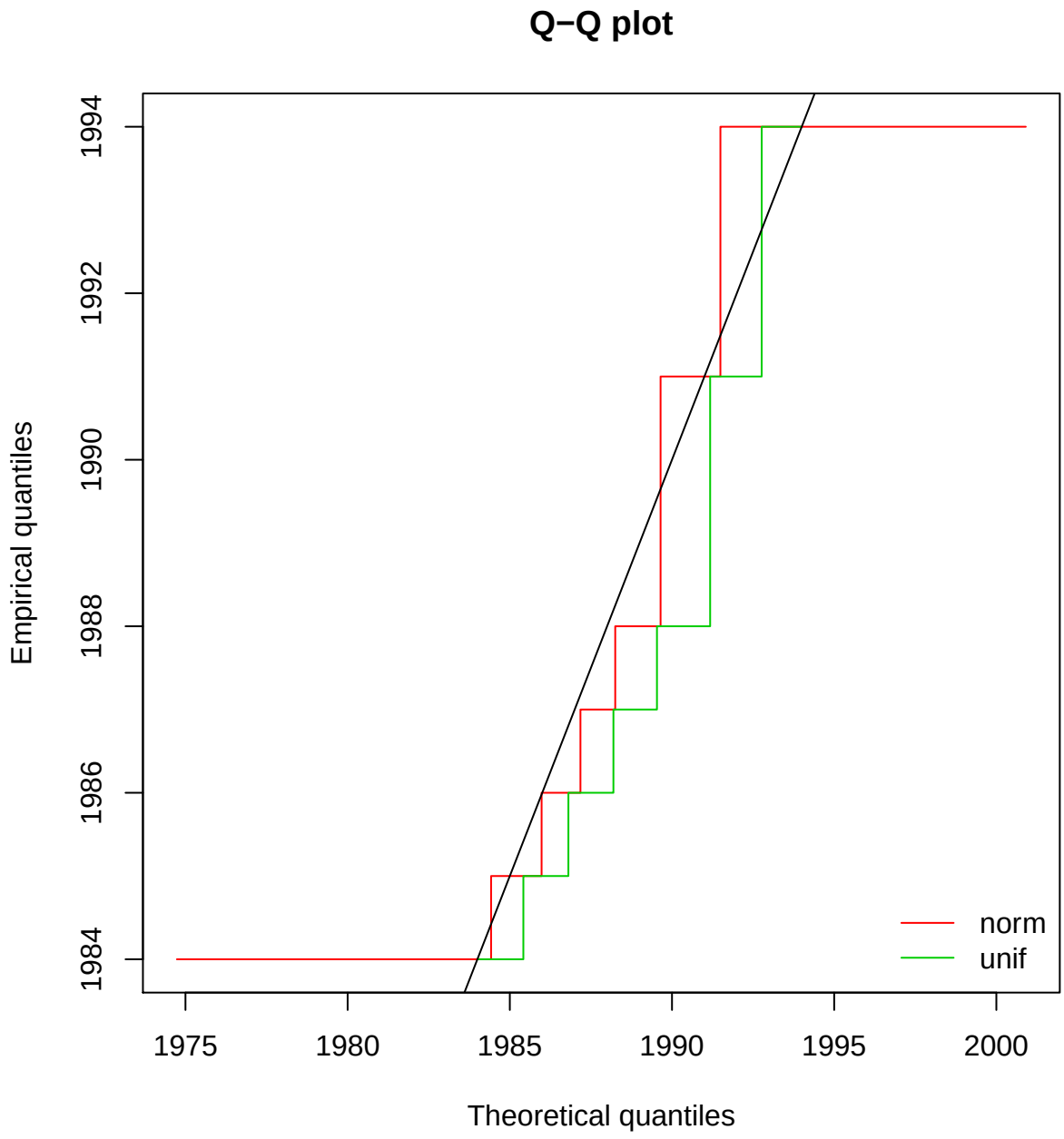
Observe that neither distribution is a good fit.

```
cdfcomp(list(YEARnorm, YEARunif), legendtext = plot.legend)
```



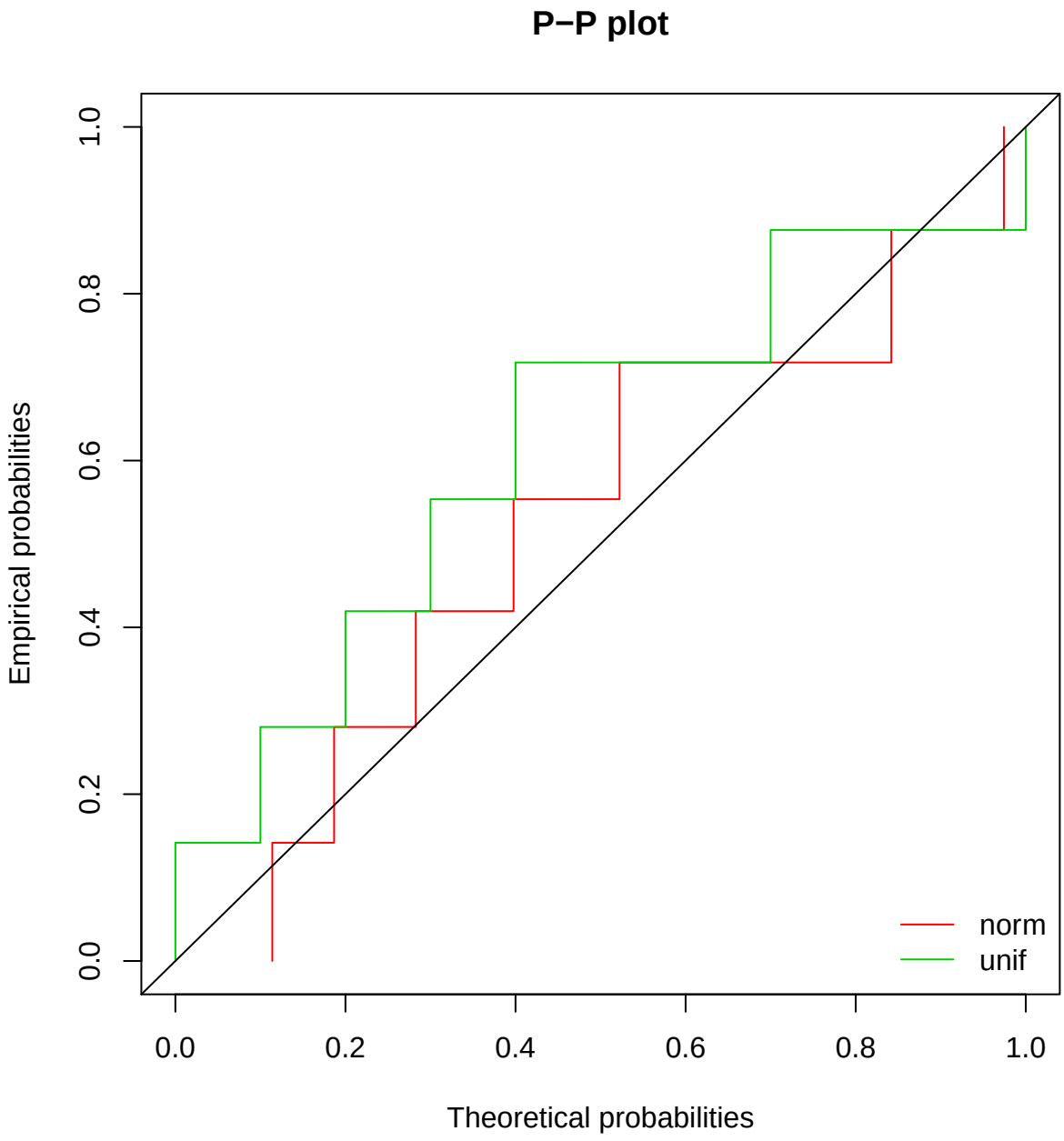
The normal distribution appears to be a better fit than the uniform distribution.

```
qqcomp(list(YEARnorm, YEARunif), legendtext = plot.legend)
```



The Q-Q plot does not provide a clear indication of which distribution is a better fit.

```
ppcomp(list(YEARnorm, YEARunif), legendtext = plot.legend)
```



The P-P plot shows the normal distribution is a marginally better fit.

We analyze goodness-of-fit statistics to get a clearer idea of which distribution is a better approximation, keeping in mind that neither distribution appears to be a particularly good fit to the dataset

```
gofstat(list(YEARnorm, YEARunif),fitnames=c("normal", "uniform"))
```

```
## Goodness-of-fit statistics
##                                  normal     uniform
## Kolmogorov-Smirnov statistic    0.1952603   0.3176246
## Cramer-von Mises statistic    173.6687948 628.8865289
## Anderson-Darling statistic   1100.8922139         Inf
##
## Goodness-of-fit criteria
##                                 normal uniform
## Akaike's Information Criterion 140462.9      NA
## Bayesian Information Criterion 140479.4      NA
```

The goodness-of-fit statistics suggest that the normal distribution is the best fitted distribution.

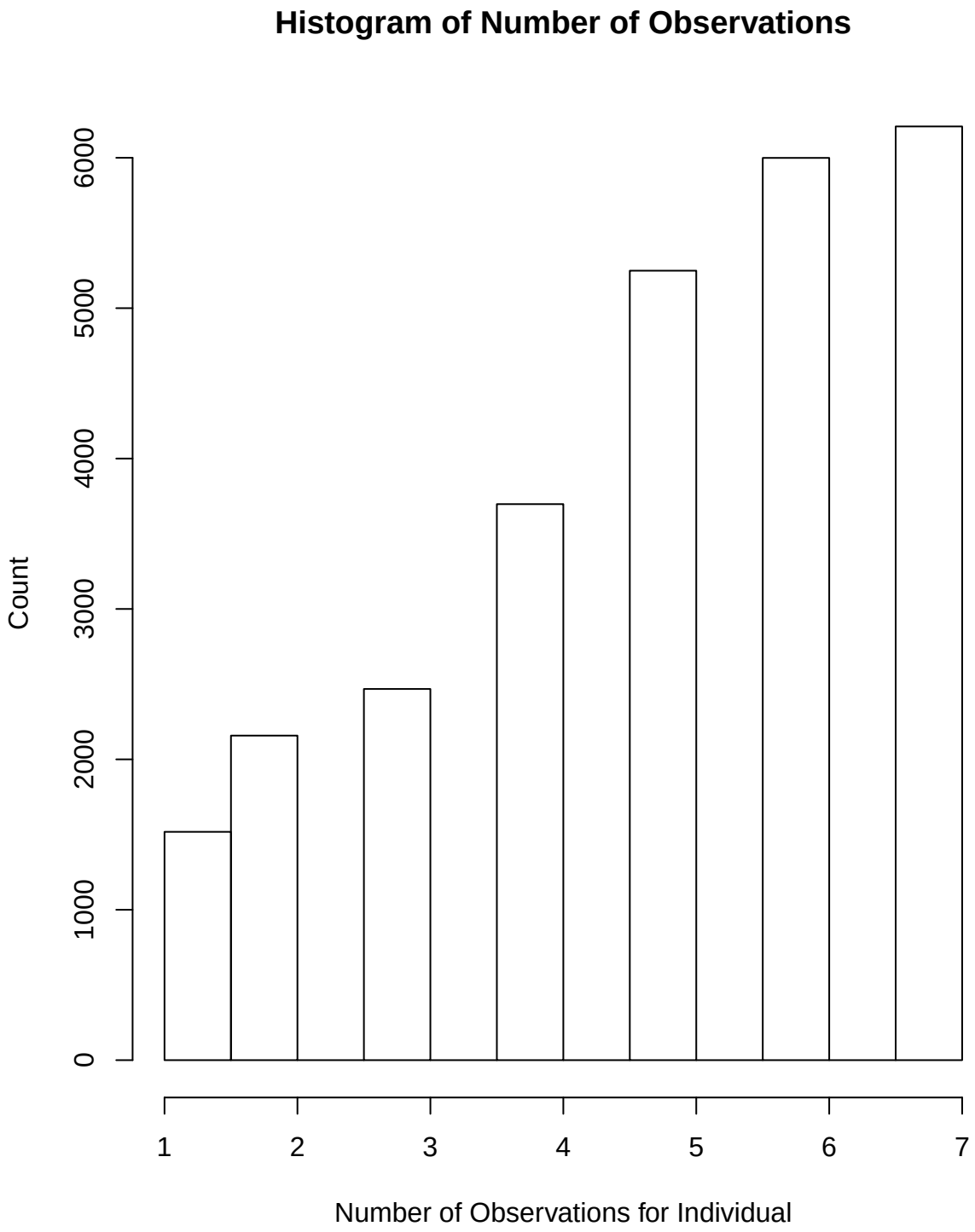
Conclusion about YEAR

We conclude that the normal distribution provides the best approximation of the YEAR variable, among the distributions considered through the Cullen and Frey graph.

Histogram and Density Curve for NUMOBS

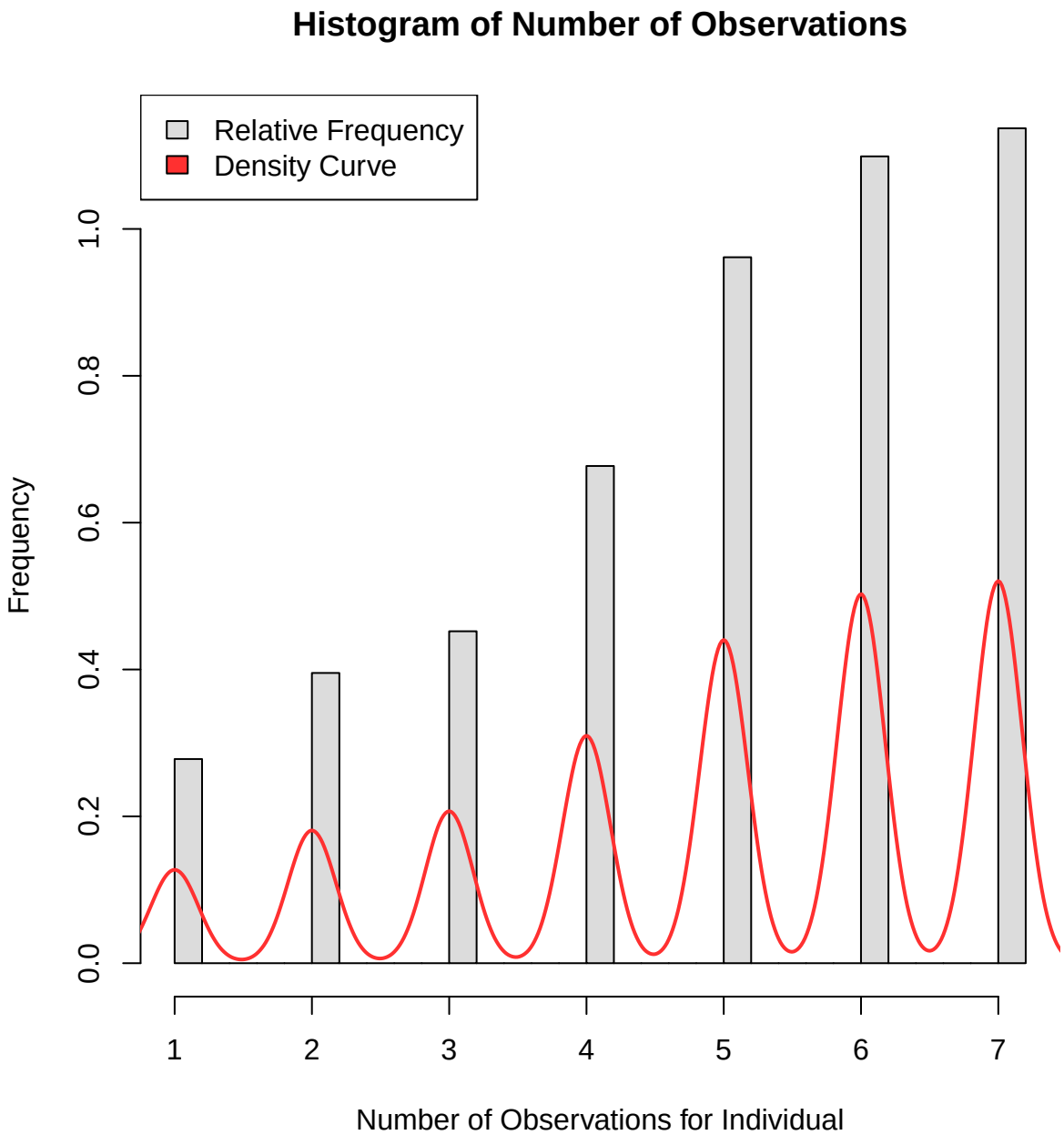
Histogram

```
hist(Pr3c$NUMOBS, xlab= "Number of Observations for Individual",
     ylab= "Count", main= "Histogram of Number of Observations")
```



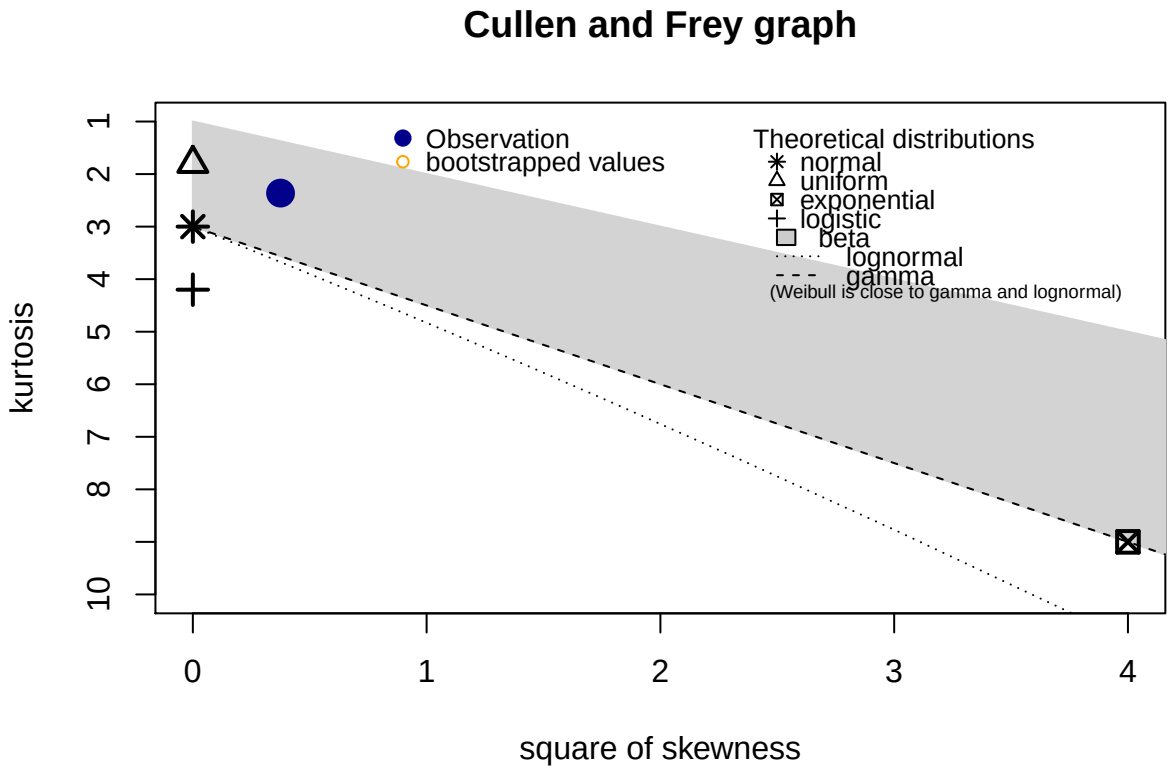
Histogram and Density Curve

```
truehist(Pr3c$NUMOBS,col="gainsboro", ylab="Frequency",
         xlab= "Number of Observations for Individual",
         main= "Histogram of Number of Observations")
lines(density((Pr3c$NUMOBS)), lwd=2,col="firebrick1")
legend("topleft", c("Relative Frequency", "Density Curve"),
       fill=c("gainsboro", "firebrick1"))
```



We plot the Cullen and Frey graph to give us an idea of what distributions to try out:

```
descdist(Pr3c$NUMOBS, boot = 1000)
```




```
## summary statistics
## ------
## min:  1   max:  7
## median:  5
## mean:  4.898743
## estimated sd:  1.793793
## estimated skewness:  -0.6124735
## estimated kurtosis:  2.361859
```

Observe that the normal and uniform distributions are basically the only possibilities, based on both the Cullen and Frey graph.

Note that the beta distribution does not make sense due to the range of the values of the dataset not being between [0,1].

Also, note that due to the shape of the histogram, gamma and lognormal distributions do not make sense.

We will attempt to fit a normal distribution and a uniform distribution.

Testing fits for distributions

Testing fit for a normal distribution

```
NUMOBSnorm <- fitdist(Pr3c$NUMOBS, "norm")
```

Testing fit for a uniform distribution

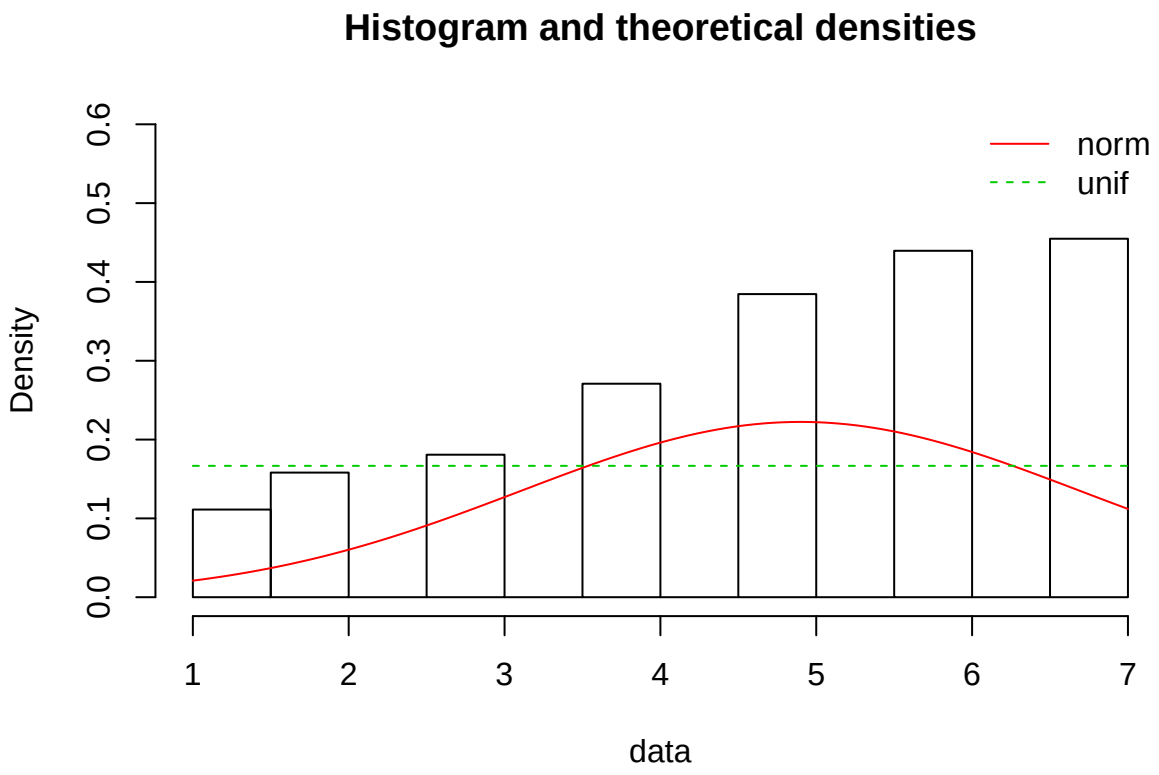
```
NUMOBSunif <- fitdist(Pr3c$NUMOBS, "unif")
```

Setting Legend

```
plot.legend <- c("norm", "unif")
```

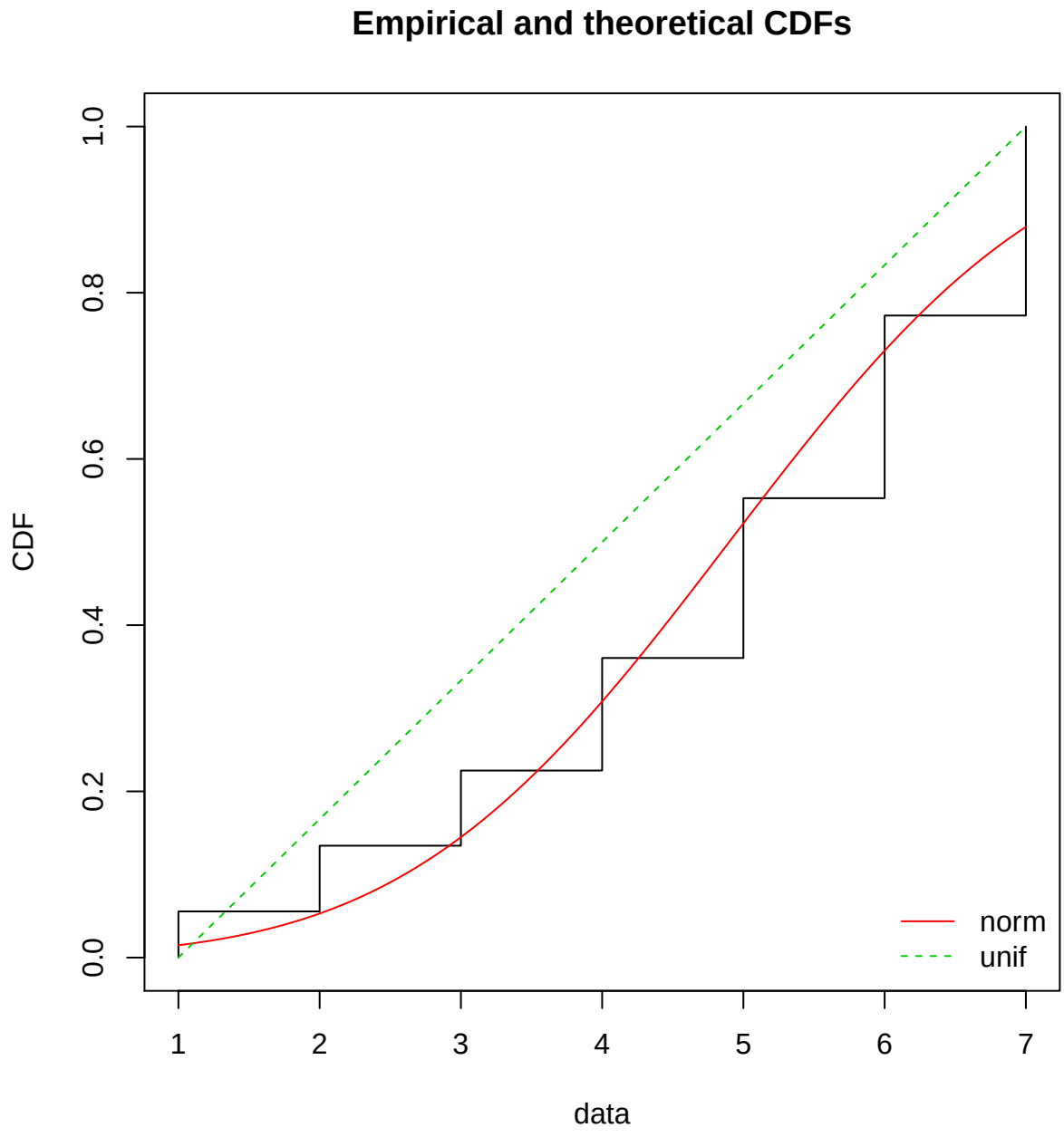
Comparing Histogram and Theoretical Densities

```
denscomp(list(NUMOBSnorm, NUMOBSunif), legendtext = plot.legend, ylim = c(0,0.6))
```



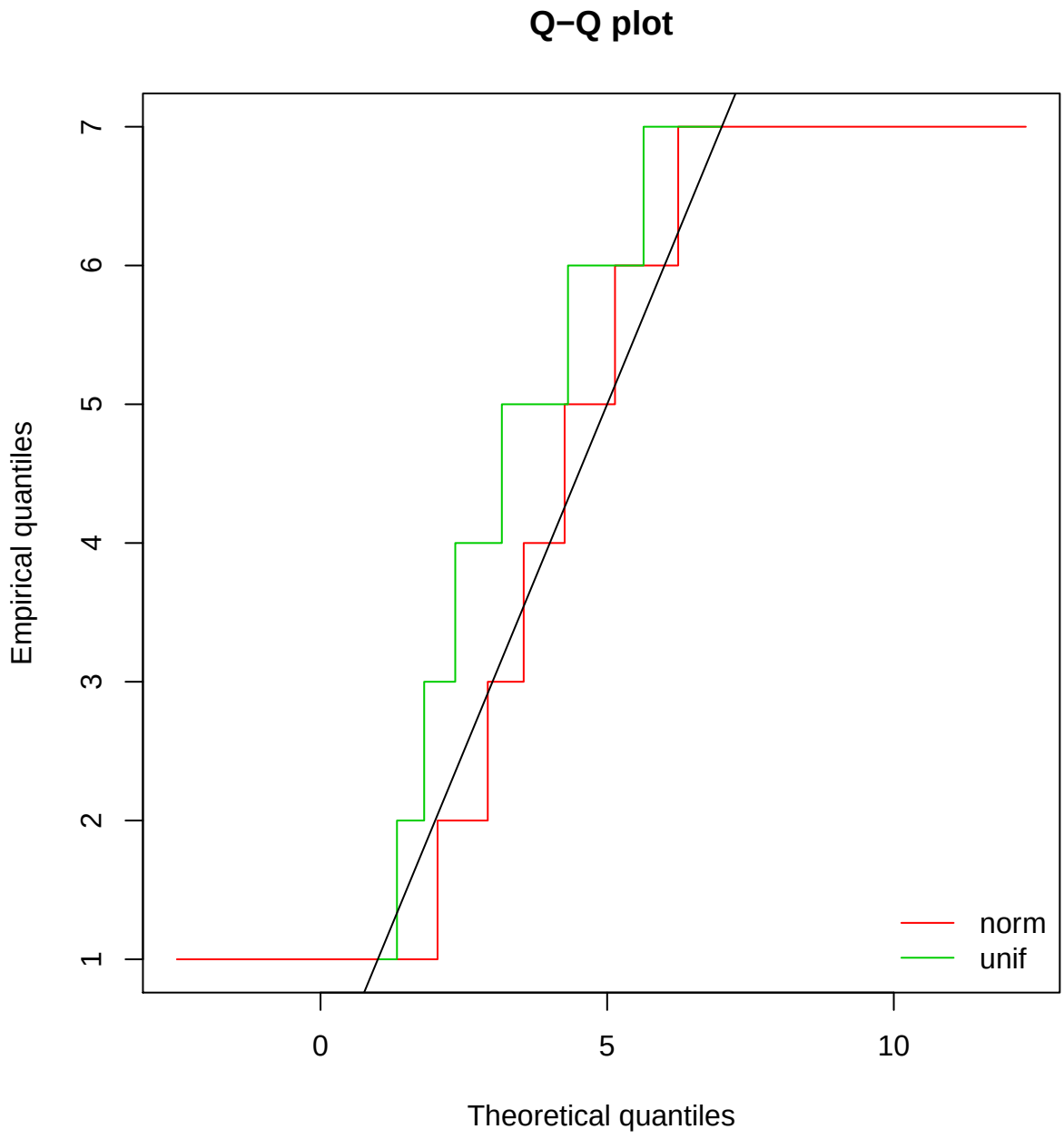
Observe that neither distribution is a good fit.

```
cdfcomp(list(NUMOBSnorm, NUMOBSunif), legendtext = plot.legend)
```



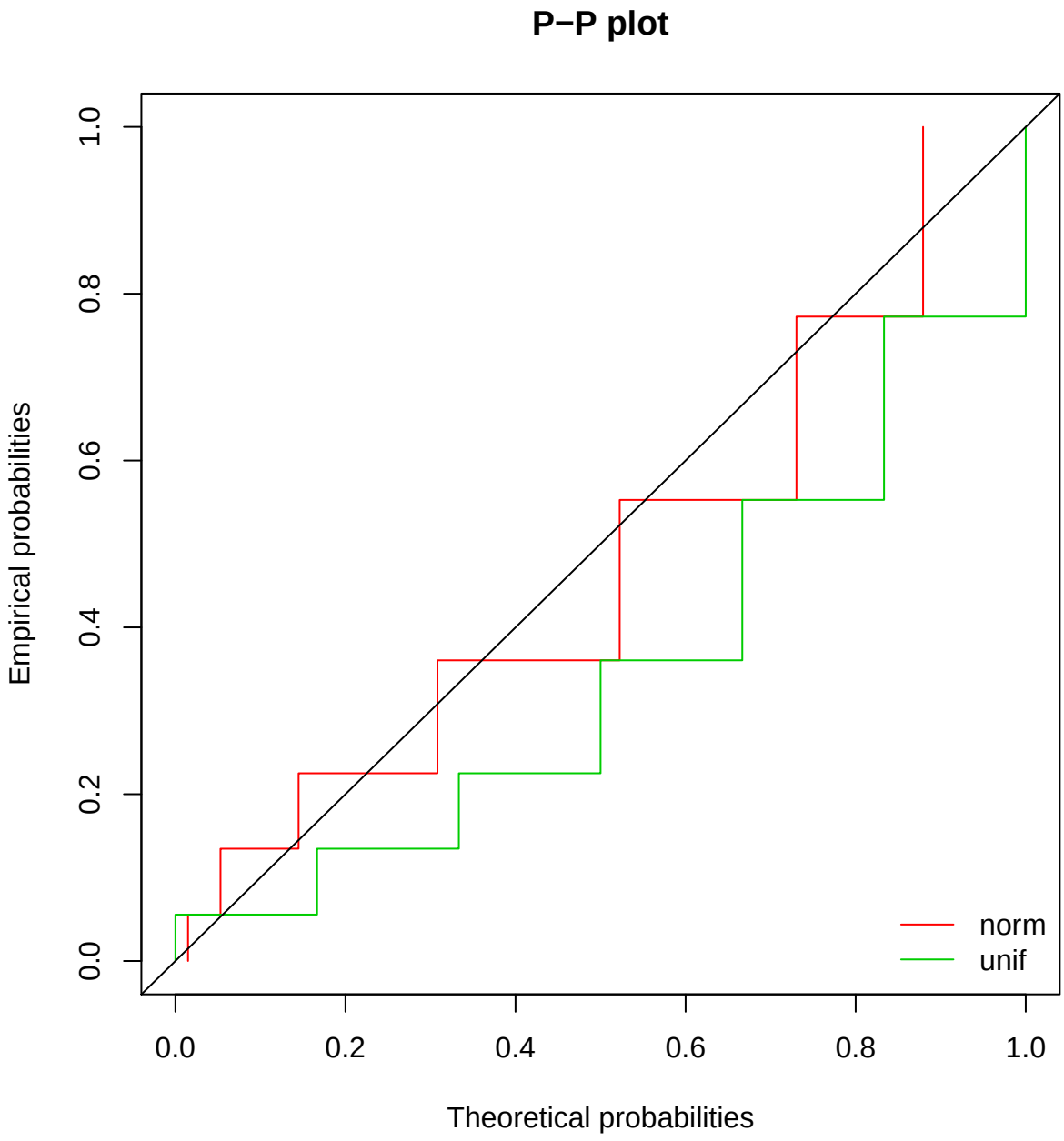
The normal distribution appears to be a better fit than the uniform distribution.

```
qqcomp(list(NUMOBSnorm, NUMOBSunif), legendtext = plot.legend)
```



The Q-Q plot indicates that the normal distribution is a better fit.

```
ppcomp(list(NUMOBSnorm, NUMOBSunif), legendtext = plot.legend)
```



The P-P plot also indicates that the normal distribution is a better fit.

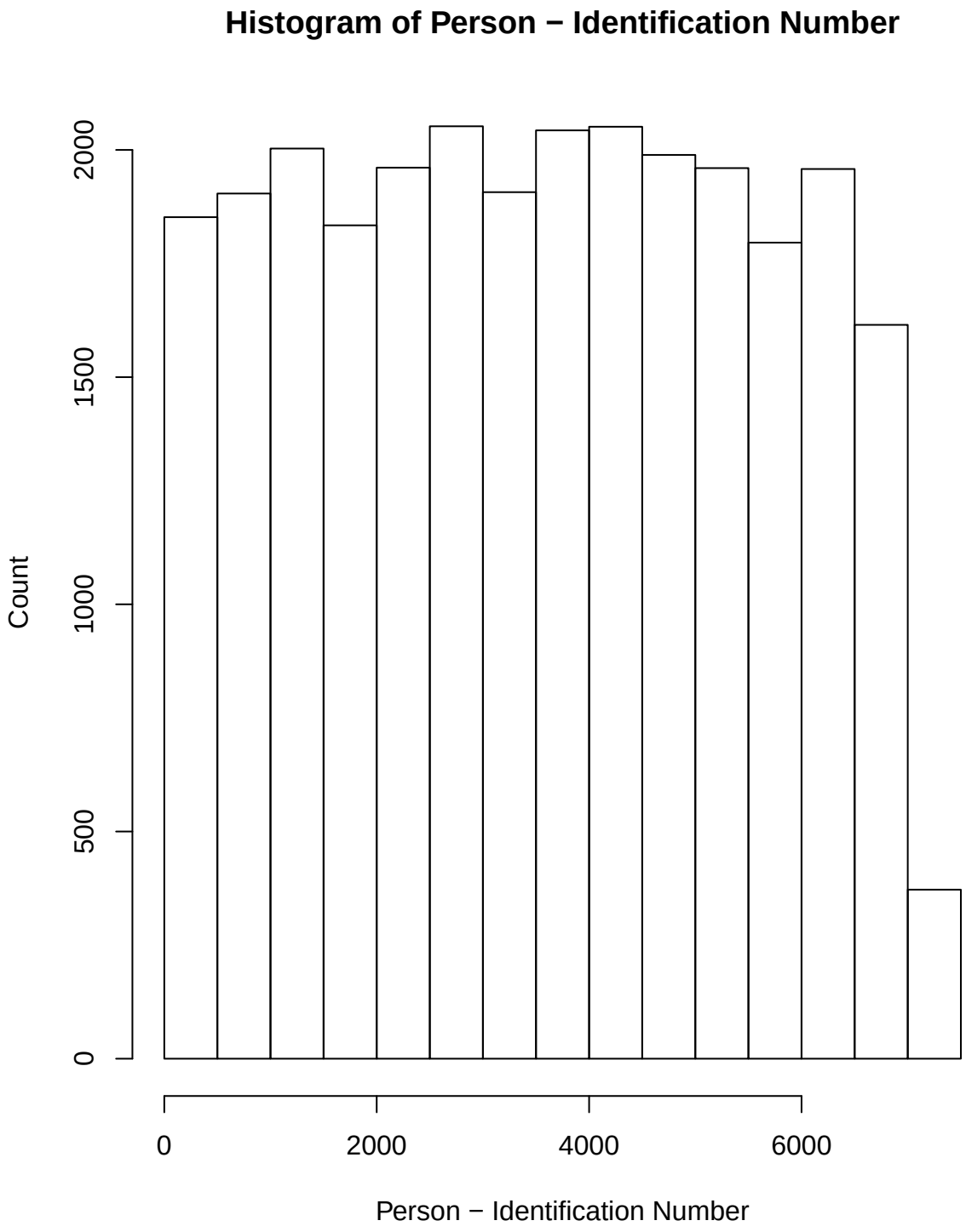
Conclusion about NUMOBS

We conclude that the normal distribution provides the best approximation of the NUMOBS variable, among the distributions considered through the Cullen and Frey graph.

Histogram and Density Curve for ID

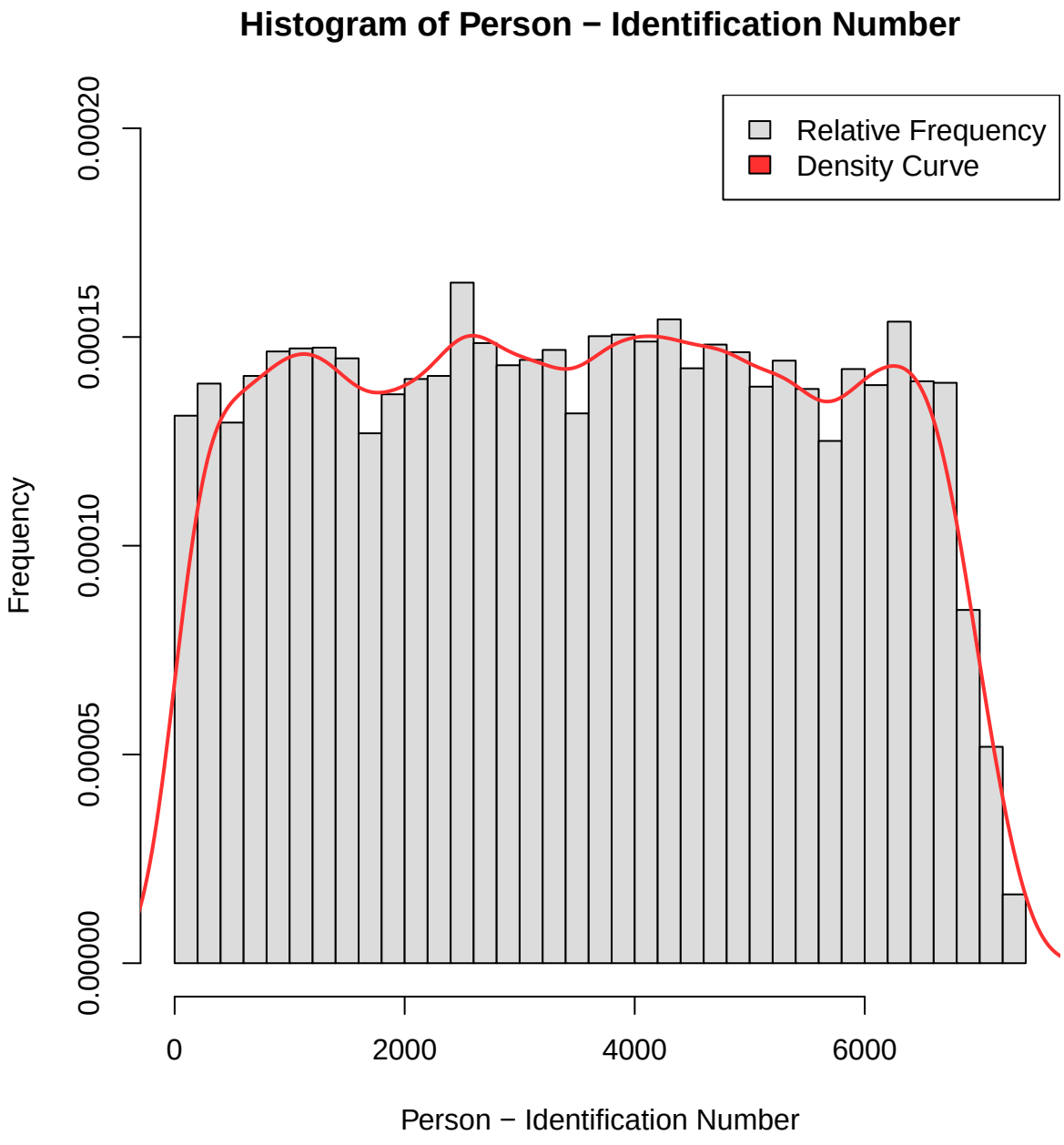
Histogram

```
hist(Pr3c$ID, xlab= "Person - Identification Number", ylab= "Count",
     main= "Histogram of Person - Identification Number")
```



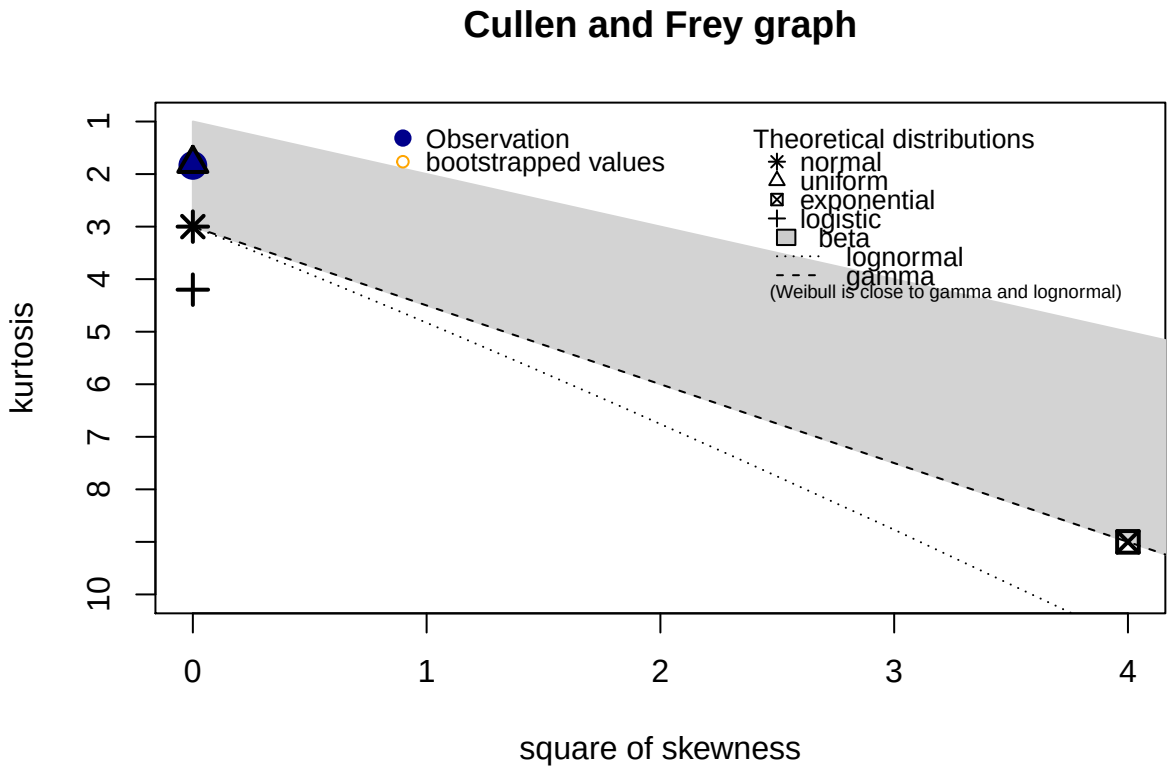
Histogram and Density Curve

```
truehist(Pr3c$ID,col="gainsboro", ylab="Frequency",
         xlab= "Person - Identification Number",
         main= "Histogram of Person - Identification Number", ylim = c(0,0.00020))
lines(density((Pr3c$ID)), lwd=2,col="firebrick1")
legend("topright", c("Relative Frequency", "Density Curve"),
       fill=c("gainsboro", "firebrick1"))
```



We plot the Cullen and Frey graph to give us an idea of what distributions to try out:

```
descdist(Pr3c$ID, boot = 1000)
```




```
## summary statistics
## ------
## min:  1   max:  7292
## median:  3538
## mean:  3519.828
## estimated sd:  2012.385
## estimated skewness:  0.007844538
## estimated kurtosis:  1.834849
```

Observe that the uniform distribution is basically the only possibility, based on both the Cullen and Frey graph and the histogram.

We will attempt to fit a uniform distribution.

Testing fits for distributions

Testing fit for a uniform distribution

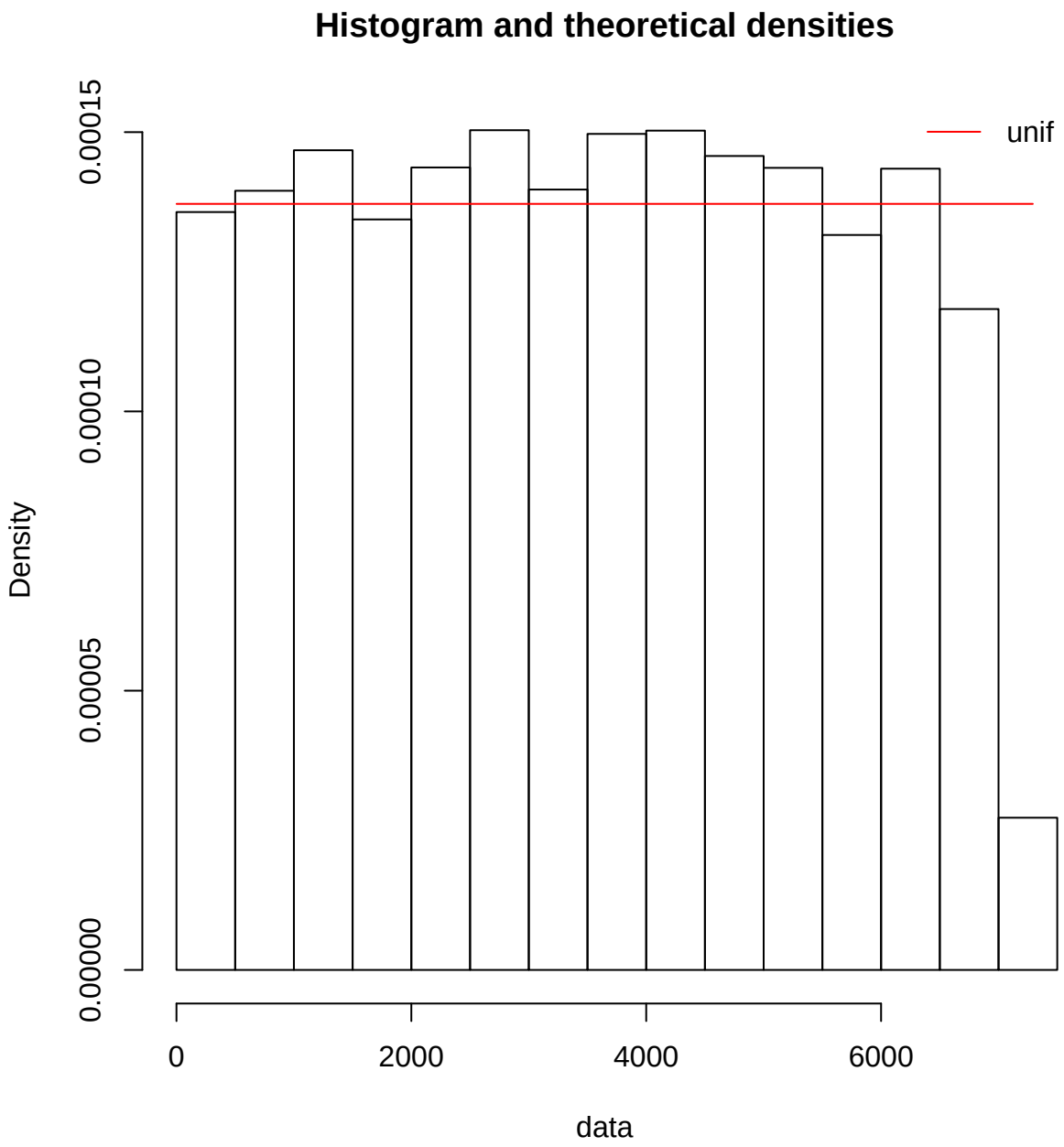
```
IDunif <- fitdist(Pr3c$ID, "unif")
```

Setting Legend

```
plot.legend <- c("unif")
```

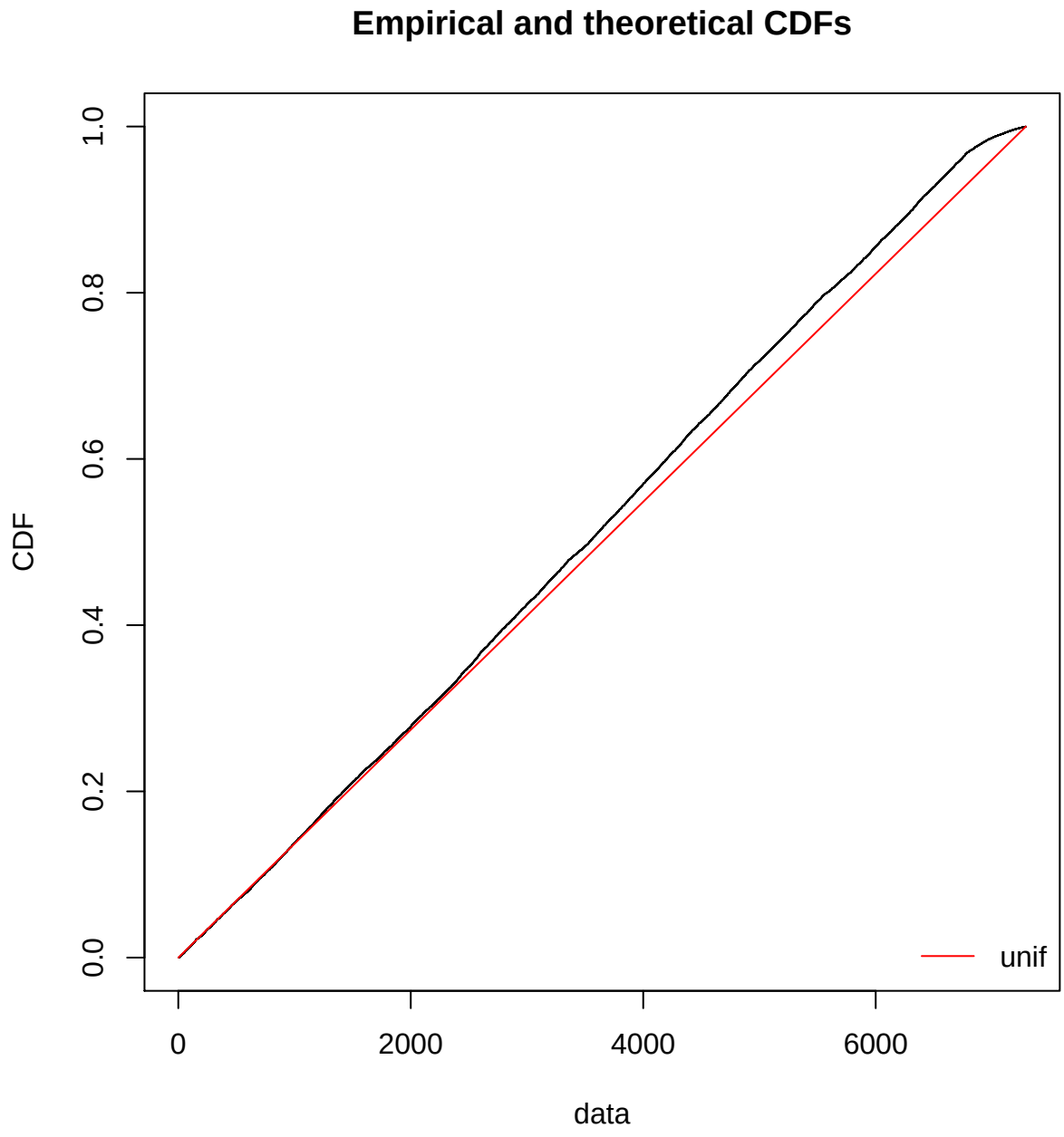
Comparing Histogram and Theoretical Densities

```
denscomp(list(IDunif), legendtext = plot.legend)
```



Observe that this is almost certainly a uniform distribution.

```
cdfcomp(list(IDunif), legendtext = plot.legend)
```



The data set is best approximated by a uniform distribution.

Conclusion about ID

We conclude that the uniform distribution provides the best approximation of the ID variable.

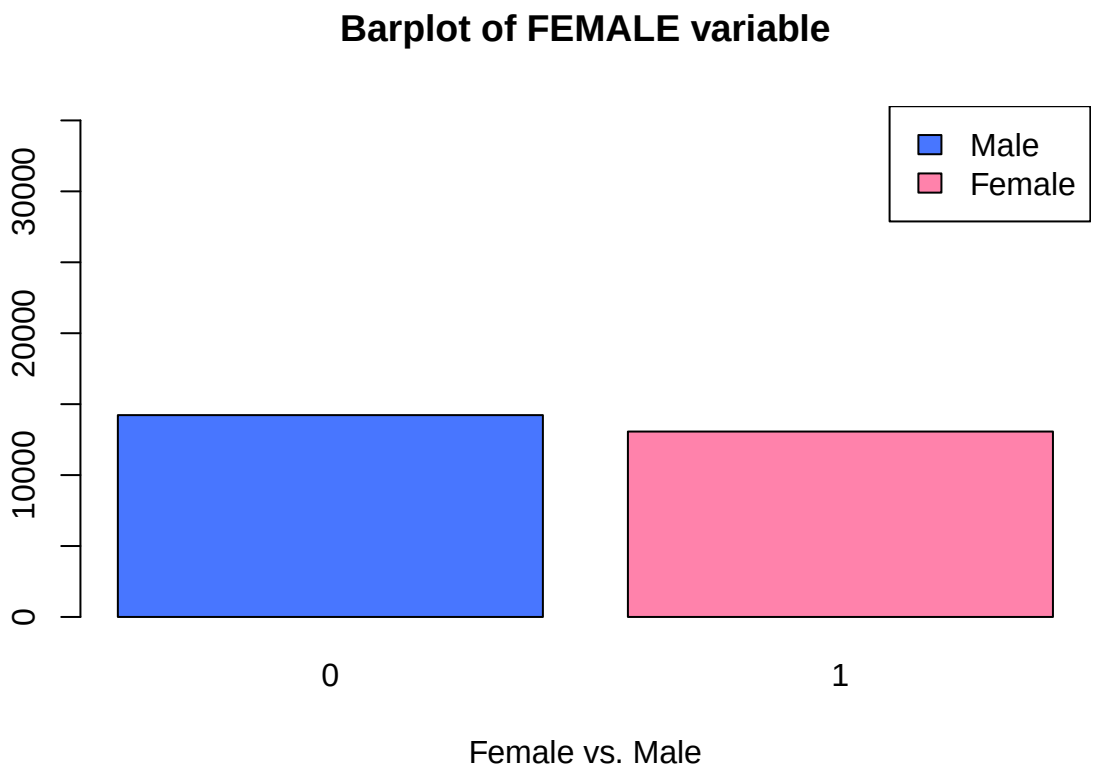
Part I (b)

Note that the categorical variables are:

- FEMALE
- HANDDUM
- FAMHIST
- HHKIDS
- MARRIED
- HAUPTS
- REALS
- FACHHS
- ABITUR
- UNIV
- WORKING
- BLUEC
- WHITEC
- SELF
- BEAMT
- UNEMPLOY
- PUBLIC
- ADDON
- DOCTOR
- HOSPITAL
- HEALTHY
- YEAR1984
- YEAR1985
- YEAR1986
- YEAR1987
- YEAR1988
- YEAR1991
- YEAR1994

Barplot of variable FEMALE

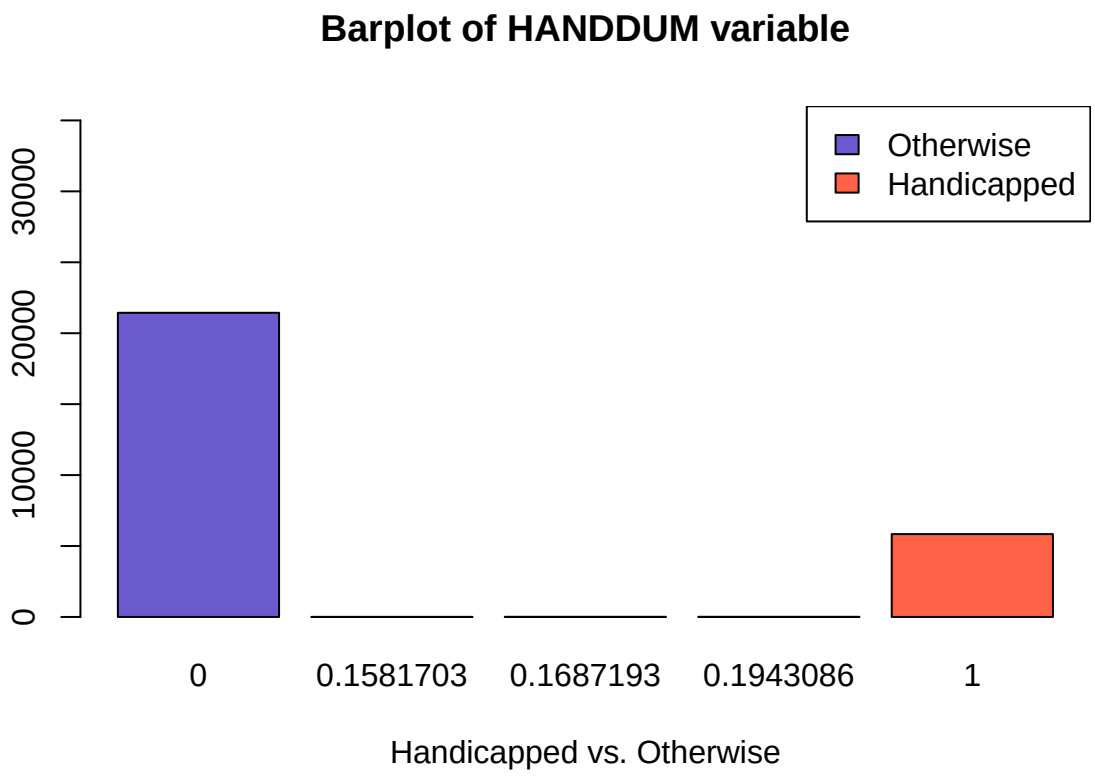
```
counts<-table(Pr3c$FEMALE)
barplot(counts,main="Barplot of FEMALE variable",
        xlab="Female vs. Male",
        col=c("royalblue1","palevioletred1"),
        ylim=c(0,36000))
legend('topright', c("Male","Female"),fill=c("royalblue1","palevioletred1"))
```



There appear to be more males in the dataset than females. However, the numbers are not so skewed that they significantly bias the data set.

Barplot of variable HANDDUM

```
counts<-table(Pr3c$HANDDUM)
barplot(counts,main="Barplot of HANDDUM variable",
        xlab="Handicapped vs. Otherwise",
        col=c("slateblue","red","red","red","tomato"),
        ylim=c(0,36000))
legend('topright',c("Otherwise","Handicapped"),fill=c("slateblue","tomato"))
```

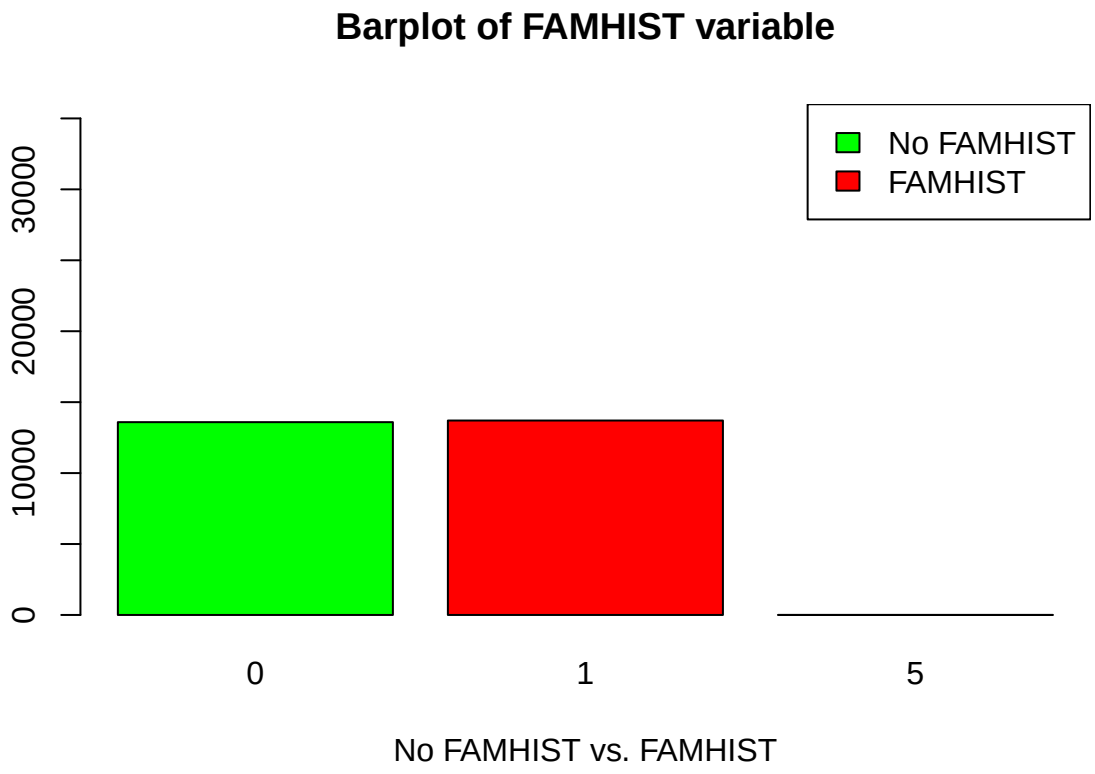


Observe that there are about four more times the individuals who are not handicapped than who are handicapped.

Also note that there are some obvious data errors.

Barplot of variable FAMHIST

```
counts<-table(Pr3c$FAMHIST)
barplot(counts,main="Barplot of FAMHIST variable",
        xlab="No FAMHIST vs. FAMHIST",
        col=c("green","red"),
        ylim = c(0,36000))
legend('topright', c("No FAMHIST","FAMHIST"),fill=c("green","red"))
```

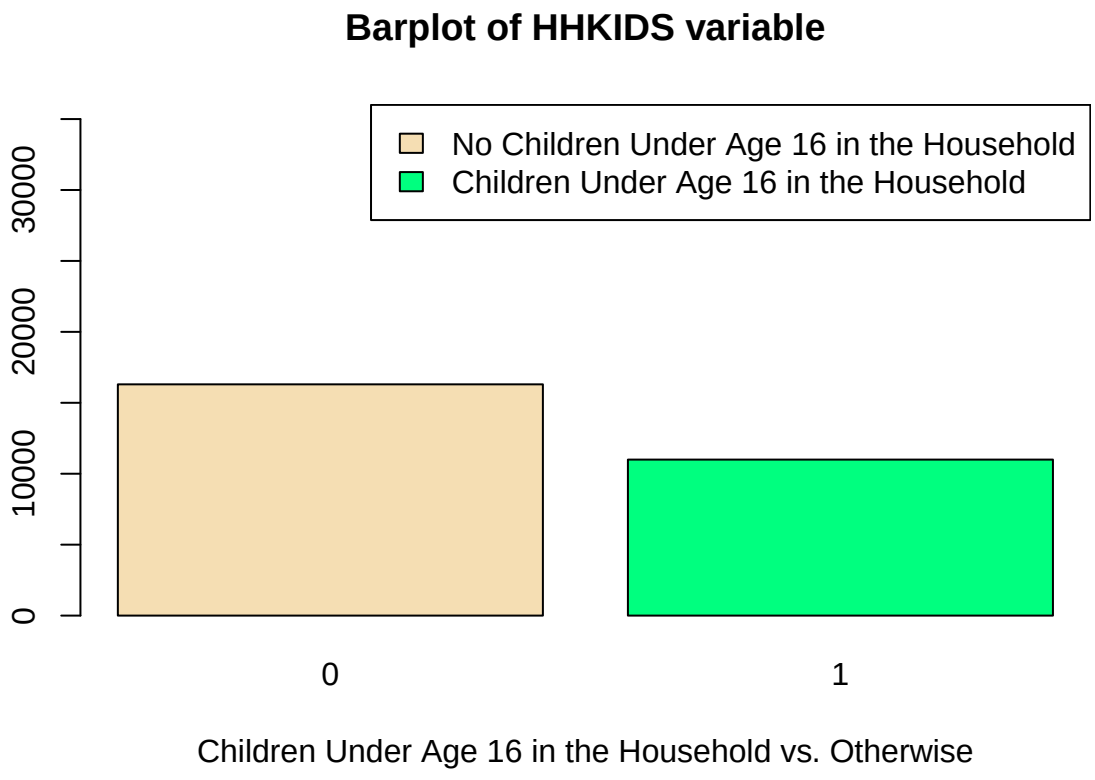


There are about as many individuals in the dataset with FAMHIST as with no FAMHIST.

Note that there is a likely clerical error in the dataset.

Barplot of variable HHKIDS

```
counts<-table(Pr3c$HHKIDS)
barplot(counts,main="Barplot of HHKIDS variable",
        xlab="Children Under Age 16 in the Household vs. Otherwise",
        col=c("wheat","springgreen"),
        ylim = c(0,36000))
legend('topright',
       c("No Children Under Age 16 in the Household",
         "Children Under Age 16 in the Household"),
       fill=c("wheat","springgreen"))
```

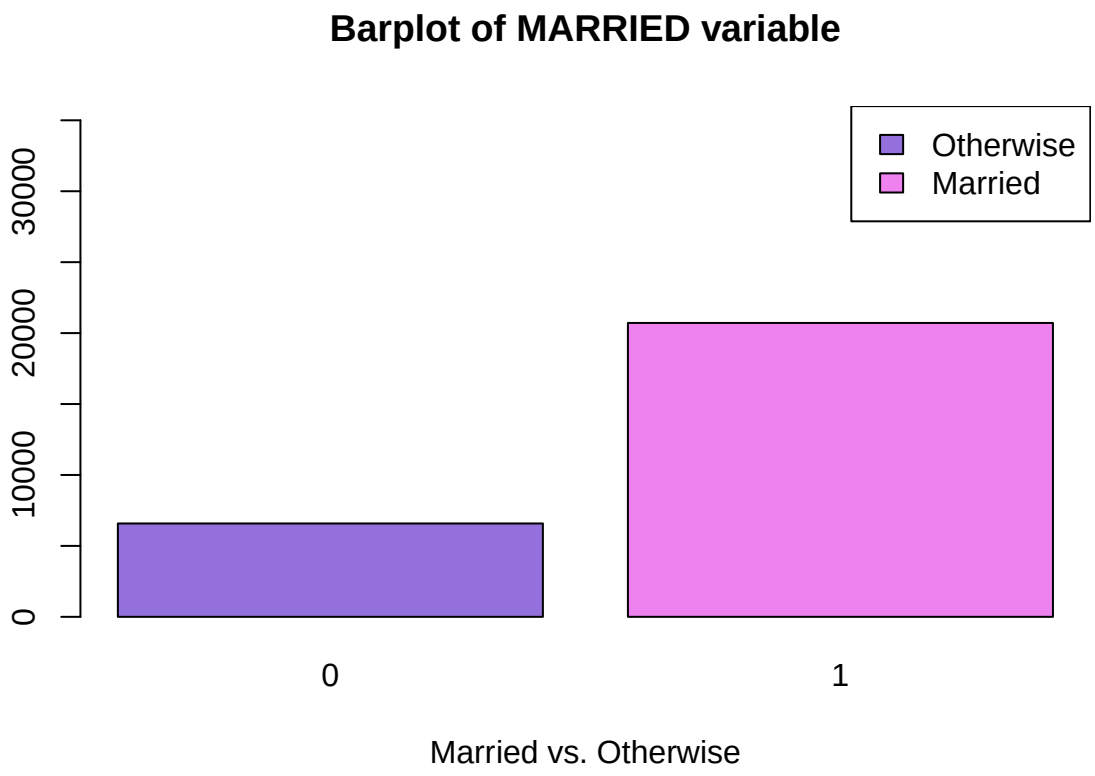


Observe that there are less individuals who have children in the household than who have no children in the household.

Around 40% of individuals have children in the household and the rest do not.

Barplot of variable MARRIED

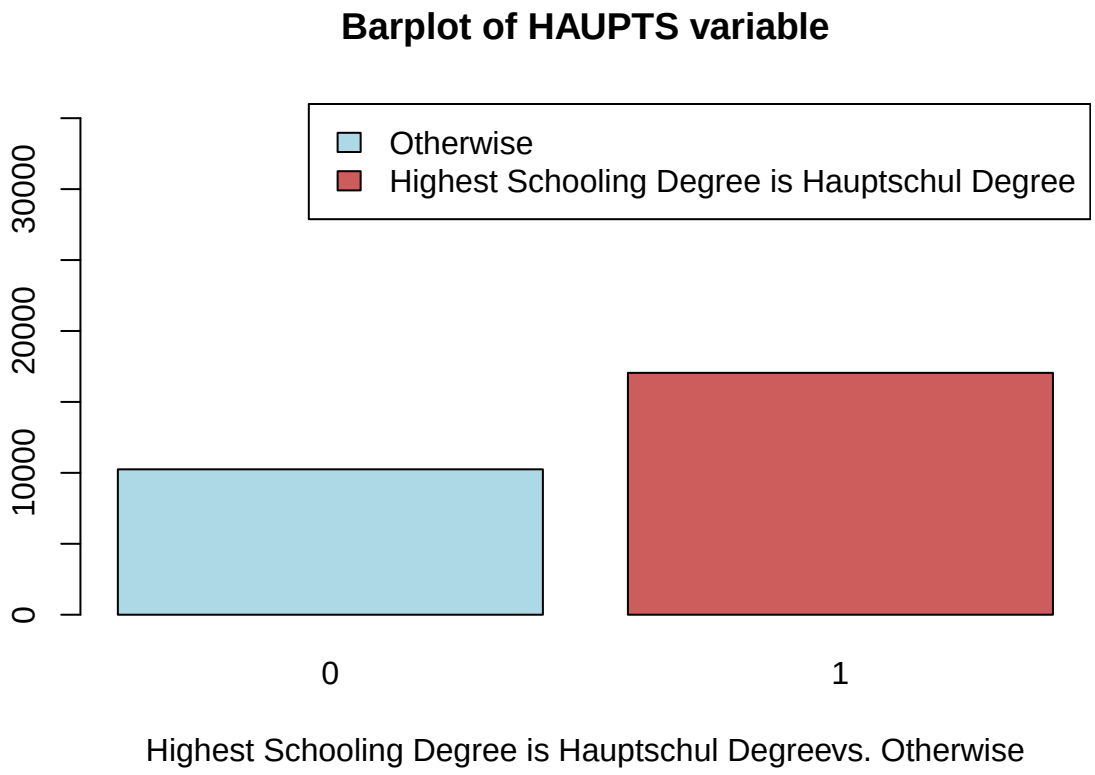
```
counts<-table(Pr3c$MARRIED)
barplot(counts,main="Barplot of MARRIED variable",
        xlab="Married vs. Otherwise",
        col=c("mediumpurple","violet"),
        ylim = c(0,36000))
legend('topright',c("Otherwise","Married"),
       fill=c("mediumpurple","violet"))
```



There are substantially more individuals in the dataset who are married than who are not married.

Barplot of variable HAUPTS

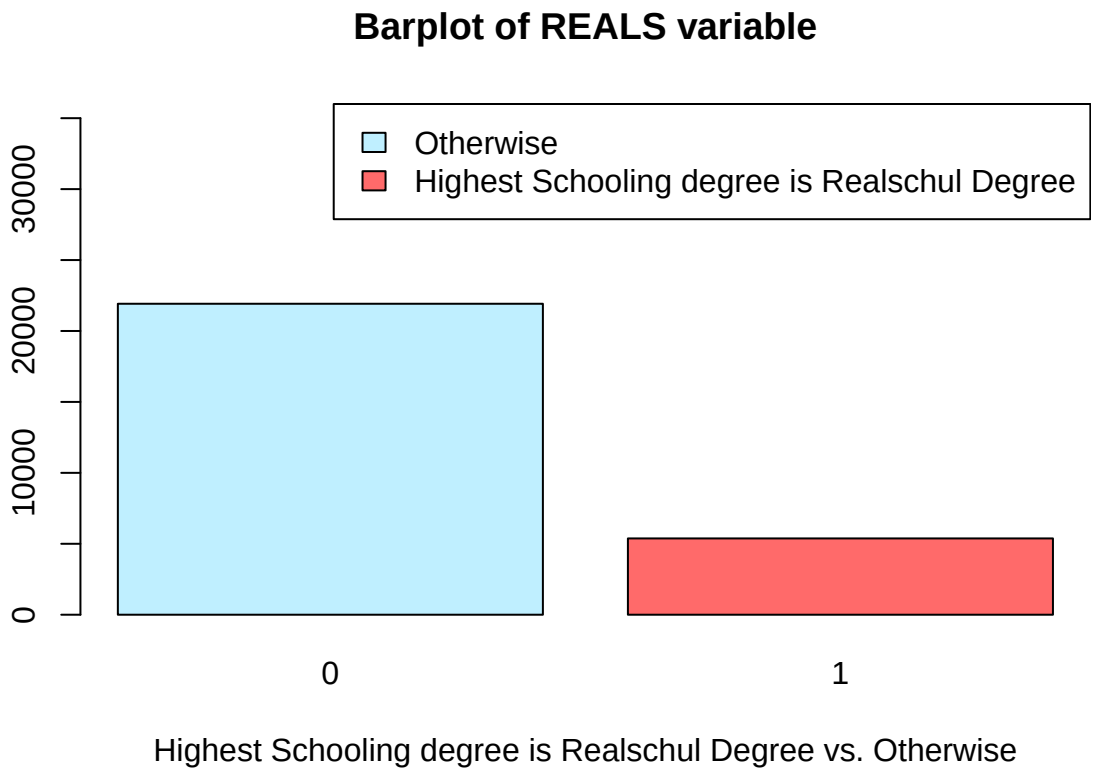
```
counts<-table(Pr3c$HAUPTS)
barplot(counts,main="Barplot of HAUPTS variable",
        xlab="Highest Schooling Degree is Hauptschul Degreevs. Otherwise",
        col=c("lightblue","indianred"),
        ylim = c(0,36000))
legend('topright',
       c("Otherwise", "Highest Schooling Degree is Hauptschul Degree"),
       fill=c("lightblue","indianred"))
```



There are substantially more individuals who have the Hauptschul Degree as their highest educational degree than those who do not.

Barplot of variable REALS

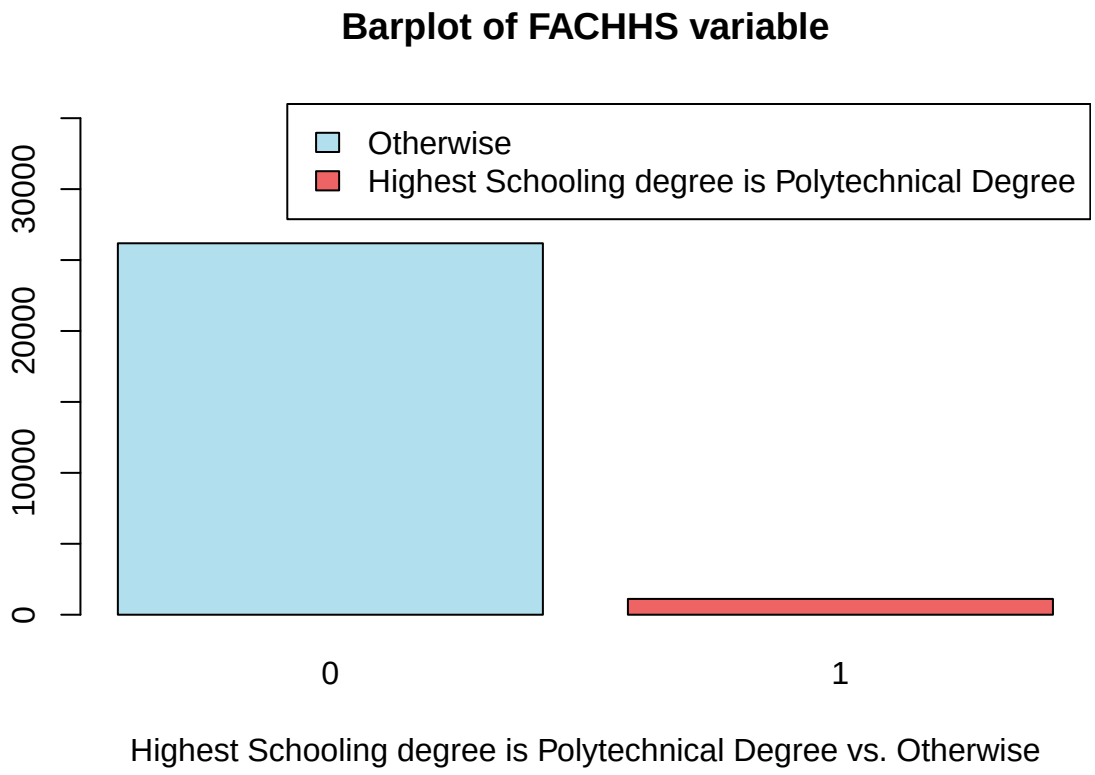
```
counts<-table(Pr3c$REALS)
barplot(counts,main="Barplot of REALS variable",
        xlab="Highest Schooling degree is Realschul Degree vs. Otherwise",
        col=c("lightblue1","indianred1"),
        ylim = c(0, 36000))
legend('topright',
       c("Otherwise","Highest Schooling degree is Realschul Degree"),
       fill=c("lightblue1","indianred1"))
```



The vast majority of individuals do not have the Realschul Degree as their highest educational degree.

Barplot of variable FACHHS

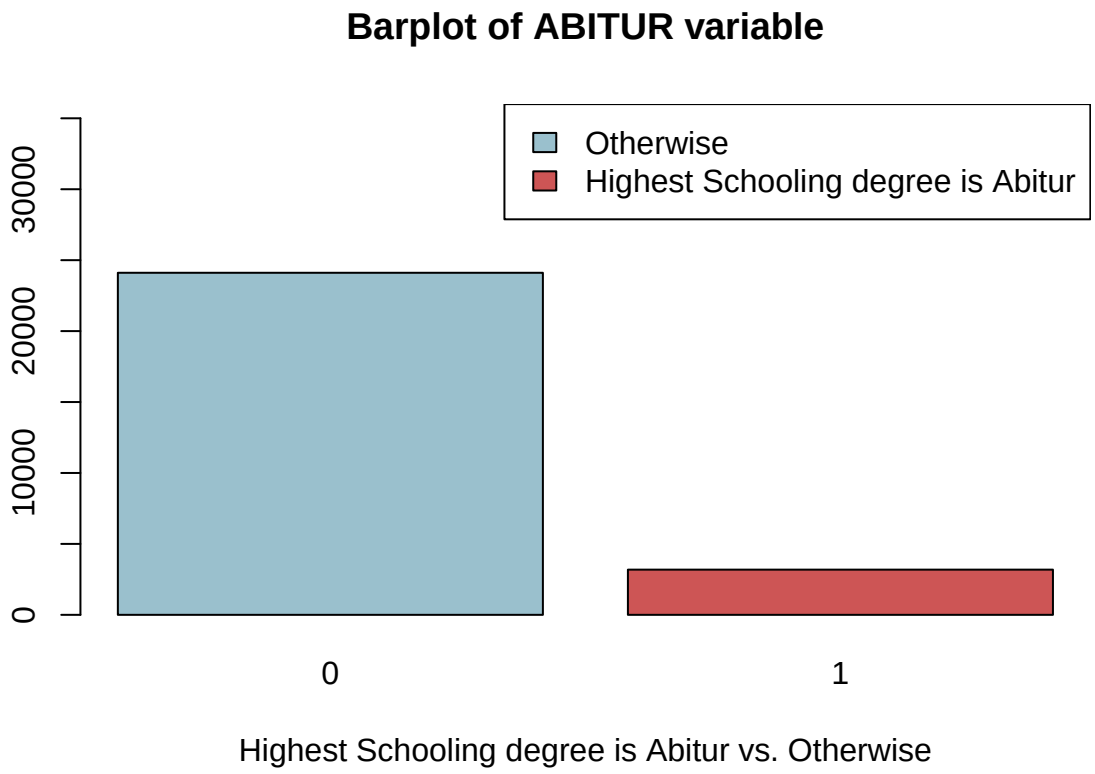
```
counts<-table(Pr3c$FACHHS)
barplot(counts,main="Barplot of FACHHS variable",
        xlab="Highest Schooling degree is Polytechnical Degree vs. Otherwise",
        col=c("lightblue2","indianred2"),
        ylim = c(0,36000))
legend('topright',
       c("Otherwise","Highest Schooling degree is Polytechnical Degree"),
       fill=c("lightblue2","indianred2"))
```



A very small percentage of all individuals have a Polytechnical Degree as their highest educational degree.

Barplot of variable ABITUR

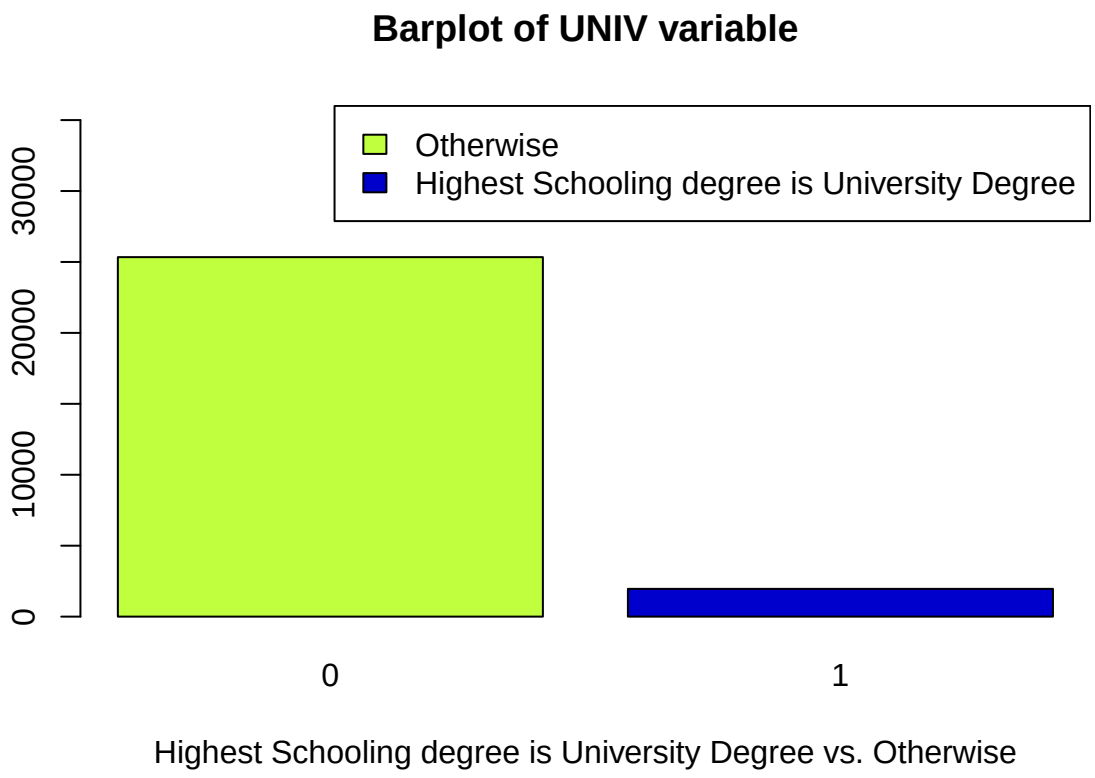
```
counts<-table(Pr3c$ABITUR)
barplot(counts,main="Barplot of ABITUR variable",
        xlab="Highest Schooling degree is Abitur vs. Otherwise",
        col=c("lightblue3","indianred3"),
        ylim = c(0,36000))
legend('topright', c("Otherwise","Highest Schooling degree is Abitur"),
       fill=c("lightblue3","indianred3"))
```



A small proportion of all individuals have an Abitur Degree as their highest educational degree.

Barplot of variable UNIV

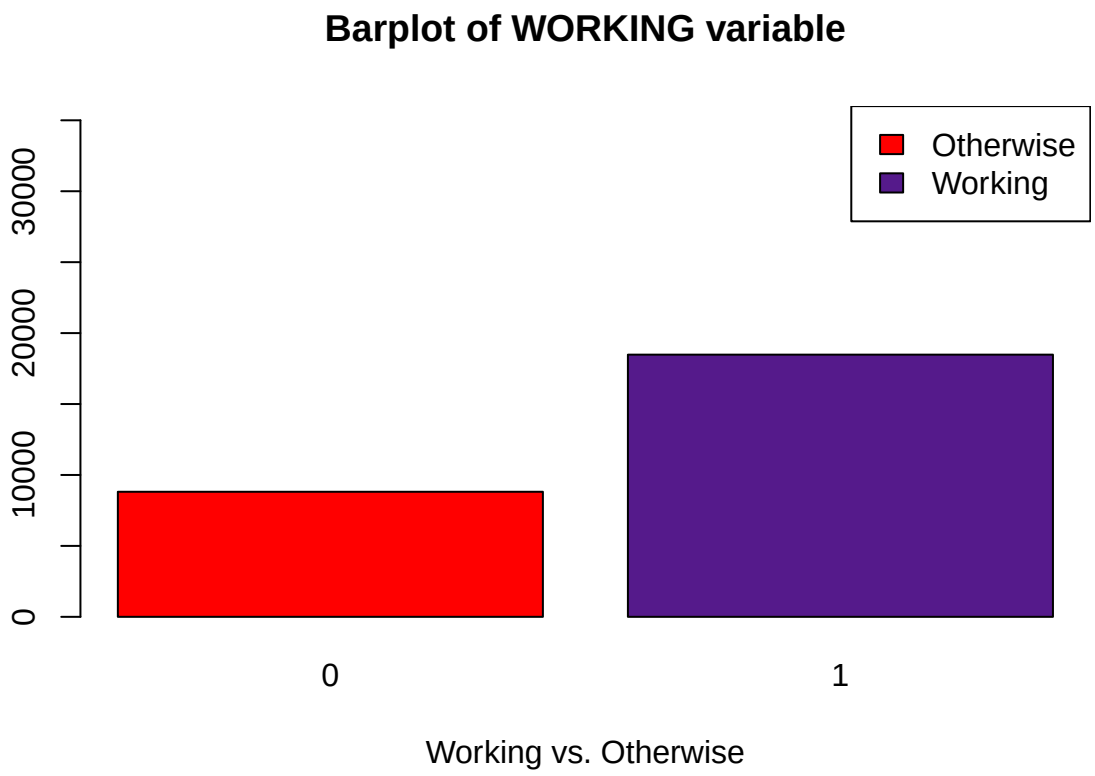
```
counts<-table(Pr3c$UNIV)
barplot(counts,main="Barplot of UNIV variable",
        xlab="Highest Schooling degree is University Degree vs. Otherwise",
        col=c("olivedrab1","mediumblue"),
        ylim = c(0,36000))
legend('topright', c("Otherwise","Highest Schooling degree is University Degree"),
       fill=c("olivedrab1","mediumblue"))
```



The vast majority of individuals do not have a University degree as their highest educational degree.

Barplot of variable WORKING

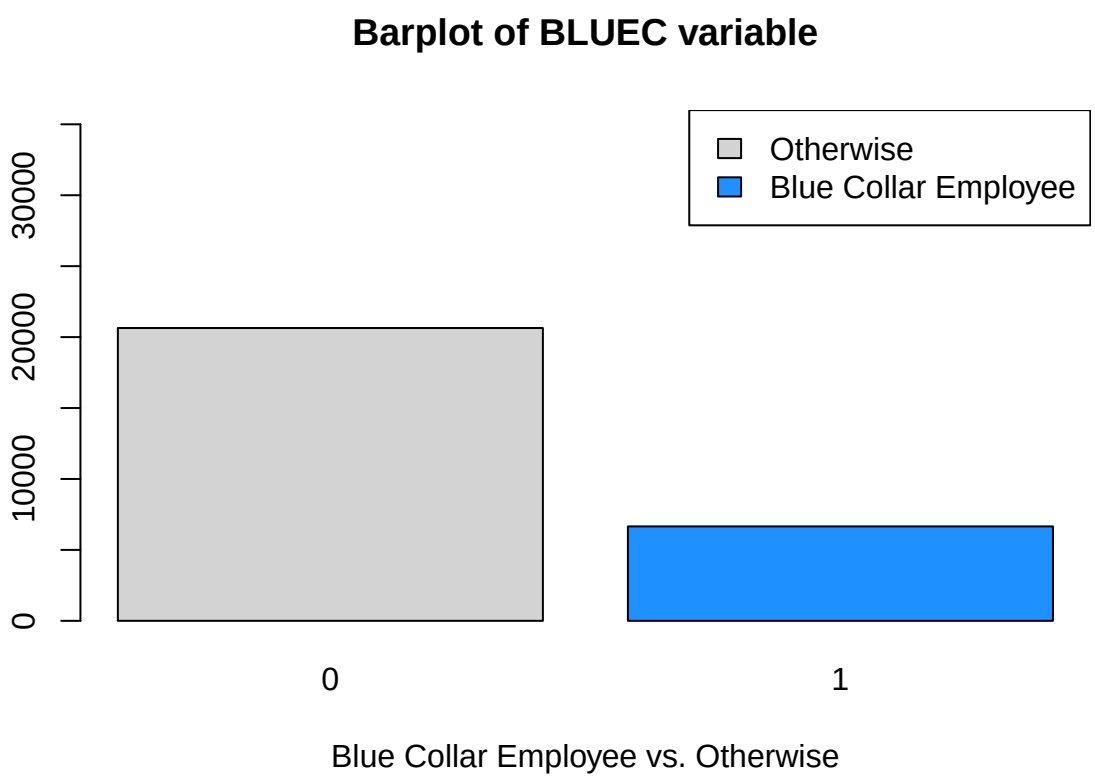
```
counts<-table(Pr3c$WORKING)
barplot(counts,main="Barplot of WORKING variable",
        xlab="Working vs. Otherwise",
        col=c("red","purple4"),
        ylim = c(0,36000))
legend('topright',c("Otherwise","Working"),
       fill=c("red","purple4"))
```



The majority of the individuals are employed.

Barplot of variable BLUEC

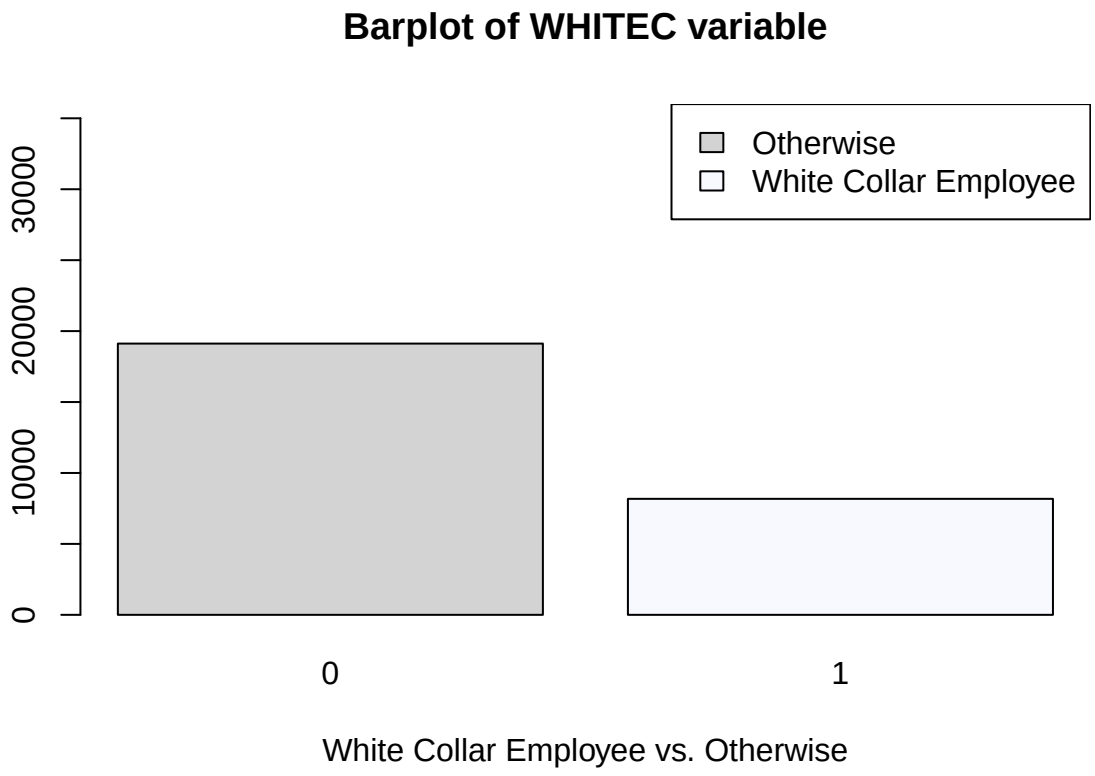
```
counts<-table(Pr3c$BLUEC)
barplot(counts,main="Barplot of BLUEC variable",
        xlab="Blue Collar Employee vs. Otherwise",
        col=c("lightgrey","dodgerblue"),
        ylim = c(0,36000))
legend('topright',c("Otherwise","Blue Collar Employee"),
       fill=c("lightgrey","dodgerblue"))
```



The majority of individuals are not blue collar employees.

Barplot of variable WHITEC

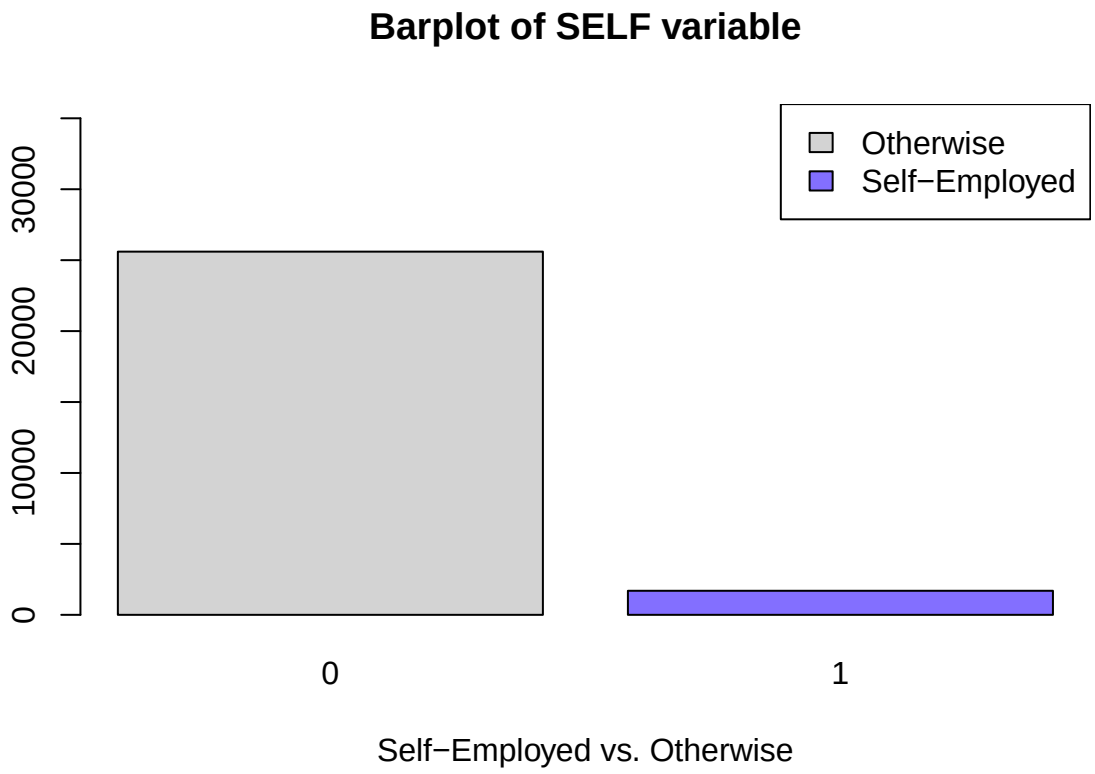
```
counts<-table(Pr3c$WHITEC)
barplot(counts,main="Barplot of WHITEC variable",
        xlab="White Collar Employee vs. Otherwise",
        col=c("lightgrey","ghostwhite"),
        ylim = c(0,36000))
legend('topright',c("Otherwise","White Collar Employee"),
       fill=c("lightgrey","ghostwhite"))
```



The majority of employees are not white collar employees.

Barplot of variable SELF

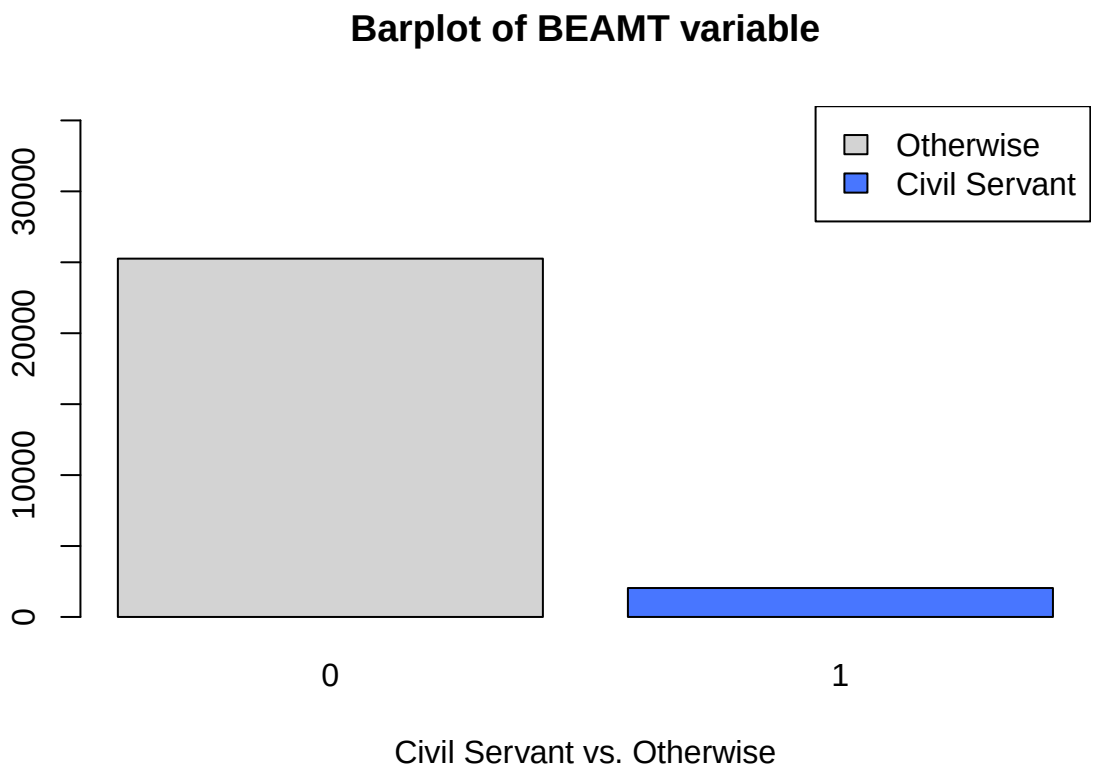
```
counts<-table(Pr3c$SELF)
barplot(counts,main="Barplot of SELF variable",
        xlab="Self-Employed vs. Otherwise",
        col=c("lightgrey","slateblue1"),
        ylim = c(0,36000))
legend('topright',c("Otherwise","Self-Employed"),
       fill=c("lightgrey","slateblue1"))
```



The vast majority of individuals are not self-employed.

Barplot of variable BEAMT

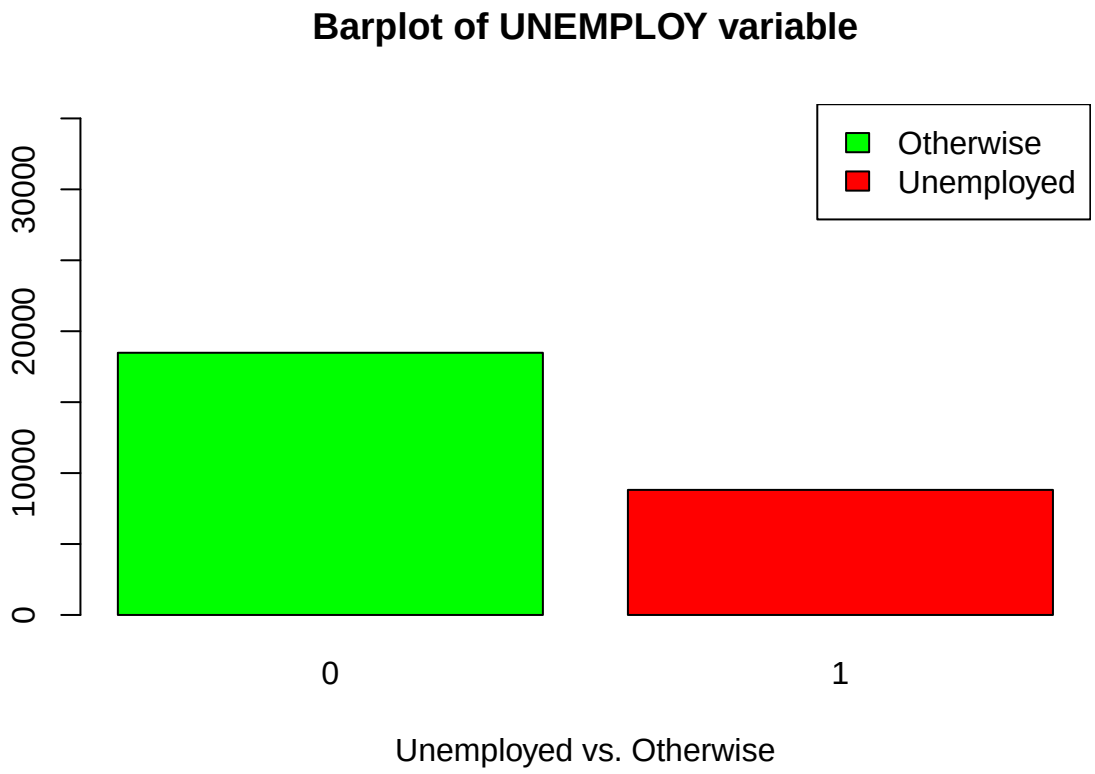
```
counts<-table(Pr3c$BEAMT)
barplot(counts,main="Barplot of BEAMT variable",
        xlab="Civil Servant vs. Otherwise",
        col=c("lightgrey","royalblue1"),
        ylim = c(0,36000))
legend('topright',c("Otherwise","Civil Servant"),fill=c("lightgrey","royalblue1"))
```



The vast majority of individuals are not civil servants.

Barplot of variable UNEMPLOY

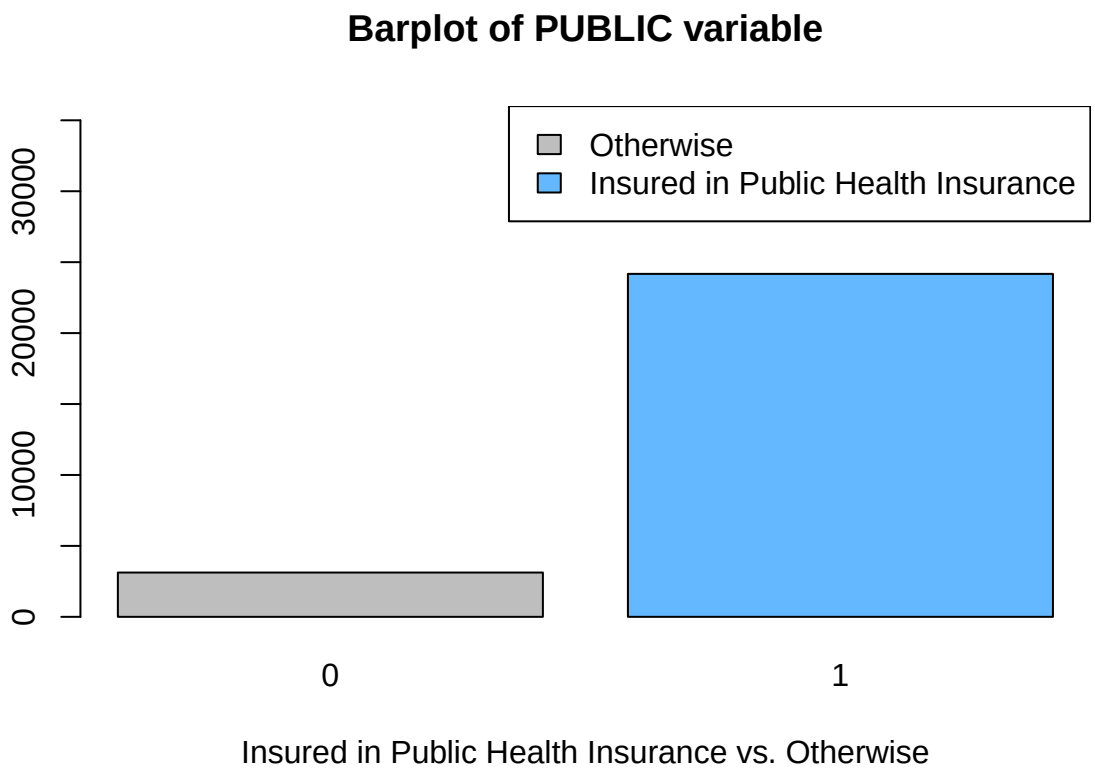
```
counts<-table(Pr3c$UNEMPLOY)
barplot(counts,main="Barplot of UNEMPLOY variable",
        xlab="Unemployed vs. Otherwise",
        col=c("green","red"),
        ylim = c(0,36000))
legend('topright',c("Otherwise","Unemployed"),fill=c("green","red"))
```



About a third of all individuals in the sample are unemployed.

Barplot of variable PUBLIC

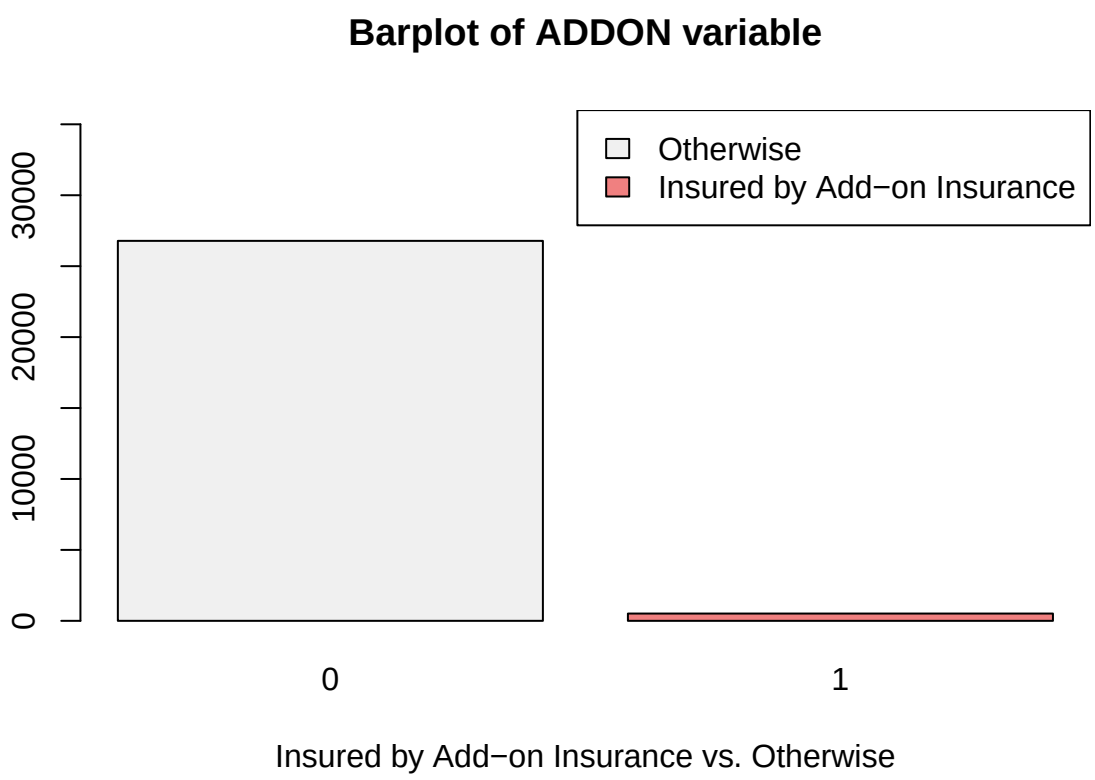
```
counts<-table(Pr3c$PUBLIC)
barplot(counts,main="Barplot of PUBLIC variable",
        xlab="Insured in Public Health Insurance vs. Otherwise",
        col=c("grey","steelblue1"),
        ylim = c(0,36000))
legend('topright',c("Otherwise","Insured in Public Health Insurance"),
       fill=c("grey","steelblue1"))
```



The vast majority of individuals are insured in public health insurance.

Barplot of variable ADDON

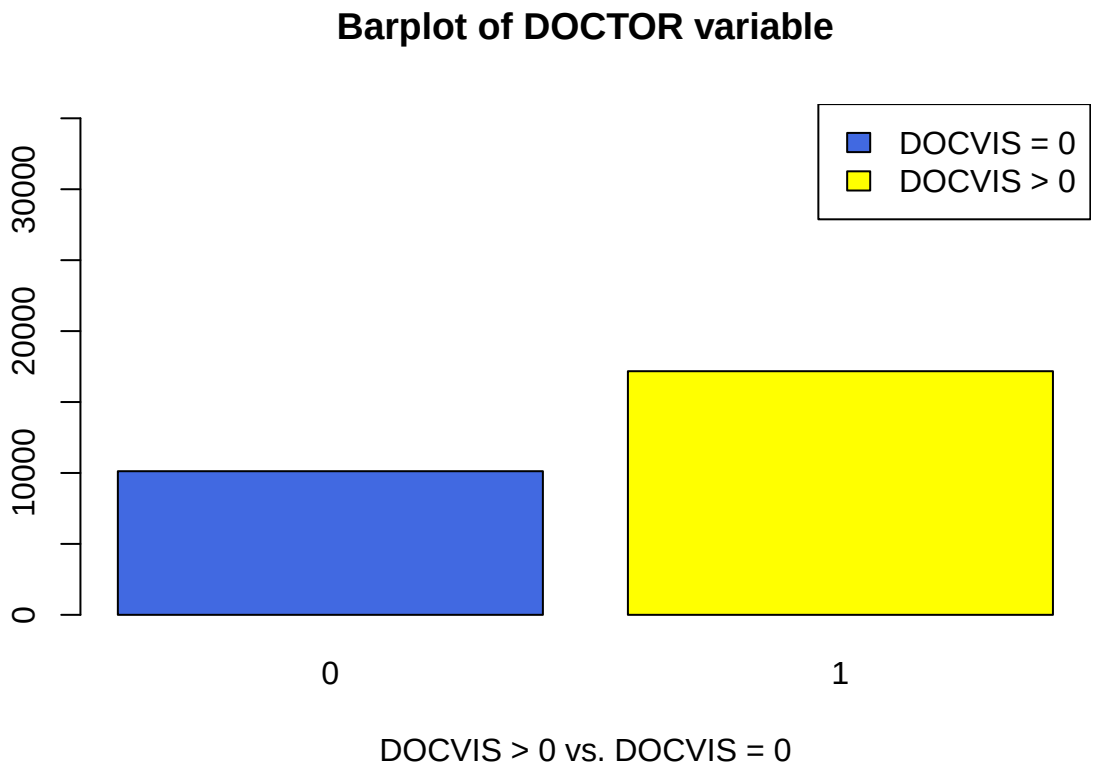
```
counts<-table(Pr3c$ADDON)
barplot(counts,main="Barplot of ADDON variable",
        xlab= "Insured by Add-on Insurance vs. Otherwise",
        col=c("gray94","lightcoral"),
        ylim = c(0,36000))
legend('topright',c("Otherwise","Insured by Add-on Insurance"),
       fill=c("gray94","lightcoral"))
```



The vast majority of individuals were not insured by add-on insurance.

Barplot of variable DOCTOR

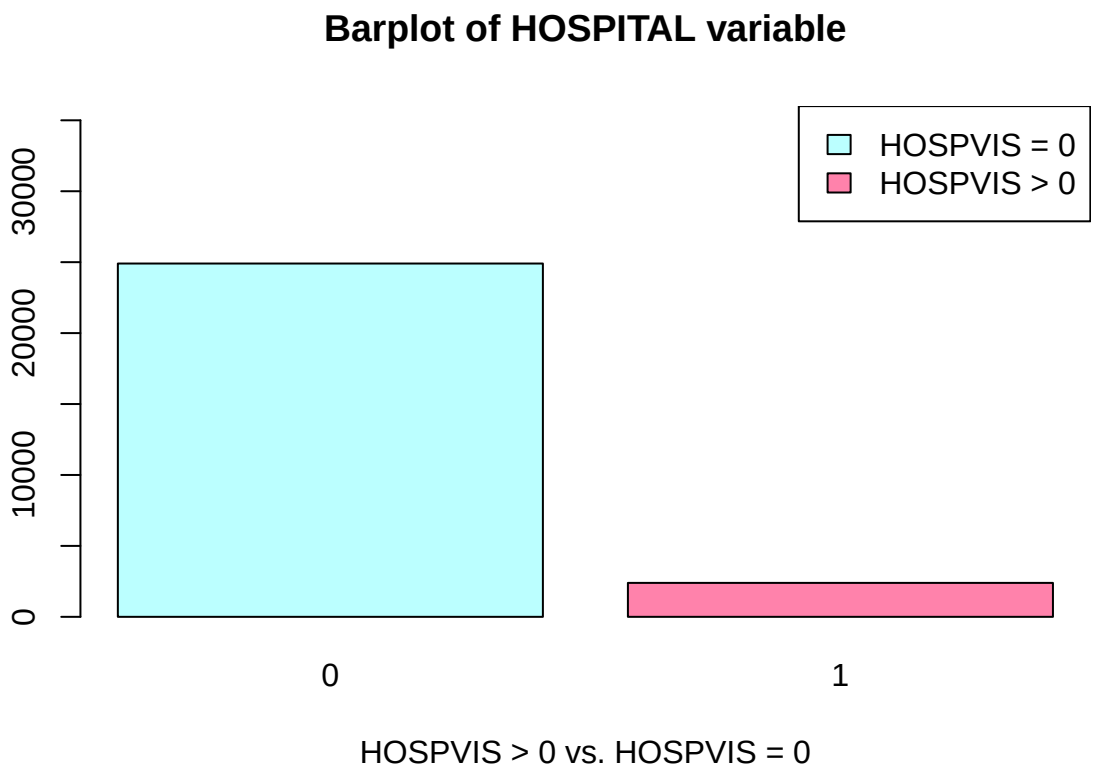
```
counts<-table(Pr3c$DOCTOR)
barplot(counts,main="Barplot of DOCTOR variable",
        xlab="DOCVIS > 0 vs. DOCVIS = 0",
        col=c("royalblue","yellow"),
        ylim = c(0,36000))
legend('topright',c("DOCVIS = 0","DOCVIS > 0"),
       fill=c("royalblue","yellow"))
```



The majority of individuals have visited a doctor in the last three months.

Barplot of variable HOSPITAL

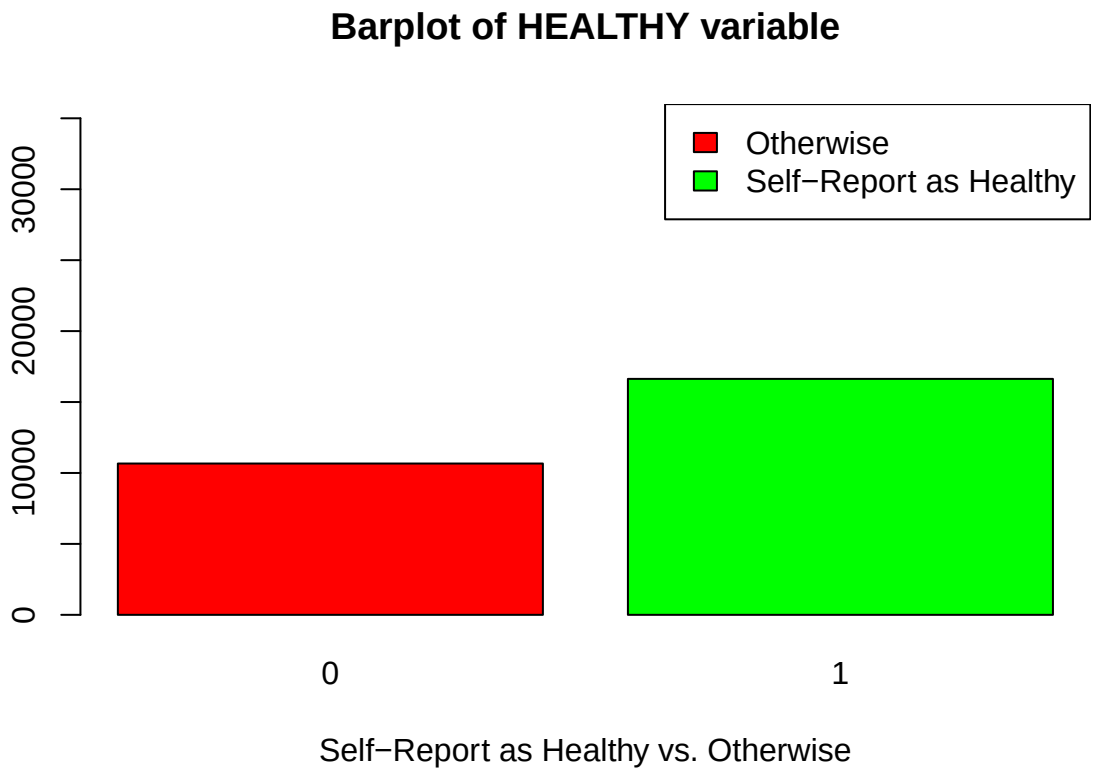
```
counts<-table(Pr3c$HOSPITAL)
barplot(counts,main="Barplot of HOSPITAL variable",
        xlab="HOSPVIS > 0 vs. HOSPVIS = 0",
        col=c("paleturquoise1","palevioletred1"),
        ylim = c(0,36000))
legend('topright',c("HOSPVIS = 0","HOSPVIS > 0"),
       fill=c("paleturquoise1","palevioletred1"))
```



The vast majority of individuals have not gone to a hospital in the last calendar year.

Barplot of variable HEALTHY

```
counts<-table(Pr3c$HEALTHY)
barplot(counts,main="Barplot of HEALTHY variable",
        xlab="Self-Report as Healthy vs. Otherwise",
        col=c("red","green"),
        ylim = c(0,36000))
legend('topright',c("Otherwise","Self-Report as Healthy"),
       fill=c("red","green"))
```

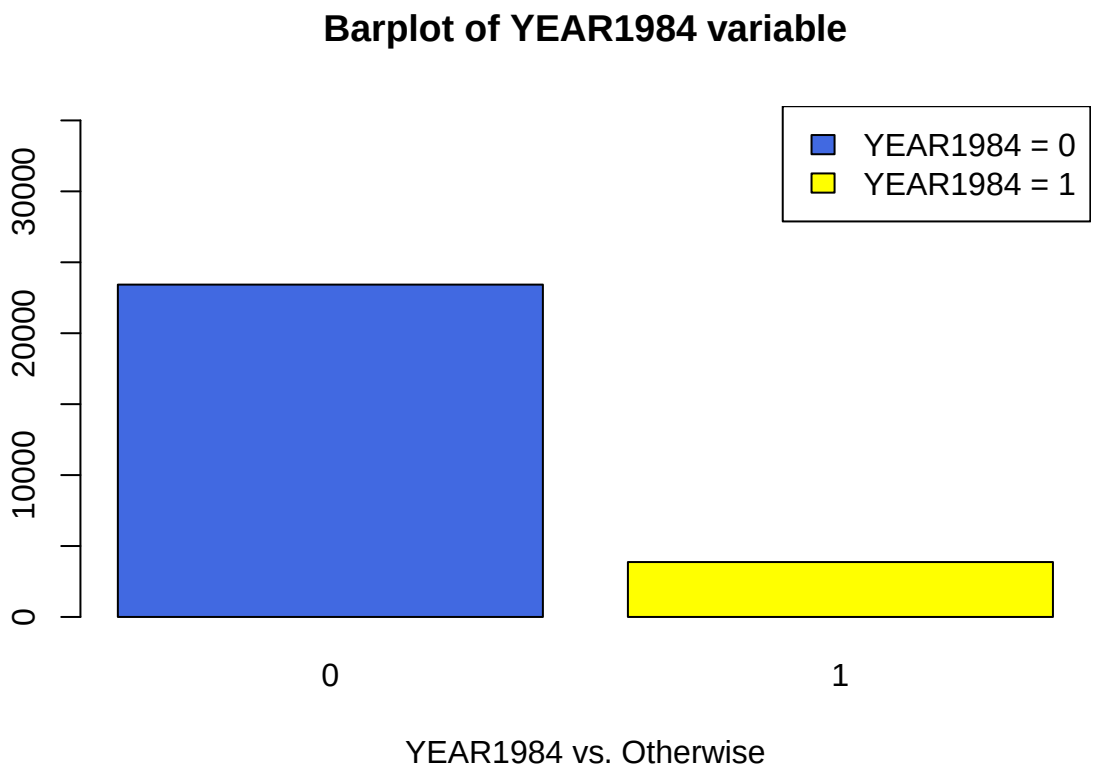


The majority of individuals in the sample self-reported as being healthy.

Note that this means around 35% of all individuals in the sample self-reported as not being healthy, which seems high.

Barplot of variable YEAR1984

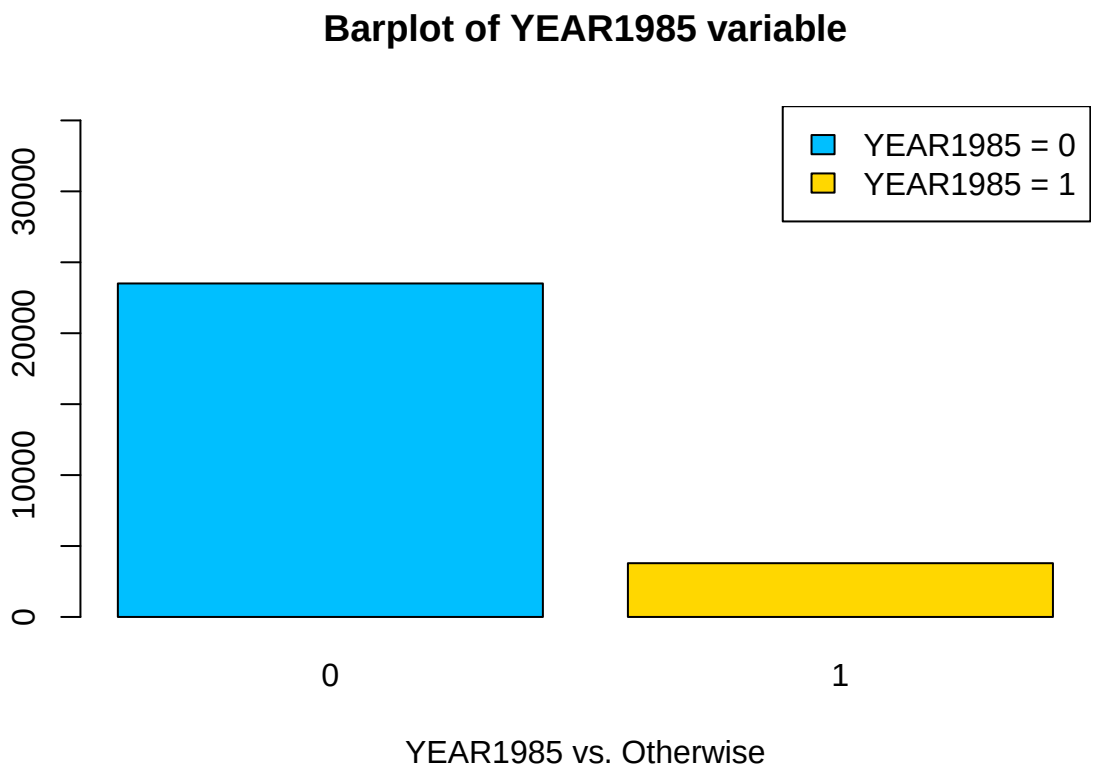
```
counts<-table(Pr3c$YEAR1984)
barplot(counts,main="Barplot of YEAR1984 variable",
        xlab="YEAR1984 vs. Otherwise",
        col=c("royalblue","yellow"),
        ylim = c(0,36000))
legend('topright',c("YEAR1984 = 0","YEAR1984 = 1"),fill=c("royalblue","yellow"))
```



A small proportion of the data is from 1984.

Barplot of variable YEAR1985

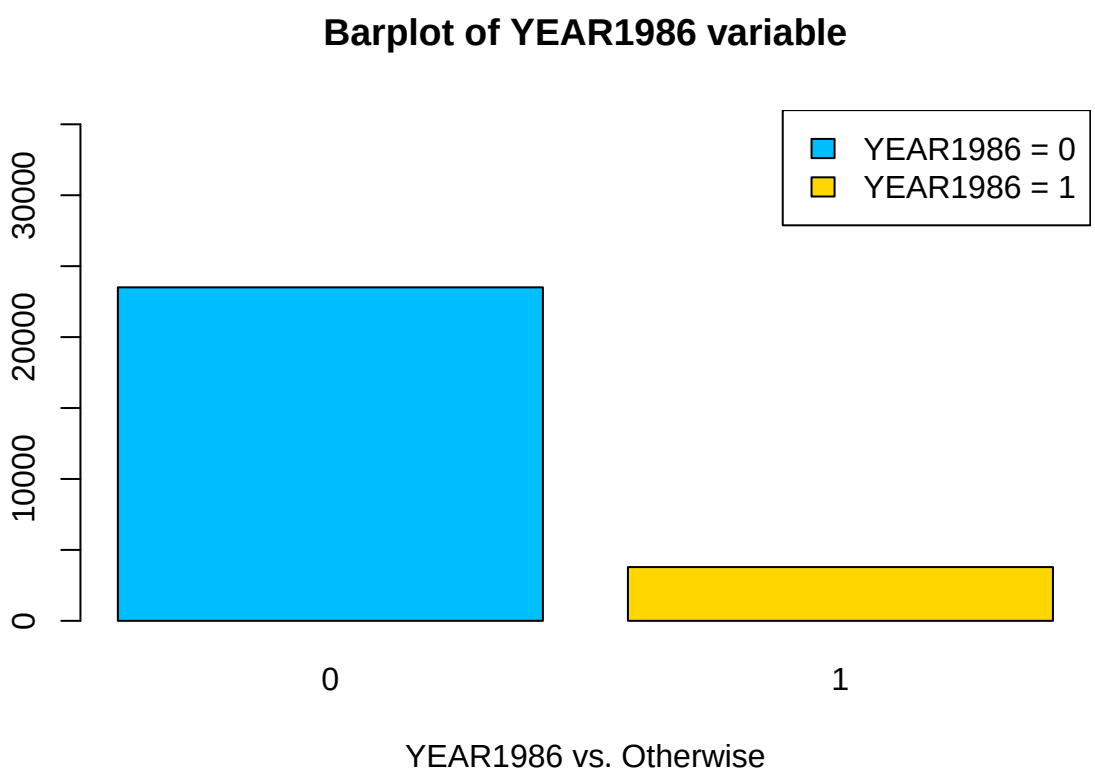
```
counts<-table(Pr3c$YEAR1985)
barplot(counts,main="Barplot of YEAR1985 variable",
        xlab="YEAR1985 vs. Otherwise",col=c("deepskyblue","gold"),
        ylim = c(0,36000))
legend('topright',c("YEAR1985 = 0","YEAR1985 = 1"),
       fill=c("deepskyblue","gold"))
```



A small proportion of the data is from 1985.

Barplot of variable YEAR1986

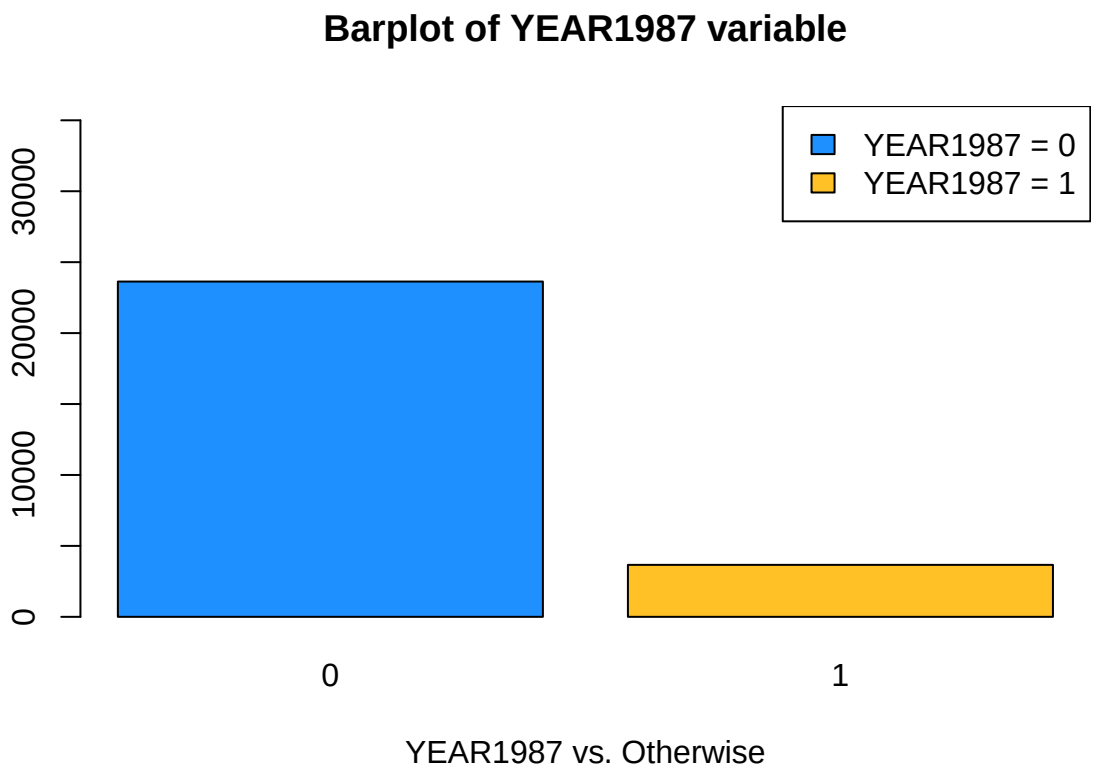
```
counts<-table(Pr3c$YEAR1986)
barplot(counts,main="Barplot of YEAR1986 variable",
        xlab="YEAR1986 vs. Otherwise",
        col=c("deepskyblue","gold"),
        ylim = c(0,36000))
legend('topright',c("YEAR1986 = 0","YEAR1986 = 1"),
       fill=c("deepskyblue","gold"))
```



A small proportion of the data is from 1986, and it is practically the same proportion of the data as is from 1985.

Barplot of variable YEAR1987

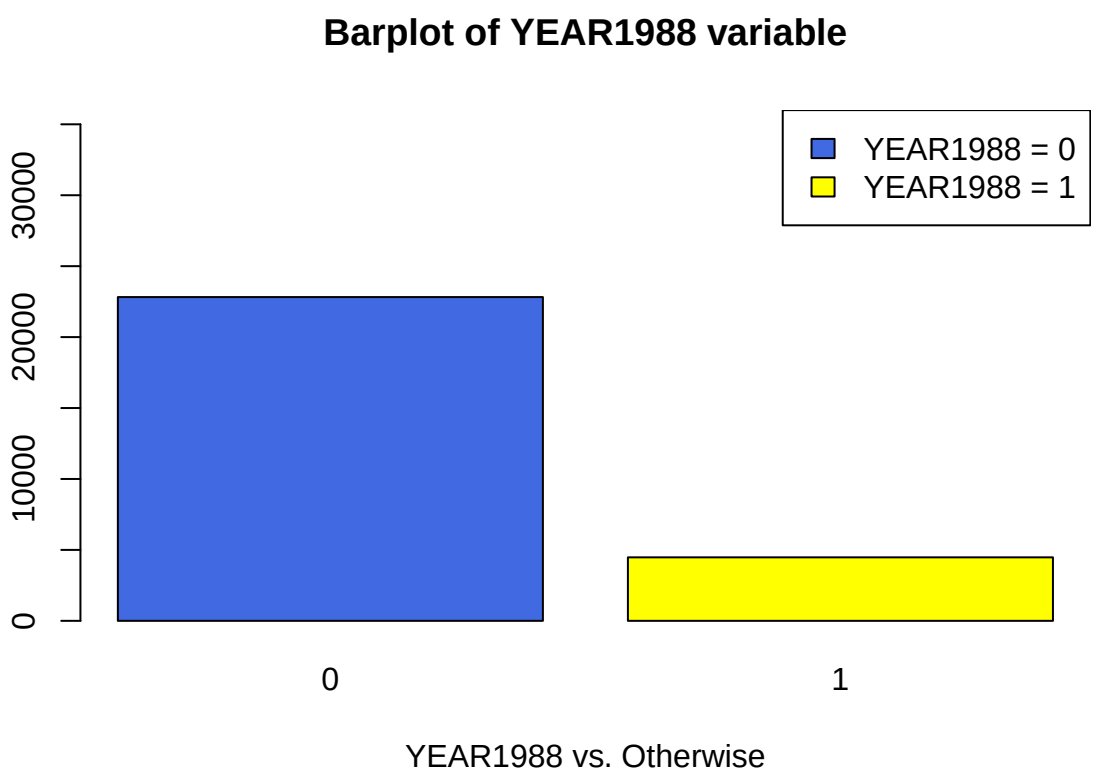
```
counts<-table(Pr3c$YEAR1987)
barplot(counts,main="Barplot of YEAR1987 variable",
        xlab="YEAR1987 vs. Otherwise",
        col=c("dodgerblue","goldenrod1"),
        ylim = c(0,36000))
legend('topright',c("YEAR1987 = 0","YEAR1987 = 1"),
       fill=c("dodgerblue","goldenrod1"))
```



A small proportion of the data is from 1987.

Barplot of variable YEAR1988

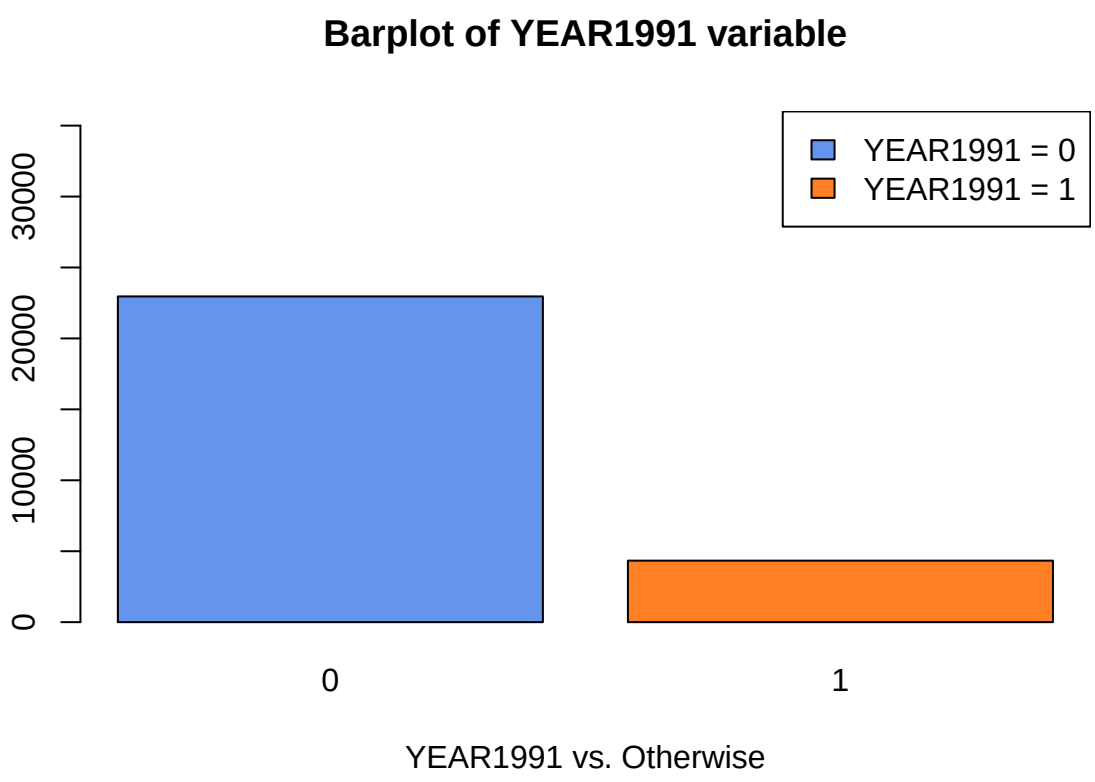
```
counts<-table(Pr3c$YEAR1988)
barplot(counts,main="Barplot of YEAR1988 variable",
        xlab="YEAR1988 vs. Otherwise",
        col=c("royalblue","yellow"),
        ylim = c(0,36000))
legend('topright',c("YEAR1988 = 0","YEAR1988 = 1"),
       fill=c("royalblue","yellow"))
```



A small proportion of the data is from 1988.

Barplot of variable YEAR1991

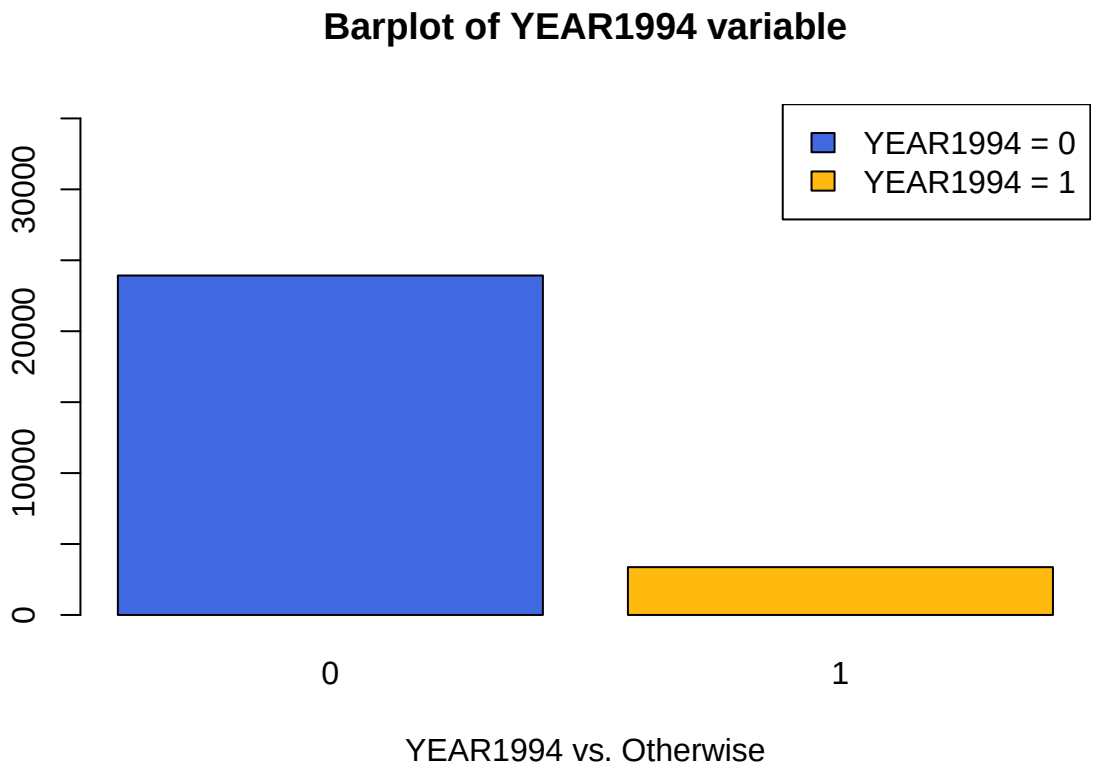
```
counts<-table(Pr3c$YEAR1991)
barplot(counts,main="Barplot of YEAR1991 variable",
        xlab="YEAR1991 vs. Otherwise",
        col=c("cornflowerblue","chocolate1"),
        ylim = c(0,36000))
legend('topright',c("YEAR1991 = 0","YEAR1991 = 1"),
       fill=c("cornflowerblue","chocolate1"))
```



A small proportion of the data is from 1991.

Barplot of variable YEAR1994

```
counts<-table(Pr3c$YEAR1994)
barplot(counts,main="Barplot of YEAR1994 variable",
        xlab="YEAR1994 vs. Otherwise",col=c("royalblue","darkgoldenrod1"),
        ylim = c(0,36000))
legend('topright', c("YEAR1994 = 0","YEAR1994 = 1"),
       fill=c("royalblue","darkgoldenrod1"))
```



A small proportion of the data is from 1994.

Part I (c)

The following are the names of the variables:

```
names(Pr3)
```

```
##  [1] "ID"       "FEMALE"   "YEAR"     "AGE"      "HANDDUM"  "ALC"
##  [7] "FAMHIST"  "HANDPER"  "HHKIDS"   "EDUC"     "MARRIED"  "HAUPTS"
## [13] "REALS"    "FACHHS"   "ABITUR"   "UNIV"     "WORKING"  "BLUEC"
## [19] "WHITEC"   "SELF"     "BEAMT"    "DOCVIS"   "HOSPVIS"  "UNEMPLOY"
## [25] "PUBLIC"   "ADDON"    "NUMOBS"   "HSAT"     "DOCTOR"   "HEALTHY"
## [31] "YEAR1984" "YEAR1985" "YEAR1986" "YEAR1987" "YEAR1988" "YEAR1991"
## [37] "YEAR1994" "LOGINC"   "TI"       "HOSPITAL" "HHNINC"   "NEWHSAT"
## [43] "PRESCRIP"
```

Estimating a linear regression model that includes all the variables:

- Note that we will not be using the variable **HSAT** but only **NEWHSAT** as the variable **HSAT** has 40 coding errors, and the Variable **NEWHSAT** fixes them.
- We also will not use the variable **TI**, as noted perviously, as we are explicitly asked to ignore it in the Data Description.
- Lastly, note that we are told that we can ignore all missing observations.

Code for linear regression:

```
Pr3Ic <-lm(DOCVIS ~ ID + FEMALE + YEAR + AGE + HANDDUM
                  + ALC + FAMHIST + HANDPER + HHKIDS + EDUC
                  + MARRIED + HAUPTS + REALS + FACHHS + ABITUR
                  + UNIV + WORKING + BLUEC + WHITEC + SELF
                  + BEAMT + HOSPVIS + UNEMPLOY + PUBLIC + ADDON
                  + NUMOBS + DOCTOR + HEALTHY + YEAR1984 + YEAR1985
                  + YEAR1986 + YEAR1987 + YEAR1988 + YEAR1991
                  + YEAR1994 + LOGINC + HOSPITAL + HHNINC + NEWHSAT
                  + PRESCRIP, data = Pr3c)
```

Showing results of linear regression model in table:

Note that the results for the residuals, coefficients, etc. are all displayed on the following page.

```
summary(Pr3Ic)
```

```
##
## Call:
## lm(formula = DOCVIS ~ ID + FEMALE + YEAR + AGE + HANDDUM + ALC +
##     FAMHIST + HANDPER + HHKIDS + EDUC + MARRIED + HAUPTS + REALS +
##     FACHHS + ABITUR + UNIV + WORKING + BLUEC + WHITEC + SELF +
##     BEAMT + HOSPVIS + UNEMPLOY + PUBLIC + ADDON + NUMOBS + DOCTOR +
##     HEALTHY + YEAR1984 + YEAR1985 + YEAR1986 + YEAR1987 + YEAR1988 +
##     YEAR1991 + YEAR1994 + LOGINC + HOSPITAL + HHNINC + NEWHSAT +
##     PRESCRIP, data = Pr3c)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -12.661  -2.198  -0.596   0.917 109.769
##
## Coefficients: (2 not defined because of singularities)
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -7.304e+02  7.405e+01  -9.863  < 2e-16 ***
## ID          -9.143e-05  1.470e-05  -6.221 5.00e-10 ***
## FEMALE       4.167e-01  6.843e-02   6.090 1.15e-09 ***
## YEAR         3.681e-01  3.716e-02   9.903  < 2e-16 ***
## AGE          5.836e-03  3.253e-03   1.794 0.072846 .
## HANDDUM      3.546e-01  1.293e-01   2.743 0.006090 **
## ALC          4.159e-03  2.533e-03   1.642 0.100554
## FAMHIST     -4.500e-02  5.848e-02  -0.770 0.441594
## HANDPER      1.621e-02  2.225e-03   7.285 3.31e-13 ***
## HHKIDS      -2.246e-01  7.010e-02  -3.204 0.001359 **
## EDUC         1.801e-01  3.998e-02   4.504 6.70e-06 ***
## MARRIED     -7.743e-02  8.084e-02  -0.958 0.338196
## HAUPTS       6.767e-02  2.112e-01   0.320 0.748646
## REALS       -2.496e-01  2.299e-01  -1.085 0.277735
## FACHHS      -8.039e-01  3.089e-01  -2.602 0.009266 **
## ABITUR      -8.237e-01  3.077e-01  -2.677 0.007441 **
## UNIV        -7.130e-01  2.106e-01  -3.386 0.000711 ***
## WORKING     -9.693e-02  2.539e-01  -0.382 0.702642
## BLUEC       -1.526e-01  1.702e-01  -0.896 0.370005
## WHITEC      -1.895e-01  1.687e-01  -1.124 0.261204
## SELF        -3.029e-01  1.945e-01  -1.557 0.119403
## BEAMT        3.534e-02  2.105e-01   0.168 0.866675
## HOSPVIS      2.531e-01  3.834e-02   6.601 4.16e-11 ***
## UNEMPLOY    -5.921e-02  2.057e-01  -0.288 0.773410
## PUBLIC       2.100e-01  1.189e-01   1.767 0.077265 .
## ADDON       -2.491e-01  2.184e-01  -1.141 0.253978
## NUMOBS      -2.124e-02  1.738e-02  -1.222 0.221633
## DOCTOR       3.904e+00  6.485e-02  60.195  < 2e-16 ***
## HEALTHY      7.116e-01  1.087e-01   6.547 5.96e-11 ***
## YEAR1984     3.370e+00  3.231e-01  10.429  < 2e-16 ***
## YEAR1985     2.857e+00  2.878e-01   9.927  < 2e-16 ***
## YEAR1986     2.824e+00  2.531e-01  11.160  < 2e-16 ***
## YEAR1987     1.862e+00  2.383e-01   7.814 5.75e-15 ***
## YEAR1988     1.247e+00  1.839e-01   6.777 1.25e-11 ***
## YEAR1991            NA         NA      NA       NA
## YEAR1994            NA         NA      NA       NA
## LOGINC       1.254e-03  1.512e-01   0.008 0.993381
## HOSPITAL     1.781e+00  1.214e-01  14.667  < 2e-16 ***
## HHNINC      -7.139e-01  4.051e-01  -1.762 0.078035 .
## NEWHSAT     -6.740e-01  2.391e-02 -28.186  < 2e-16 ***
## PRESCRIP    -3.788e-03  9.296e-03  -0.407 0.683667
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.834 on 27258 degrees of freedom
## Multiple R-squared:  0.2796, Adjusted R-squared:  0.2786
## F-statistic: 278.4 on 38 and 27258 DF,  p-value: < 2.2e-16
```

Discussion: Results of Linear Regression

First, we observe that 2 coefficients are not defined due to singularities. This is because of multicollinearity:

```
summary(lm(YEAR1991 ~ ID + FEMALE + YEAR + AGE + HANDDUM  + ALC + FAMHIST + HANDPER
+ HHKIDS + EDUC + MARRIED + HAUPTS + REALS + FACHHS + ABITUR + UNIV + WORKING + BLUEC
+ WHITEC + SELF + BEAMT + HOSPVIS + UNEMPLOY + PUBLIC + ADDON + NUMOBS + DOCTOR
+ HEALTHY + YEAR1984 + YEAR1985 + YEAR1986 + YEAR1987 + YEAR1988 + YEAR1994 + LOGINC
+ HOSPITAL + HHNINC + NEWHSAT + PRESCRIP, data = Pr3c))$r.sq
```

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

We observe that the R-squared value of a model for **YEAR1991** is equal to 1, confirming multicollinearity.

```
summary(lm(YEAR1994 ~ ID + FEMALE + YEAR + AGE + HANDDUM + ALC + FAMHIST + HANDPER
+ HHKIDS + EDUC + MARRIED + HAUPTS + REALS + FACHHS + ABITUR + UNIV + WORKING + BLUEC
+ WHITEC + SELF + BEAMT + HOSPVIS + UNEMPLOY + PUBLIC + ADDON + NUMOBS + DOCTOR
+ HEALTHY + YEAR1984 + YEAR1985 + YEAR1986 + YEAR1987 + YEAR1988 + YEAR1991 + LOGINC
+ HOSPITAL + HHNINC + NEWHSAT + PRESCRIP, data = Pr3c))$r.sq
```

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

We observe that the R-squared value of a model for **YEAR1994** is also equal to 1, confirming multicollinearity.

We observe that the following explanatory variables are significant at the 0.001 level:

- ID
- FEMALE
- YEAR
  - YEAR1984
  - YEAR1985
  - YEAR1986
  - YEAR1987
  - YEAR1988
- HANDPER
- EDUC
- UNIV
- HOSPVIS
- DOCTOR
- HEALTHY
- HOSPITAL
- NEWHSAT

We also observe that the following explanatory variables are significant at the 0.01 level:

- HANDDUM
- HHKIDS
- FACHHS
- ABITUR

Lastly, we observe that the following explanatory variables are significant at the 0.1 level:

- AGE
- PUBLIC
- HHNINC

Note that the intercept is also significant at the 0.001 level.

The model's R-squared value is 0.2796 and Adjusted R-squared value is 0.2786. This means that the model is not particularly effective at estimating the number of a patient's doctor visits over a 3 month period.

Comments about specific explanatory variables and their relative impacts on the model's estimates:

Miscellaneous Observation

We note that we find the statistical significance of **ID** very odd, given that identification numbers are usually completely random and since we saw in Part (a) that **ID** follows a relatively uniform distribution. However, we do note that the coefficient of the **ID** variable is negligible compared to the other explanatory variables.

Gender

Considering the **FEMALE** variable, we observe that an individual being female increases the extimated number of doctor visits over a 3 month period by 0.4167 visits.

Time/Year

Considering time, we observe that the coefficient of the **YEAR** variable is 0.3681, and the respective coefficients of the categorical year variables are:

- **YEAR1984**: 3.370
- **YEAR1985**: 2.857
- **YEAR1986**: 2.824
- **YEAR1987**: 1.862
- **YEAR1988**: 1.247

This means that as the time increases from the year 1984 to the year 1994, the quantitative **YEAR** variable results in an individual going to the doctor an estimated 0.3681(YEAR) additional times every 3 months. Just for the year 1984, this leads to an estimated 730.3104 additional doctor visits - *which does not make sense unless one takes into account the* ***Intercept****, which has a value of -730.4.*

Also, on the particular years that we have a categorical year variable for (1984, 1985, 1986, 1987, 1988), every 3 months the individual is estimated to go to the doctor an additional amount corresponding to the coefficient of each of the categorical year variables, *on top of the estimated effect* of the **YEAR** quantitative variable for that year (which, as previously noted, should be considered in the context of the intercept).

Individual's Age

Considering **AGE**, an individual goes to the doctor an estimated additional 0.005836 times every 3 months for each year of their age.

Disability

Considering the **HANDDUM** variable, we observe that being handicapped increases an individual's estimated number of doctor visits in a 3 month period by 0.3546 visits. Considering the **HANDPER** variable, we observe that the degree of handicap a person has increases their estimated number of doctor visits in a 3 month period by 0.01621 *for each additional percent* of their degree of handicap. In particular, for someone who is 100 percent handicapped, the **HANDPER** variable's coefficient leads to that person going to the doctor an estimated 1.621 additional times in a 3 month period.

Children/Family

Considering **HHKIDS**, having children under 16 years of age in the household leads to an estimated *decrease* in doctor visits of 0.2246 in a 3 month period.

Education

Considering **EDUC**, an individual is estimated to go to the doctor 0.1801 additional times every 3 months for each year of schooling they have.

Considering **FACHHS**, **ABITUR**, **UNIV**, an individual having a Polytechnical degree, an Abitur, or a university degree as their highest level of schooling results in an estimated *decrease* in doctor visits over a 3 month period by 0.8039 visits, 0.8237 visits, and 0.7130 visits for each of the respective highest levels of schooling.

Health Information from Prior Health Behavior

Considering **HOSPVIS**, an individual's estimated number of hospital visits in the last calendar year increases the estimated number of times an individual goes to the doctor in a 3 month period by 0.2531 visits for each of their hospital visits in the last calendar year. This is in addition to the estimated impact of **HOSPITAL**, which increases an individual's estimated number of doctor visits in a 3 month period by 1.781 if an individual has gone to the hospital at all in the past calendar year. Considering **DOCTOR**, an individual having gone to the doctor in the past 3 months increases their estimated number of doctor visits in a 3 month period by 3.904 visits.

Also, considering **PUBLIC**, being insured in public health insurance has an estimated increase on doctor visits in a 3 month period by 0.2100 visits.

Self-Reported Health Information

Considering **NEWHSAT**, an individual is estimated to visit the doctor 0.6740 *less times* in a 3 month period for each additional point of health satisfaction (above zero) that they report. Also, someone who has self-reported as **HEALTHY** is estimated to visit the doctor 0.7116 *less times* in a 3 month period than someone who has not.

II. Model Building

Part II (a)

Proposed nested regression model:

Based on our findings in Part (I), we propose a model that only includes the explanatory variables that had a p-value less than 0.05. So we remove from our model any variables that have a greater than 5% probability of not having any predictive power whatsoever (within our model) for estimating the number of doctor visits a patient has over a 3 month period.

We therefore include the following variables in our proposed model:

- ID
- FEMALE
- YEAR
- HANDDUM
- HANDPER
- HHKIDS
- EDUC
- FACHHS
- ABITUR
- UNIV
- HOSPVIS
- DOCTOR
- HEALTHY
- YEAR1984
- YEAR1985
- YEAR1986
- YEAR1987
- YEAR1988
- HOSPITAL
- NEWHSAT

Code for linear regression to create proposed nested model:

```
Pr3IIa <- lm(DOCVIS ~ ID + FEMALE + YEAR + HANDDUM + HANDPER + HHKIDS + EDUC + FACHHS
                  + ABITUR + UNIV + HOSPVIS + DOCTOR + HEALTHY + YEAR1984 + YEAR1985
                  + YEAR1986 + YEAR1987 + YEAR1988 + HOSPITAL + NEWHSAT, data = Pr3c)
```

Showing results of nested model in table:

```
summary(Pr3IIa)
```

```
##
## Call:
## lm(formula = DOCVIS ~ ID + FEMALE + YEAR + HANDDUM + HANDPER +
##     HHKIDS + EDUC + FACHHS + ABITUR + UNIV + HOSPVIS + DOCTOR +
##     HEALTHY + YEAR1984 + YEAR1985 + YEAR1986 + YEAR1987 + YEAR1988 +
##     HOSPITAL + NEWHSAT, data = Pr3c)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -12.864  -2.191  -0.604   0.911 109.924
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -7.164e+02  7.383e+01  -9.704  < 2e-16 ***
## ID          -9.162e-05  1.463e-05  -6.264 3.82e-10 ***
## FEMALE       4.525e-01  6.098e-02   7.419 1.21e-13 ***
## YEAR         3.616e-01  3.705e-02   9.759  < 2e-16 ***
## HANDDUM      3.637e-01  1.292e-01   2.815  0.00489 **
## HANDPER      1.812e-02  2.175e-03   8.331  < 2e-16 ***
## HHKIDS      -2.871e-01  6.100e-02  -4.707 2.53e-06 ***
## EDUC         5.700e-02  3.133e-02   1.820  0.06883 .
## FACHHS      -4.189e-01  1.873e-01  -2.237  0.02533 *
## ABITUR      -3.405e-01  1.780e-01  -1.912  0.05585 .
## UNIV        -4.177e-01  1.927e-01  -2.168  0.03018 *
## HOSPVIS      2.516e-01  3.836e-02   6.560 5.48e-11 ***
## DOCTOR       3.911e+00  6.471e-02  60.441  < 2e-16 ***
## HEALTHY      6.988e-01  1.086e-01   6.433 1.27e-10 ***
## YEAR1984     3.433e+00  3.225e-01  10.645  < 2e-16 ***
## YEAR1985     2.911e+00  2.875e-01  10.126  < 2e-16 ***
## YEAR1986     2.868e+00  2.528e-01  11.343  < 2e-16 ***
## YEAR1987     1.873e+00  2.380e-01   7.870 3.68e-15 ***
## YEAR1988     1.272e+00  1.838e-01   6.920 4.61e-12 ***
## HOSPITAL     1.785e+00  1.215e-01  14.695  < 2e-16 ***
## NEWHSAT     -6.809e-01  2.389e-02 -28.502  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.838 on 27276 degrees of freedom
## Multiple R-squared:  0.2779, Adjusted R-squared:  0.2774
## F-statistic:   525 on 20 and 27276 DF,  p-value: < 2.2e-16
```

We note that the R-squared value is 0.2770 and Adjusted R-squared value is 0.2774.

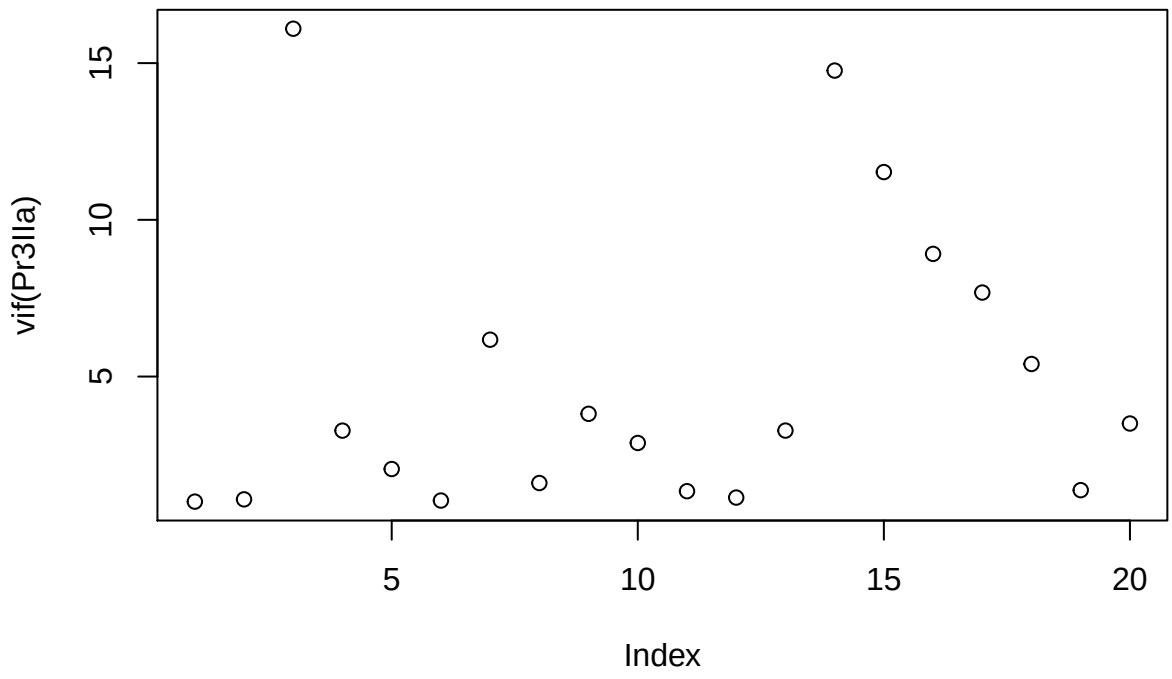
We observe that within our new nested model, the variables relating to education, such as years of education or highest educational degree, have the highest p-values. It is worth noting, however, that their p-values are relatively low, with **EDUC** having a p-value of 0.06883 and **ABITUR** having a p-value of 0.05585. The other two variables relating to education have p-values less than 0.05. Nonetheless, all four variables experienced an increase in their p-values after removing the variables with p-values greater than 0.05 from the original model.

As in our orginal model, the nested model also has a very high value for its intercept (-716.4).

Part II (b)

Testing for multicollinearity via VIF:

```
plot(vif(Pr3IIa))
```



We observe that we have some incredibly high VIFs. This suggests that there is multicollinearity between the explanatory variables in our nested model.

Analysis of VIFs:

```
vif(Pr3IIa)
```

```
##       ID   FEMALE     YEAR  HANDDUM  HANDPER   HHKIDS     EDUC   FACHHS
##   1.0106   1.0825  16.0970   3.2754   2.0495   1.0441   6.1760   1.6033
##   ABITUR     UNIV  HOSPVIS   DOCTOR  HEALTHY YEAR1984 YEAR1985 YEAR1986
##   3.8116   2.8815   1.3432   1.1396   3.2763  14.7620  11.5240   8.9133
## YEAR1987 YEAR1988 HOSPITAL  NEWHSAT
##   7.6801   5.4012   1.3761   3.5018
```

We note that the variables related to time (**YEAR**, **YEAR1984**, **YEAR1985**, **YEAR1986**, **YEAR1987**, **YEAR1988**) have particularly high VIFs. Multicollinearity between the quantitative and categorical time variables does make sense.

We also note that **EDUC** has a VIF of 6.1760. It is possible that the variables **FACHHS**, **ABITUR**, and **UNIV** predict doctor visits in a 3 month period just as well, if not better than, **EDUC**. This may indicate that the impact of education on doctor visits in a 3 month period may be better predicted by a stepwise function through the categorical variables than a linear function through the quantitative variable **EDUC**.

Conducting auxiliary regressions (for variables 1 - 10):

```
# Auxiliary Regression 1: **ID**
Pr3IIb.IDar <- lm(ID ~ + FEMALE + YEAR + HANDDUM + HANDPER + HHKIDS + EDUC + FACHHS
                  + ABITUR + UNIV + HOSPVIS + DOCTOR + HEALTHY + YEAR1984 + YEAR1985
                  + YEAR1986 + YEAR1987 + YEAR1988 + HOSPITAL + NEWHSAT, data = Pr3c)
# Auxiliary Regression 2: **FEMALE**
Pr3IIb.FEMALEar <- lm(FEMALE ~ ID + YEAR + HANDDUM + HANDPER + HHKIDS + EDUC + FACHHS
                  + ABITUR + UNIV + HOSPVIS + DOCTOR + HEALTHY + YEAR1984 + YEAR1985
                  + YEAR1986 + YEAR1987 + YEAR1988 + HOSPITAL + NEWHSAT, data = Pr3c)
# Auxiliary Regression 3: **YEAR**
Pr3IIb.YEARar <- lm(YEAR ~ ID + FEMALE + HANDDUM + HANDPER + HHKIDS + EDUC + FACHHS
                  + ABITUR + UNIV + HOSPVIS + DOCTOR + HEALTHY + YEAR1984 + YEAR1985
                  + YEAR1986 + YEAR1987 + YEAR1988 + HOSPITAL + NEWHSAT, data = Pr3c)
# Auxiliary Regression 4: **HANDDUM**
Pr3IIb.HANDDUMar <- lm(HANDDUM ~ ID + FEMALE + YEAR + HANDPER + HHKIDS + EDUC + FACHHS
                  + ABITUR + UNIV + HOSPVIS + DOCTOR + HEALTHY + YEAR1984 + YEAR1985
                  + YEAR1986 + YEAR1987 + YEAR1988 + HOSPITAL + NEWHSAT, data = Pr3c)
# Auxiliary Regression 5: **HANDPER**
Pr3IIb.HANDPERar <- lm(HANDPER ~ ID + FEMALE + YEAR + HANDDUM + HHKIDS + EDUC + FACHHS
                  + ABITUR + UNIV + HOSPVIS + DOCTOR + HEALTHY + YEAR1984 + YEAR1985
                  + YEAR1986 + YEAR1987 + YEAR1988 + HOSPITAL + NEWHSAT, data = Pr3c)
# Auxiliary Regression 6: **HHKIDS**
Pr3IIb.HHKIDSar <- lm(HHKIDS ~ ID + FEMALE + YEAR + HANDDUM + HANDPER + EDUC + FACHHS
                  + ABITUR + UNIV + HOSPVIS + DOCTOR + HEALTHY + YEAR1984 + YEAR1985
                  + YEAR1986 + YEAR1987 + YEAR1988 + HOSPITAL + NEWHSAT, data = Pr3c)
# Auxiliary Regression 7: **EDUC**
Pr3IIb.EDUCar <- lm(EDUC ~ ID + FEMALE + YEAR + HANDDUM + HANDPER + HHKIDS + FACHHS
                  + ABITUR + UNIV + HOSPVIS + DOCTOR + HEALTHY + YEAR1984 + YEAR1985
                  + YEAR1986 + YEAR1987 + YEAR1988 + HOSPITAL + NEWHSAT, data = Pr3c)
# Auxiliary Regression 8: **FACHHS**
Pr3IIb.FACHHSar <- lm(FACHHS ~ ID + FEMALE + YEAR + HANDDUM + HANDPER + HHKIDS + EDUC
                  + ABITUR + UNIV + HOSPVIS + DOCTOR + HEALTHY + YEAR1984 + YEAR1985
                  + YEAR1986 + YEAR1987 + YEAR1988 + HOSPITAL + NEWHSAT, data = Pr3c)
# Auxiliary Regression 9: **ABITUR**
Pr3IIb.ABITURar <- lm(ABITUR ~ ID + FEMALE + YEAR + HANDDUM + HANDPER + HHKIDS + EDUC
                  + FACHHS
                  + UNIV + HOSPVIS + DOCTOR + HEALTHY + YEAR1984 + YEAR1985
                  + YEAR1986 + YEAR1987 + YEAR1988 + HOSPITAL + NEWHSAT, data = Pr3c)
# Auxiliary Regression 10: **UNIV**
Pr3IIb.UNIVar <- lm(UNIV ~ ID + FEMALE + YEAR + HANDDUM + HANDPER + HHKIDS + EDUC
                  + FACHHS
                  + ABITUR + HOSPVIS + DOCTOR + HEALTHY + YEAR1984 + YEAR1985
                  + YEAR1986 + YEAR1987 + YEAR1988 + HOSPITAL + NEWHSAT, data = Pr3c)
```

Conducting auxiliary regressions (for variables 11 - 20):

```
# Auxiliary Regression 11: **HOSPVIS**
Pr3IIb.HOSPVISar <- lm(HOSPVIS ~ ID + FEMALE + YEAR + HANDDUM + HANDPER + HHKIDS + EDUC
                  + FACHHS
                  + ABITUR + UNIV + DOCTOR + HEALTHY + YEAR1984 + YEAR1985
                  + YEAR1986 + YEAR1987 + YEAR1988 + HOSPITAL + NEWHSAT, data = Pr3c)
# Auxiliary Regression 12: **DOCTOR**
Pr3IIb.DOCTORar <- lm(DOCTOR ~ ID + FEMALE + YEAR + HANDDUM + HANDPER + HHKIDS + EDUC
                  + FACHHS
                  + ABITUR + UNIV + HOSPVIS + HEALTHY + YEAR1984 + YEAR1985
                  + YEAR1986 + YEAR1987 + YEAR1988 + HOSPITAL + NEWHSAT, data = Pr3c)
# Auxiliary Regression 13: **HEALTHY**
Pr3IIb.HEALTHYar <- lm(HEALTHY ~ ID + FEMALE + YEAR + HANDDUM + HANDPER + HHKIDS + EDUC
                  + FACHHS
                  + ABITUR + UNIV + HOSPVIS + DOCTOR + YEAR1984 + YEAR1985
                  + YEAR1986 + YEAR1987 + YEAR1988 + HOSPITAL + NEWHSAT, data = Pr3c)
# Auxiliary Regression 14: **YEAR1984**
Pr3IIb.YEAR1984ar <- lm(YEAR1984 ~ ID + FEMALE + YEAR + HANDDUM + HANDPER + HHKIDS
                  + EDUC + FACHHS
                  + ABITUR + UNIV + HOSPVIS + DOCTOR + HEALTHY + YEAR1985
                  + YEAR1986 + YEAR1987 + YEAR1988 + HOSPITAL + NEWHSAT, data = Pr3c)
# Auxiliary Regression 15: **YEAR1985**
Pr3IIb.YEAR1985ar <- lm(YEAR1985 ~ ID + FEMALE + YEAR + HANDDUM + HANDPER + HHKIDS
                  + EDUC + FACHHS
                  + ABITUR + UNIV + HOSPVIS + DOCTOR + HEALTHY + YEAR1984
                  + YEAR1986 + YEAR1987 + YEAR1988 + HOSPITAL + NEWHSAT, data = Pr3c)
# Auxiliary Regression 16: **YEAR1986**
Pr3IIb.YEAR1986ar <- lm(YEAR1986 ~ ID + FEMALE + YEAR + HANDDUM + HANDPER + HHKIDS
                  + EDUC + FACHHS
                  + ABITUR + UNIV + HOSPVIS + DOCTOR + HEALTHY + YEAR1984 + YEAR1985
                  + YEAR1987 + YEAR1988 + HOSPITAL + NEWHSAT, data = Pr3c)
# Auxiliary Regression 17: **YEAR1987**
Pr3IIb.YEAR1987ar <- lm(YEAR1987 ~ ID + FEMALE + YEAR + HANDDUM + HANDPER + HHKIDS
                  + EDUC + FACHHS
                  + ABITUR + UNIV + HOSPVIS + DOCTOR + HEALTHY + YEAR1984 + YEAR1985
                  + YEAR1986 + YEAR1988 + HOSPITAL + NEWHSAT, data = Pr3c)
# Auxiliary Regression 18: **YEAR1988**
Pr3IIb.YEAR1988ar <- lm(YEAR1988 ~ ID + FEMALE + YEAR + HANDDUM + HANDPER + HHKIDS
                  + EDUC + FACHHS
                  + ABITUR + UNIV + HOSPVIS + DOCTOR + HEALTHY + YEAR1984 + YEAR1985
                  + YEAR1986 + YEAR1987 + HOSPITAL + NEWHSAT, data = Pr3c)
# Auxiliary Regression 19: **HOSPITAL**
Pr3IIb.HOSPITALar <- lm(HOSPITAL ~ ID + FEMALE + YEAR + HANDDUM + HANDPER + HHKIDS
                  + EDUC + FACHHS
                  + ABITUR + UNIV + HOSPVIS + DOCTOR + HEALTHY + YEAR1984 + YEAR1985
                  + YEAR1986 + YEAR1987 + YEAR1988 + NEWHSAT, data = Pr3c)
# Auxiliary Regression 20: **NEWHSAT**
Pr3IIb.NEWHSATar <- lm(NEWHSAT ~ ID + FEMALE + YEAR + HANDDUM + HANDPER + HHKIDS
                  + EDUC + FACHHS
                  + ABITUR + UNIV + HOSPVIS + DOCTOR + HEALTHY + YEAR1984 + YEAR1985
                  + YEAR1986 + YEAR1987 + YEAR1988 + HOSPITAL, data = Pr3c)
```

R-squared values of auxiliary regressions on the following page.

Summary of Auxiliary Regression Results

```
Pr3IIb.a <- rownames(summary(Pr3IIa)$coefficients)[2:21]
Pr3IIb.b <- c(summary(Pr3IIb.IDar)$r.squared, summary(Pr3IIb.FEMALEar)$r.squared,
summary(Pr3IIb.YEARar)$r.squared, summary(Pr3IIb.HANDDUMar)$r.squared,
summary(Pr3IIb.HANDPERar)$r.squared, summary(Pr3IIb.HHKIDSar)$r.squared,
summary(Pr3IIb.EDUCar)$r.squared, summary(Pr3IIb.FACHHSar)$r.squared,
summary(Pr3IIb.ABITURar)$r.squared, summary(Pr3IIb.UNIVar)$r.squared,
summary(Pr3IIb.HOSPVISar)$r.squared, summary(Pr3IIb.DOCTORar)$r.squared,
summary(Pr3IIb.HEALTHYar)$r.squared, summary(Pr3IIb.YEAR1984ar)$r.squared,
summary(Pr3IIb.YEAR1985ar)$r.squared, summary(Pr3IIb.YEAR1986ar)$r.squared,
summary(Pr3IIb.YEAR1987ar)$r.squared, summary(Pr3IIb.YEAR1988ar)$r.squared,
summary(Pr3IIb.HOSPITALar)$r.squared, summary(Pr3IIb.NEWHSATar)$r.squared)
```

Below are the R-squared values that resulted from each of the auxiliary regressions:

```
Pr3IIbAuxiliaryRegressionResults <- data.frame(Pr3IIb.a, Pr3IIb.b)
colnames(Pr3IIbAuxiliaryRegressionResults) <- c("Variable","R-squared")
format(Pr3IIbAuxiliaryRegressionResults, justify = "left")
```

```
##    Variable  R-squared
## 1  ID       0.01050385
## 2  FEMALE   0.07623384
## 3  YEAR     0.93787520
## 4  HANDDUM  0.69469170
## 5  HANDPER  0.51206932
## 6  HHKIDS   0.04225866
## 7  EDUC     0.83808379
## 8  FACHHS   0.37628558
## 9  ABITUR   0.73764197
## 10 UNIV     0.65296050
## 11 HOSPVIS  0.25549290
## 12 DOCTOR   0.12248469
## 13 HEALTHY  0.69478150
## 14 YEAR1984 0.93225689
## 15 YEAR1985 0.91322587
## 16 YEAR1986 0.88780856
## 17 YEAR1987 0.86979351
## 18 YEAR1988 0.81485763
## 19 HOSPITAL 0.27329019
## 20 NEWHSAT  0.71443566
```

We recall that the R-squared value of the nested model is 0.2779, and so any of the auxiliary regression R-squared values being above 0.2779 means that multicollinearity is a concern (due to Klein's rule).

Analysis of Nested Model p-values, VIF, and Auxiliary Regression Results

Time Variables

We note that the **YEAR** variable regression resulted in an R-squared value of 0.93787520, and that all the time categorical variable (**YEAR1984**, **YEAR1985**, **YEAR1986**, **YEAR1987**, **YEAR1988**) regressions have R-squared values above 0.8, with two of them having R-squared values above 0.9. Since the time categorical variables can likely be predicted by the time quantitative variable **YEAR**, we will keep the variable **YEAR** in our model and discard the time categorical variables.

Education Variables

First, recall that in our nested model, **EDUC** had a p-value of 0.06883 and **ABITUR** had a p-value of 0.05585. So, without even looking at the R-squared values of the auxiliary regressions, we know that these are variables we may want to discard.

We then observe that the regression for **EDUC** has a high R-squared value of 0.83808379, but that the regressions of **FACHHS**, **ABITUR**, and **UNIV** have R-squared values of 0.37628558, 0.73764197, and 0.65296050, respectively. Basically, the other variables are not good predictors of **FACHHS* and possibly** UNIV** but are good predictors of **EDUC**. We note that **FACHHS** has a VIF value of 1.6033 and **UNIV** has a VIF value of 2.8815, whereas **EDUC** has a VIF value of 6.1760 and **ABITUR** has a VIF value of 3.8116.

Based on the p-values, the VIF values, and the R-squared values of the auxiliary regressions, we choose to discard **EDUC** and **ABITUR** from our model, while retaining **FACHHS** and **UNIV**.

Health Information Variables

We recall that in the nested model, **HANDDUM** had a p-value of 0.00489, which is the highest p-value outside of the education variable p-values mentioned above.

We note that the **NEWHSAT** auxiliary regression resulted in an R-squared value of 0.71443566. We also note that the **HANDDUM**, **HANDPER**, and **HEALTHY** auxiliary regressions resulted in R-squared values of 0.69469170, 0.51206932, and 0.69478150, respectively. So **HANDDUM** is better predicted by the other variables than **HANDPER** is. Also, **HANDDUM** has a VIF value of 3.2754, whereas **HANDPER** has a VIF value of 2.0495. Looking at health specifically, we hypothesize that **HEALTHY** is can be predicted by **NEWHSAT**, in addition to variables like **HOSPVIS**, **HOSPITAL**, **DOCTOR**, and possibly even **HANDPER**. Also, both **HEALTHY** and **NEWHSAT** have R-squared values close to one another and are both self-reported measures of health or satisfaction with health.

Based on our analysis, we choose to discard the variables **HANDDUM** and **HEALTHY** from our model.

We are therefore left with the following variables for our model M_0:

- ID
- FEMALE
- YEAR
- HANDPER
- HHKIDS
- FACHHS
- UNIV
- HOSPVIS
- DOCTOR
- HOSPITAL
- NEWHSAT

Part II (c)

Analysis of M_0:

```
M0 <- lm(DOCVIS ~ ID + FEMALE + YEAR + HANDPER + HHKIDS + FACHHS
                  + UNIV + HOSPVIS + DOCTOR + HOSPITAL + NEWHSAT, data = Pr3c)
summary(M0)
```

```
##
## Call:
## lm(formula = DOCVIS ~ ID + FEMALE + YEAR + HANDPER + HHKIDS +
##     FACHHS + UNIV + HOSPVIS + DOCTOR + HOSPITAL + NEWHSAT, data = Pr3c)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -12.194  -2.223  -0.548   0.904 110.410
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.648e+01  1.854e+01   3.047  0.00231 **
## ID          -8.857e-05  1.467e-05  -6.037 1.59e-09 ***
## FEMALE       4.190e-01  6.024e-02   6.956 3.59e-12 ***
## YEAR        -2.620e-02  9.326e-03  -2.809  0.00497 **
## HANDPER      2.256e-02  1.630e-03  13.837  < 2e-16 ***
## HHKIDS      -2.464e-01  6.092e-02  -4.044 5.26e-05 ***
## FACHHS      -1.876e-01  1.491e-01  -1.258  0.20827
## UNIV        -3.143e-01  1.150e-01  -2.733  0.00628 **
## HOSPVIS      2.557e-01  3.850e-02   6.641 3.18e-11 ***
## DOCTOR       3.892e+00  6.491e-02  59.969  < 2e-16 ***
## HOSPITAL     1.775e+00  1.219e-01  14.557  < 2e-16 ***
## NEWHSAT     -5.575e-01  1.409e-02 -39.558  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.857 on 27285 degrees of freedom
## Multiple R-squared:  0.2719, Adjusted R-squared:  0.2716
## F-statistic: 926.3 on 11 and 27285 DF,  p-value: < 2.2e-16
```

We observe that **FACHHS** is no longer statistically significant at a reasonable level and so choose to remove it from our model and denote this new model as M_1:

```
M1 <- lm(DOCVIS ~ ID + FEMALE + YEAR + HANDPER + HHKIDS + UNIV
           + HOSPVIS + DOCTOR + HOSPITAL + NEWHSAT, data = Pr3c)
```

We analyze this new model M_1 on the next page.

Analysis of M_1:

```
summary(M1)
```

```
##
## Call:
## lm(formula = DOCVIS ~ ID + FEMALE + YEAR + HANDPER + HHKIDS +
##     UNIV + HOSPVIS + DOCTOR + HOSPITAL + NEWHSAT, data = Pr3c)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -12.198  -2.220  -0.550   0.897 110.414
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.697e+01  1.853e+01   3.074  0.00212 **
## ID          -8.853e-05  1.467e-05  -6.034 1.62e-09 ***
## FEMALE       4.242e-01  6.010e-02   7.057 1.74e-12 ***
## YEAR        -2.644e-02  9.324e-03  -2.836  0.00457 **
## HANDPER      2.262e-02  1.630e-03  13.878  < 2e-16 ***
## HHKIDS      -2.456e-01  6.092e-02  -4.031 5.58e-05 ***
## UNIV        -3.122e-01  1.150e-01  -2.715  0.00664 **
## HOSPVIS      2.555e-01  3.850e-02   6.637 3.25e-11 ***
## DOCTOR       3.892e+00  6.491e-02  59.965  < 2e-16 ***
## HOSPITAL     1.774e+00  1.219e-01  14.554  < 2e-16 ***
## NEWHSAT     -5.580e-01  1.409e-02 -39.609  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.857 on 27286 degrees of freedom
## Multiple R-squared:  0.2719, Adjusted R-squared:  0.2716
## F-statistic:  1019 on 10 and 27286 DF,  p-value: < 2.2e-16
```

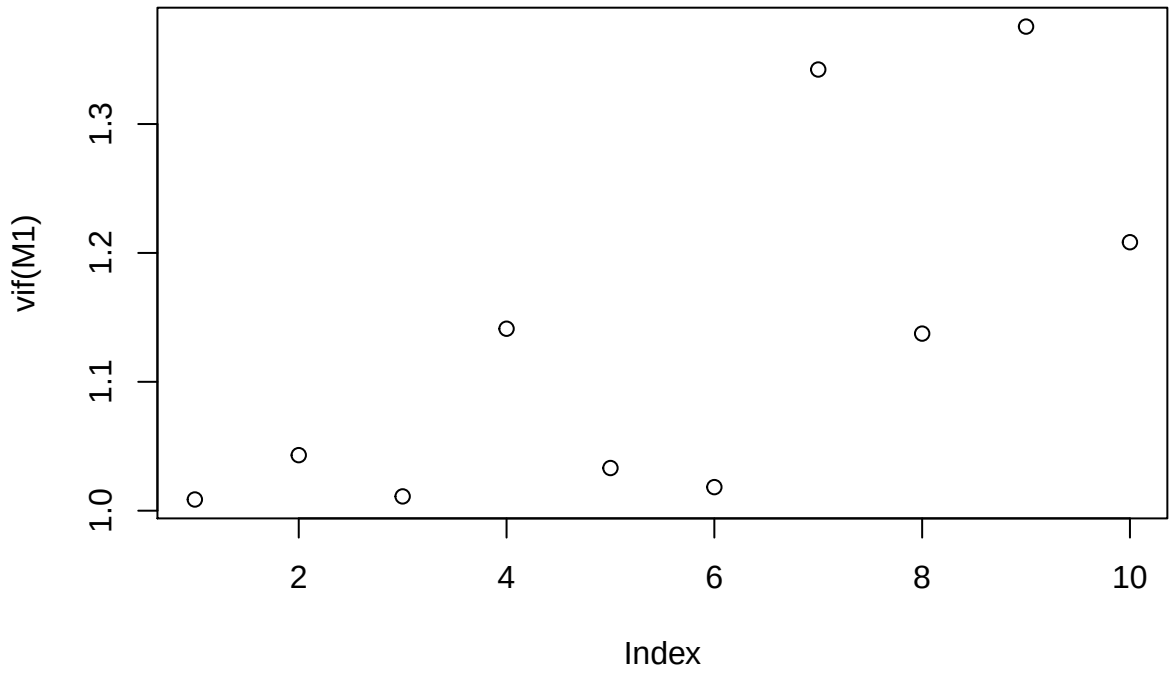
After discarding **FACHHS** from our model, all our variables are now statistically significant at the 0.01 level.

Note that **ID** is still statistically significant, although it has a negligible impact on the estimated value of **DOCVIS**.

We proceed to test for multicollinearity in model M_1 on the next page.

Testing for multicollinearity within model via VIF:

```
plot(vif(M1))
```



We observe that all the VIF values are close to 1, so we can safely say that multicollinearity will not be a problem for this model.

Note that the highest VIF values are for **HOSPVIS** and **HOSPITAL**, which makes sense, but observe that they are still low VIF values.

Thus, we choose to proceed with M_1 which we will henceforth call M_2 :.

```
M2 <- M1
```

Part II (d)

Data Imputation

NOTE: The original data set from moodle is **Pr3**. After we observed that more than 99.89% of our data in **Pr3** did not have missing values, we used the complete.cases() function to remove all the observations with missing values and called the new data set **Pr3c**. For all our analysis so far, **Pr3c** is the data set we have been using. So to impute all missing values, we must again use the original data set **Pr3**.

We proceed to impute all missing values in our predictors for the data set **Pr3new** (exact copy of **Pr3**):

```
Pr3new <- Pr3
imputedPr3 = impute(Pr3, target = character(0), cols = list(x = 99, y = imputeMode()))
# (1) Imputation for ID:
Pr3IId.na.ID <- is.na(Pr3new$ID)
Pr3new$ID[Pr3IId.na.ID] <- imputedPr3$ID[Pr3IId.na.ID]
# (2) Imputation for FEMALE:
Pr3IId.na.FEMALE <- is.na(Pr3new$FEMALE)
Pr3new$FEMALE[Pr3IId.na.FEMALE] <- imputedPr3$FEMALE[Pr3IId.na.FEMALE]
# (3) Imputation for YEAR:
Pr3IId.na.YEAR <- is.na(Pr3new$YEAR)
Pr3new$YEAR[Pr3IId.na.YEAR] <- imputedPr3$YEAR[Pr3IId.na.YEAR]
# (4) Imputation for HANDPER:
Pr3IId.na.HANDPER <- is.na(Pr3new$HANDPER)
Pr3new$HANDPER[Pr3IId.na.HANDPER] <- imputedPr3$HANDPER[Pr3IId.na.HANDPER]
# (5) Imputation for HHKIDS:
Pr3IId.na.HHKIDS <- is.na(Pr3new$HHKIDS)
Pr3new$HHKIDS[Pr3IId.na.HHKIDS] <- imputedPr3$HHKIDS[Pr3IId.na.HHKIDS]
# (6) Imputation for UNIV:
Pr3IId.na.UNIV <- is.na(Pr3new$UNIV)
Pr3new$UNIV[Pr3IId.na.UNIV] <- imputedPr3$UNIV[Pr3IId.na.UNIV]
# (7) Imputation for HOSPVIS:
Pr3IId.na.HOSPVIS <- is.na(Pr3new$HOSPVIS)
Pr3new$HOSPVIS[Pr3IId.na.HOSPVIS] <- imputedPr3$HOSPVIS[Pr3IId.na.HOSPVIS]
# (8) Imputation for DOCTOR:
Pr3IId.na.DOCTOR <- is.na(Pr3new$DOCTOR)
Pr3new$DOCTOR[Pr3IId.na.DOCTOR] <- imputedPr3$DOCTOR[Pr3IId.na.DOCTOR]
# (9) Imputation for HOSPITAL:
Pr3IId.na.HOSPITAL <- is.na(Pr3new$HOSPITAL)
Pr3new$HOSPITAL[Pr3IId.na.HOSPITAL] <- imputedPr3$HOSPITAL[Pr3IId.na.HOSPITAL]
# (10) Imputation for NEWHSAT:
Pr3IId.na.NEWHSAT <- is.na(Pr3new$NEWHSAT)
Pr3new$NEWHSAT[Pr3IId.na.NEWHSAT] <- imputedPr3$NEWHSAT[Pr3IId.na.NEWHSAT]
```

We reestimate our model using the predictors without missing values and denote this new model as M_3:

```
M3 <- lm(DOCVIS ~ ID + FEMALE + YEAR + HANDPER + HHKIDS + UNIV
           + HOSPVIS + DOCTOR + HOSPITAL + NEWHSAT, data = Pr3new)
```

We compare the AIC and BIC values of the two models:

```
print(data.frame(AIC(M3,M2), BIC(M3,M2)))
```

```
##    df      AIC df.1      BIC
## M3 12 163923.2   12 164021.8
## M2 12 163760.4   12 163859.0
```

We compare M_2 with M_3:

```
stargazer(M2, M3, title="M2 vs. M3", align = TRUE, header = FALSE, type = 'text',
digits = 7, order=c("Constant"), model.names = FALSE, column.labels = c("M2", "M3"))
```

```
##
## M2 vs. M3
## ================================================================================================
##                                                Dependent variable:
##                                 ---------------------------------------------------------------
##                                                       DOCVIS
##                                           M2                                M3
##                                           (1)                               (2)
## ------------------------------------------------------------------------------------------------
## Constant                                56.9698800***                     55.9111600***
##                                         (18.5337000)                      (18.5185600)
##
## ID                                      -0.0000885***                     -0.0000876***
##                                          (0.0000147)                       (0.0000146)
##
## FEMALE                                  0.4241774***                      0.4236615***
##                                          (0.0601037)                       (0.0600640)
##
## YEAR                                    -0.0264450***                     -0.0259153***
##                                          (0.0093237)                       (0.0093162)
##
## HANDPER                                 0.0226185***                      0.0226598***
##                                          (0.0016298)                       (0.0016292)
##
## HHKIDS                                  -0.2455580***                     -0.2431797***
##                                          (0.0609212)                       (0.0608788)
##
## UNIV                                    -0.3121601***                     -0.3132513***
##                                          (0.1149944)                       (0.1147378)
##
## HOSPVIS                                 0.2555304***                      0.2554288***
##                                          (0.0384983)                       (0.0384899)
##
## DOCTOR                                  3.8921210***                      3.8923990***
##                                          (0.0649063)                       (0.0648545)
##
## HOSPITAL                                1.7744440***                      1.7720600***
##                                          (0.1219174)                       (0.1218516)
##
## NEWHSAT                                 -0.5580376***                     -0.5577945***
##                                          (0.0140888)                       (0.0140780)
##
## ------------------------------------------------------------------------------------------------
## Observations                                27,297                            27,326
## R2                                          0.2718689                         0.2718361
## Adjusted R2                                 0.2716020                         0.2715695
## Residual Std. Error           4.8570370 (df = 27286)            4.8560450 (df = 27315)
## F Statistic              1,018.8020000*** (df = 10; 27286) 1,019.7160000*** (df = 10; 27315)
## ================================================================================================
## Note:                                                          *p<0.1; **p<0.05; ***p<0.01
```

Conclusion: Comparison of AIC, BIC, and R-squared

We observe that M_2 has a lower AIC, as well as a lower BIC, than M_3.

Also, M_2 has marginally higher R-squared and Adjusted R-squared values than M_3.

Basically, M_2 is a better model than M_3.

Part II (e)

```
resettest(M3, power=2, type="regressor")
```

```
##
##  RESET test
##
## data:  M3
## RESET = 67.188, df1 = 10, df2 = 27305, p-value < 2.2e-16
```

Due to the low p-value, we reject the null hypothesis that the functional form of the model is adequate.

```
summary(M3)
```

```
##
## Call:
## lm(formula = DOCVIS ~ ID + FEMALE + YEAR + HANDPER + HHKIDS +
##     UNIV + HOSPVIS + DOCTOR + HOSPITAL + NEWHSAT, data = Pr3new)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -12.193  -2.221  -0.549   0.896 110.418
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.591e+01  1.852e+01   3.019  0.00254 **
## ID          -8.761e-05  1.465e-05  -5.982 2.23e-09 ***
## FEMALE       4.237e-01  6.006e-02   7.054 1.79e-12 ***
## YEAR        -2.592e-02  9.316e-03  -2.782  0.00541 **
## HANDPER      2.266e-02  1.629e-03  13.909  < 2e-16 ***
## HHKIDS      -2.432e-01  6.088e-02  -3.994 6.50e-05 ***
## UNIV        -3.133e-01  1.147e-01  -2.730  0.00633 **
## HOSPVIS      2.554e-01  3.849e-02   6.636 3.28e-11 ***
## DOCTOR       3.892e+00  6.485e-02  60.017  < 2e-16 ***
## HOSPITAL     1.772e+00  1.219e-01  14.543  < 2e-16 ***
## NEWHSAT     -5.578e-01  1.408e-02 -39.622  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.856 on 27315 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.2718, Adjusted R-squared:  0.2716
## F-statistic:  1020 on 10 and 27315 DF,  p-value: < 2.2e-16
```

We first observe from the summary of M_3 that the variables **YEAR** and **UNIV** have the highest p-values.

We choose to include the following additional terms in our model: **(1)** We note that it is possible that time may be better represented as a quadratic term than a linear term so we decide to include a $\mathbf{YEAR^2}$ term in the model. **(2)** There may be diminishing returns to health satisfaction, so we will include an $\mathbf{NEWHSAT^2}$ term in our model as well. **(3)** Since those with a higher degree of disability may exhibit different behavior in visiting the doctor than those wil a lower degree of disability, we also include a $\mathbf{HANDPER^2}$ term in our model. **(4)** We also note that it is possible that **UNIV** and **NEWHSAT** may be better represented in the model through an interaction term **(UNIV)(NEWHSAT)** as individuals who attend university and are not satisfied with their health may go to the doctor more often. **(5)** Similarly, those who are female and have children may exhibit different behavior than the rest of the population in number of doctor visits. So we include a **(FEMALE)(HHKIDS)** interaction term in our model as well.

We define a new model M_4:

```
M4 <- lm(DOCVIS ~ ID + FEMALE + YEAR + I(YEAR*YEAR) + HANDPER + I(HANDPER*HANDPER)
         + HHKIDS + UNIV + HOSPVIS + DOCTOR + HOSPITAL + NEWHSAT + I(NEWHSAT*NEWHSAT)
         + I((UNIV)*(NEWHSAT)) + I((FEMALE)*(HHKIDS)), data = Pr3new)
```

Analysing new model M_4

```
summary(M4)
```

```
##
## Call:
## lm(formula = DOCVIS ~ ID + FEMALE + YEAR + I(YEAR * YEAR) + HANDPER +
##     I(HANDPER * HANDPER) + HHKIDS + UNIV + HOSPVIS + DOCTOR +
##     HOSPITAL + NEWHSAT + I(NEWHSAT * NEWHSAT) + I((UNIV) * (NEWHSAT)) +
##     I((FEMALE) * (HHKIDS)), data = Pr3new)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -13.279  -2.115  -0.365   0.698 107.729
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)              1.133e+05  1.243e+04   9.109  < 2e-16 ***
## ID                      -8.499e-05  1.448e-05  -5.871 4.38e-09 ***
## FEMALE                   5.822e-01  7.624e-02   7.637 2.30e-14 ***
## YEAR                    -1.139e+02  1.250e+01  -9.108  < 2e-16 ***
## I(YEAR * YEAR)           2.863e-02  3.143e-03   9.107  < 2e-16 ***
## HANDPER                  6.114e-02  5.223e-03  11.706  < 2e-16 ***
## I(HANDPER * HANDPER)    -5.706e-04  6.605e-05  -8.639  < 2e-16 ***
## HHKIDS                  -1.160e-01  8.246e-02  -1.407  0.15955
## UNIV                    -1.943e+00  4.224e-01  -4.601 4.23e-06 ***
## HOSPVIS                  2.341e-01  3.803e-02   6.156 7.55e-10 ***
## DOCTOR                   3.987e+00  6.422e-02  62.082  < 2e-16 ***
## HOSPITAL                 1.642e+00  1.205e-01  13.622  < 2e-16 ***
## NEWHSAT                 -1.749e+00  5.463e-02 -32.018  < 2e-16 ***
## I(NEWHSAT * NEWHSAT)     9.811e-02  4.380e-03  22.398  < 2e-16 ***
## I((UNIV) * (NEWHSAT))    2.235e-01  5.555e-02   4.023 5.77e-05 ***
## I((FEMALE) * (HHKIDS)) -3.448e-01  1.186e-01  -2.907  0.00365 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.796 on 27310 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.2897, Adjusted R-squared:  0.2894
## F-statistic: 742.7 on 15 and 27310 DF,  p-value: < 2.2e-16
```

We observe that the **HHKIDS** term is no longer significant, but the **(FEMALE)(HHKIDS)** term is significant, as are all the other quadratic and interaction terms we added to the model.

We compare the AIC and BIC values of the two models:

```
print(data.frame(AIC(M4,M3), BIC(M4,M3)))
```

```
##    df      AIC df.1      BIC
## M4 17 163252.9   17 163392.6
## M3 12 163923.2   12 164021.8
```

We compare M_3 with M_4:

```
stargazer(M3, M4, title="M3 vs. M4", align = TRUE, header = FALSE, type = 'text',
digits = 7, order=c("Constant"), model.names = FALSE, column.labels = c("M3", "M4"),
no.space = TRUE)
```

```
##
## M3 vs. M4
## ================================================================================================
##                                                Dependent variable:
##                                  ---------------------------------------------------------------
##                                                        DOCVIS
##                                           M3                                  M4
##                                           (1)                                 (2)
## ------------------------------------------------------------------------------------------------
## Constant                                55.9111600***                  113,273.8000000***
##                                         (18.5185600)                    (12,434.7300000)
## ID                                      -0.0000876***                     -0.0000850***
##                                          (0.0000146)                       (0.0000145)
## FEMALE                                  0.4236615***                      0.5822163***
##                                          (0.0600640)                       (0.0762357)
## YEAR                                    -0.0259153***                    -113.8884000***
##                                          (0.0093162)                       (12.5039600)
## I(YEAR * YEAR)                                                            0.0286283***
##                                                                           (0.0031434)
## HANDPER                                 0.0226598***                      0.0611425***
##                                          (0.0016292)                       (0.0052231)
## I(HANDPER * HANDPER)                                                      -0.0005706***
##                                                                           (0.0000660)
## HHKIDS                                  -0.2431797***                     -0.1159919
##                                          (0.0608788)                       (0.0824609)
## UNIV                                    -0.3132513***                     -1.9432630***
##                                          (0.1147378)                       (0.4223770)
## HOSPVIS                                 0.2554288***                      0.2341411***
##                                          (0.0384899)                       (0.0380331)
## DOCTOR                                  3.8923990***                      3.9872260***
##                                          (0.0648545)                       (0.0642249)
## HOSPITAL                                1.7720600***                      1.6418500***
##                                          (0.1218516)                       (0.1205269)
## NEWHSAT                                 -0.5577945***                     -1.7489700***
##                                          (0.0140780)                       (0.0546250)
## I(NEWHSAT * NEWHSAT)                                                      0.0981111***
##                                                                           (0.0043804)
## I((UNIV) * (NEWHSAT))                                                     0.2234603***
##                                                                           (0.0555488)
## I((FEMALE) * (HHKIDS))                                                    -0.3448379***
##                                                                           (0.1186146)
## ------------------------------------------------------------------------------------------------
## Observations                              27,326                              27,326
## R2                                       0.2718361                           0.2897408
## Adjusted R2                              0.2715695                           0.2893507
## Residual Std. Error           4.8560450 (df = 27315)              4.7964100 (df = 27310)
## F Statistic             1,019.7160000*** (df = 10; 27315) 742.7168000*** (df = 15; 27310)
## ================================================================================================
## Note:                                                          *p<0.1; **p<0.05; ***p<0.01
```

Conclusion: Comparison of AIC, BIC, and R-squared

We observe that M_4 has a lower AIC, as well as a lower BIC, than M_3.
Also, M_4 has higher R-squared and Adjusted R-squared values than M_3.

Basically, M_4 is a better model than M_3. Therefore, we elect to continue with M_4.

Part II (f)

Mallows CP

Calculating Mallows CP statistics for M3 and M4:

```
M3MallowsCP = ( (anova(M3)[11,2])/(anova(M3)[11,3]) - 27326 + (2*10) )
M4MallowsCP = ( (anova(M4)[16,2])/(anova(M4)[16,3]) - 27326 + (2*15) )
print(data.frame(c("M3",M3MallowsCP),c("M4", M4MallowsCP))[1:2,])
```

```
##   c..M3...M3MallowsCP. c..M4...M4MallowsCP.
## 1                   M3                   M4
## 2                    9                   14
```

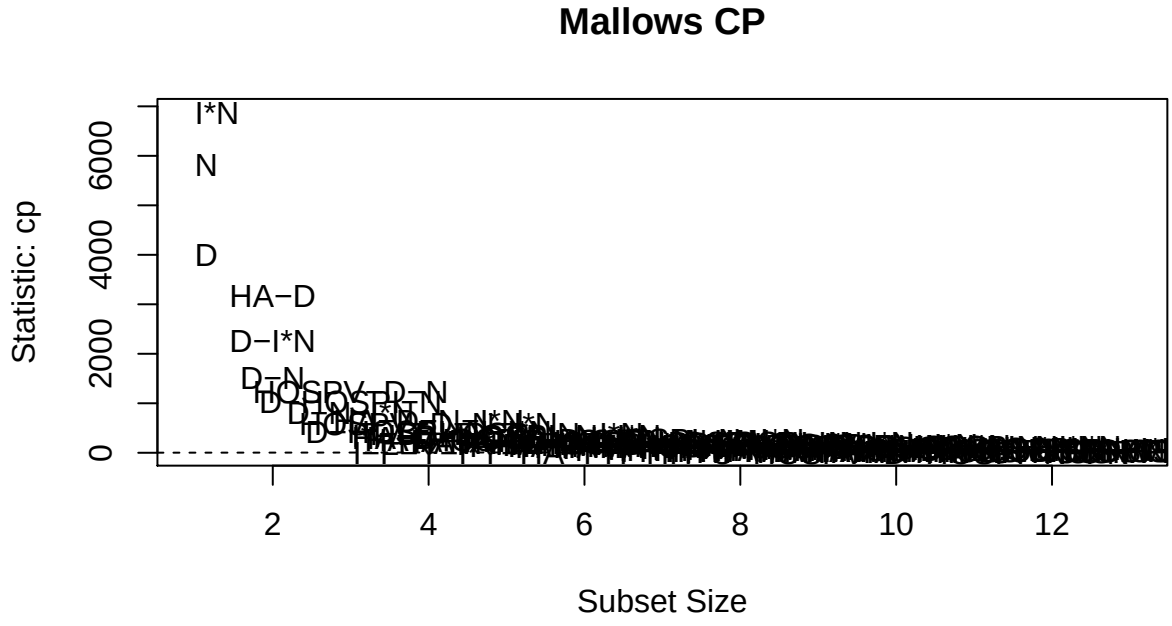
M3 has a Mallows CP value of 9 and M4 has a Mallows CP value of 14. Due to both models having CP values close to the value of the number of parameters, they are both reasonably good models.

Choosing preferred model based on Mallows CP statistic:

We know that the Mallows CP statistic will punish us for having unnecessary parameters in our model, as well as for having a lot of unnecessary parameters. In order to choose a model based on the Mallows CP statistic, we proceed first by looking at what parameters we can eliminate without substantially reducing the effectiveness of the model.

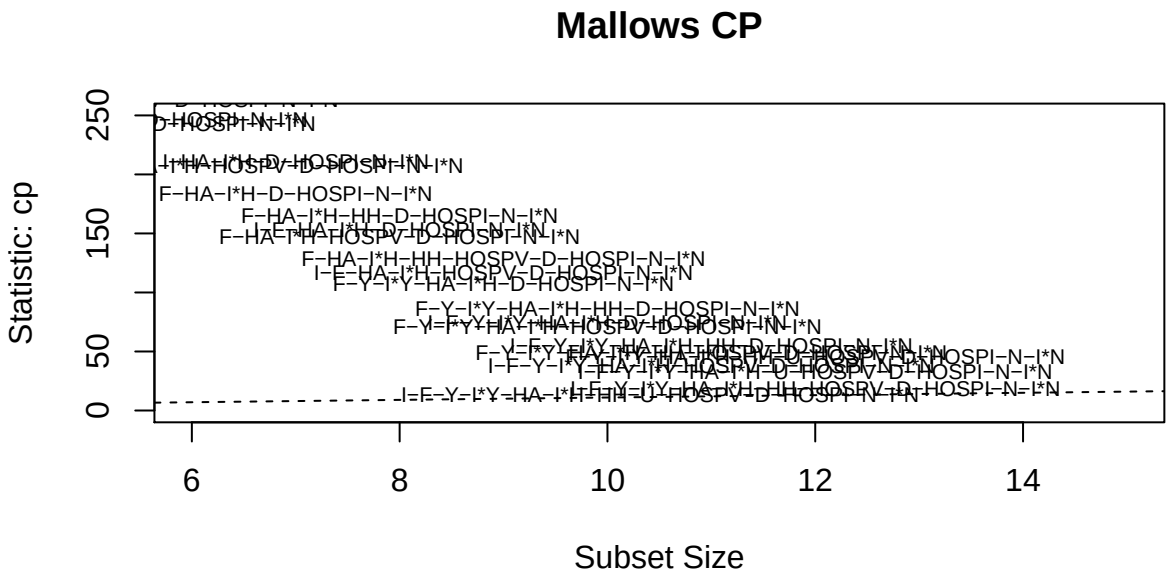
We look at Mallows CP values for subsets of the variables used in M4:

```
regsubsets.Pr3IIfi <- regsubsets(DOCVIS ~ ID + FEMALE + YEAR + I(YEAR*YEAR) + HANDPER
    + I(HANDPER*HANDPER) + HHKIDS + UNIV + HOSPVIS + DOCTOR + HOSPITAL + NEWHSAT
    + I(NEWHSAT*NEWHSAT), data = Pr3new, nbest = 3, nvmax = NULL, force.in = NULL,
    force.out = NULL, method = "exhaustive")
res.legend <- subsets(regsubsets.Pr3IIfi,statistic="cp",legend=T,main="Mallows CP")
abline(a = 1, b = 1, lty = 2)
```



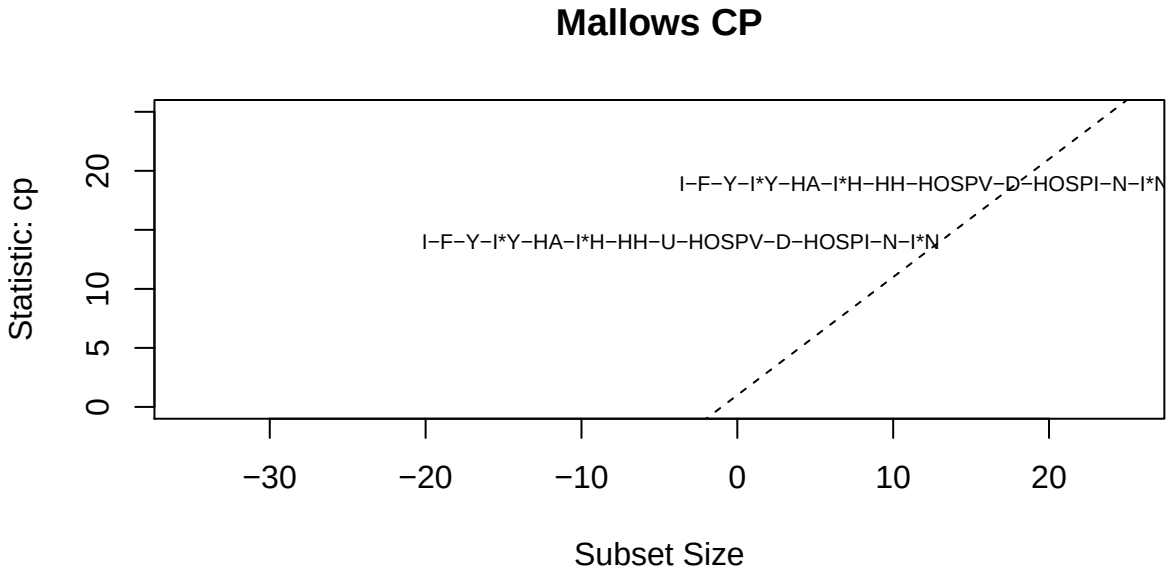
We taker a closer look at the **lower** range of Mallows CP values:

```
res.legend <- subsets(regsubsets.Pr3IIfi,statistic="cp",legend=F,main="Mallows CP",
        xlim = c(6,15), ylim = c(0,250), cex = 0.7)
abline(a = 1, b = 1, lty = 2)
```



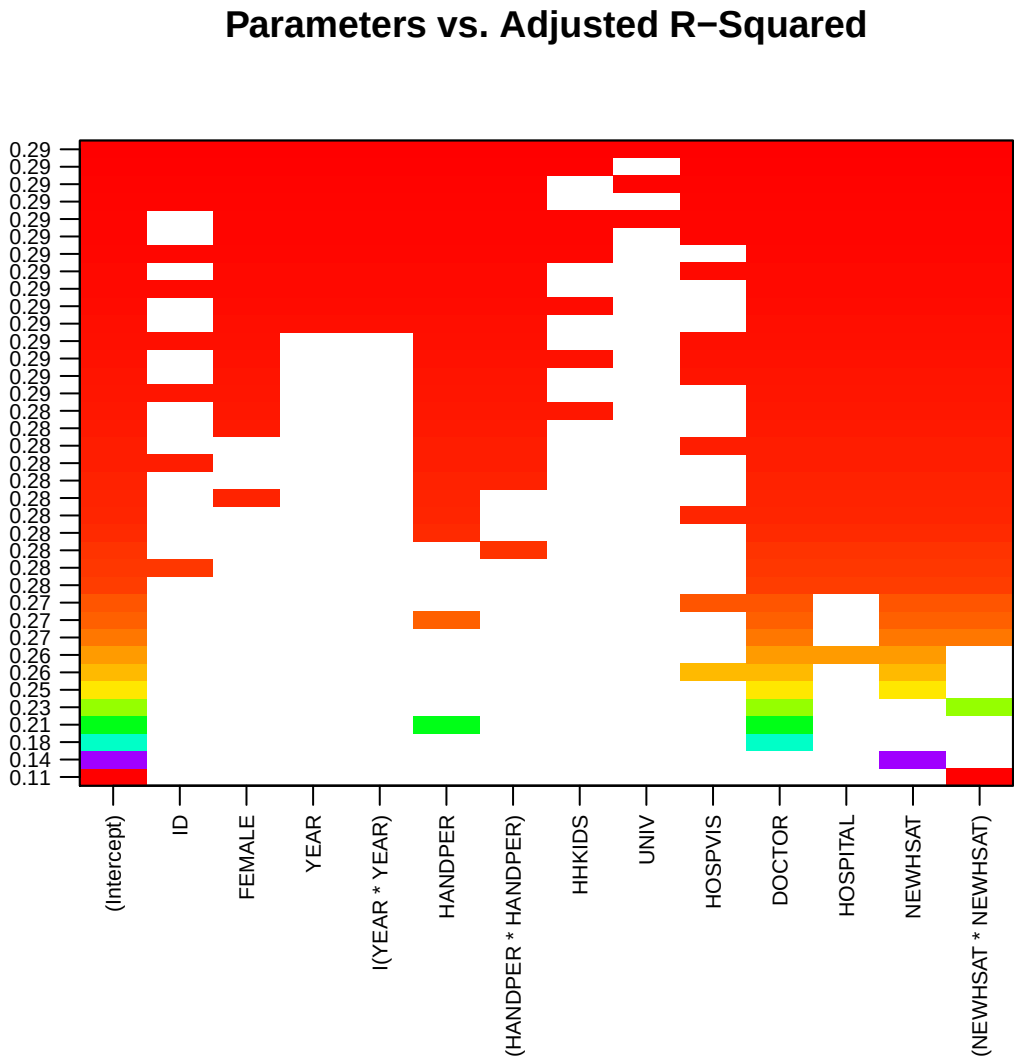
We then take a closer look at the **lowest** range of Mallows CP values:

```
res.legend <- subsets(regsubsets.Pr3IIfi,statistic="cp",legend=F,main="Mallows CP",
xlim = c(-35,25), ylim = c(0,25), cex = 0.7)
abline(a = 1, b = 1, lty = 2)
```



We analyze different subsets of parameters compared to their R-squared values to see which parameters we can remove from our model, in order to achieve a lower Mallows CP value:

```
par(cex.axis = 0.7, mai=c(1.02,0,0,0), mgp=c(10,0.75,0))
plot(regsubsets.Pr3IIfi, scale = "adjr2", main = "Parameters vs. Adjusted R-Squared",
col = rainbow(100000))
```



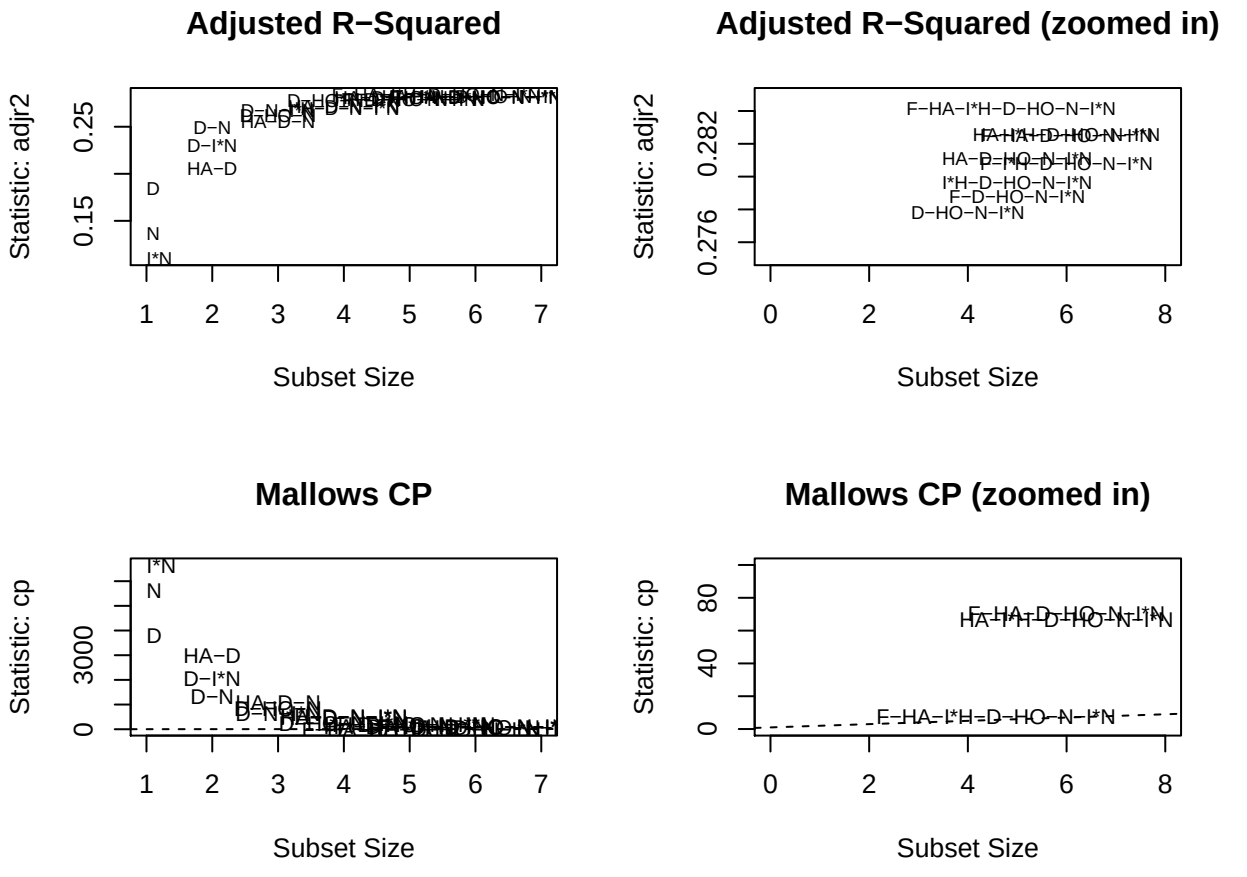
We observe that most models that reach high Adjusted R-squared values do so without the variable **UNIV**. We also note that the variables **ID**, **HHKIDS**, and **HOSPVIS** are not necessary for high Adjusted R-squared values. Lastly, we note that the **YEAR** and **(YEAR)(YEAR)** variables are also unnecessary for high Adjusted R-squared values. We therefore elect to remove all these variables from our model, and are left with:

- FEMALE
- HANDPER
- I(HANDPER*HANDPER)
- DOCTOR
- HOSPITAL
- NEWHSAT
- I(NEWHSAT*NEWHSAT)

We proceed to plot Mallows CP and Adjusted R-squared values of models based on the chosen variables:.

```
regsubsets.Pr3IIfii <- regsubsets(DOCVIS ~ FEMALE + HANDPER + I(HANDPER*HANDPER)
    + DOCTOR + HOSPITAL  + NEWHSAT + I(NEWHSAT*NEWHSAT), data = Pr3new, nbest = 3,
    nvmax = NULL, force.in = NULL, force.out = NULL, method = "exhaustive")

layout(matrix(1:4, nrow = 2))
res.legend <- subsets(regsubsets.Pr3IIfii, statistic="adjr2", cex = 0.7, legend = TRUE,
min.size = 1, main = "Adjusted R-Squared")
res.legend <- subsets(regsubsets.Pr3IIfii, statistic="cp", cex = 0.8, legend = TRUE,
min.size = 1, main = "Mallows CP")
abline(a = 1, b = 1, lty = 2)
res.legend <- subsets(regsubsets.Pr3IIfii, statistic="adjr2", cex = 0.7,
    legend = TRUE, min.size = 1, main = "Adjusted R-Squared (zoomed in)",
    xlim = c(0,8), ylim = c(0.275,0.285))
res.legend <- subsets(regsubsets.Pr3IIfii, statistic="cp", cex = 0.8,
    legend = TRUE, min.size = 1, main = "Mallows CP (zoomed in)", xlim = c(0,8),
     ylim = c(0,100))
abline(a = 1, b = 1, lty = 2)
```



By using the Mallows CP statistic, we thus observe that the model we have chosen maximizes its Adjusted R-squared value while minimizing the number of parameters it uses. We denote this model as M_5:

```
M5 <- lm(DOCVIS ~ FEMALE + HANDPER + I(HANDPER*HANDPER) + DOCTOR
       + HOSPITAL  + NEWHSAT + I(NEWHSAT*NEWHSAT), data = Pr3new)
```

III. Model Validation

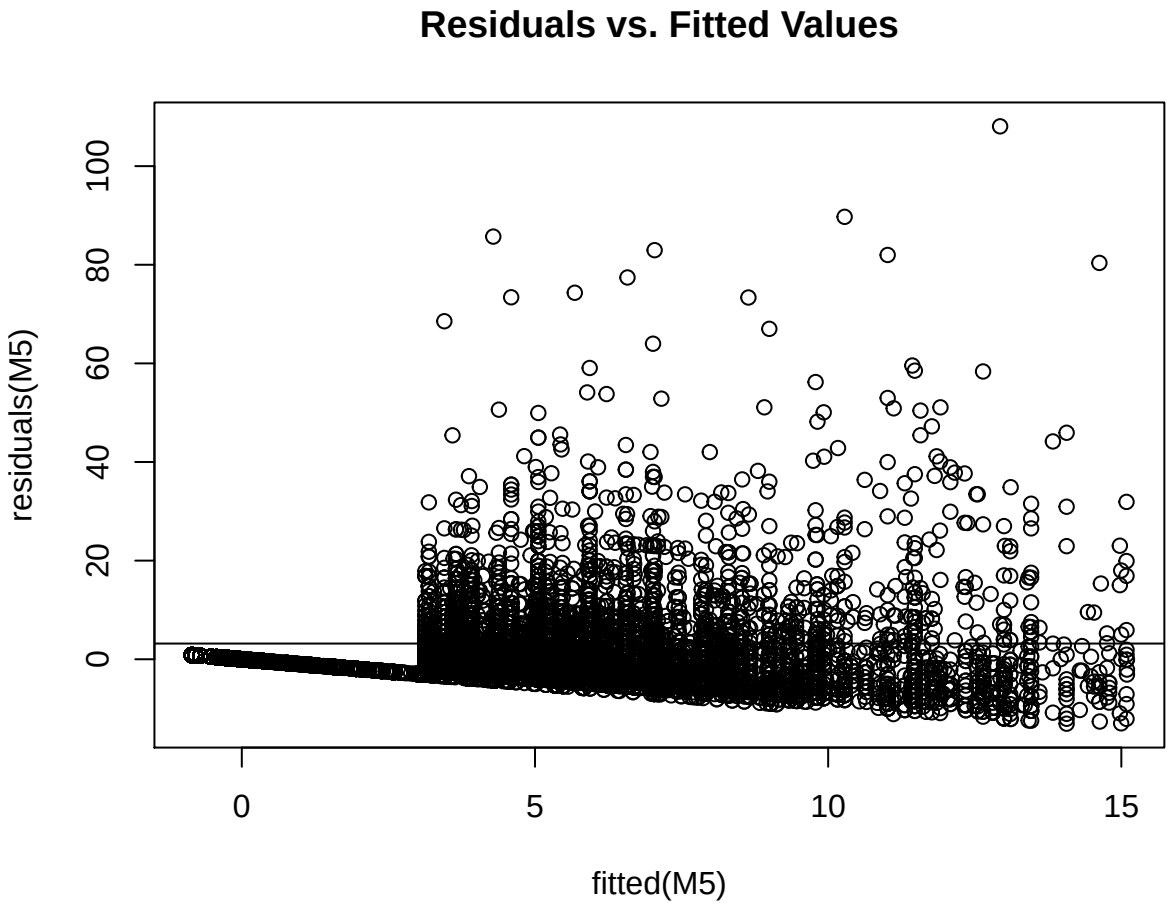
Part III (a)

Plotting Residuals vs. Fitted Values

```
plot(fitted(M5),residuals(M5), main = "Residuals vs. Fitted Values")
lm(fitted(M5)~residuals(M5))
```

```
##
## Call:
## lm(formula = fitted(M5) ~ residuals(M5))
##
## Coefficients:
##   (Intercept)  residuals(M5)
##     3.184e+00     -1.429e-16
```

```
abline(lm(fitted(M5)~residuals(M5)))
```

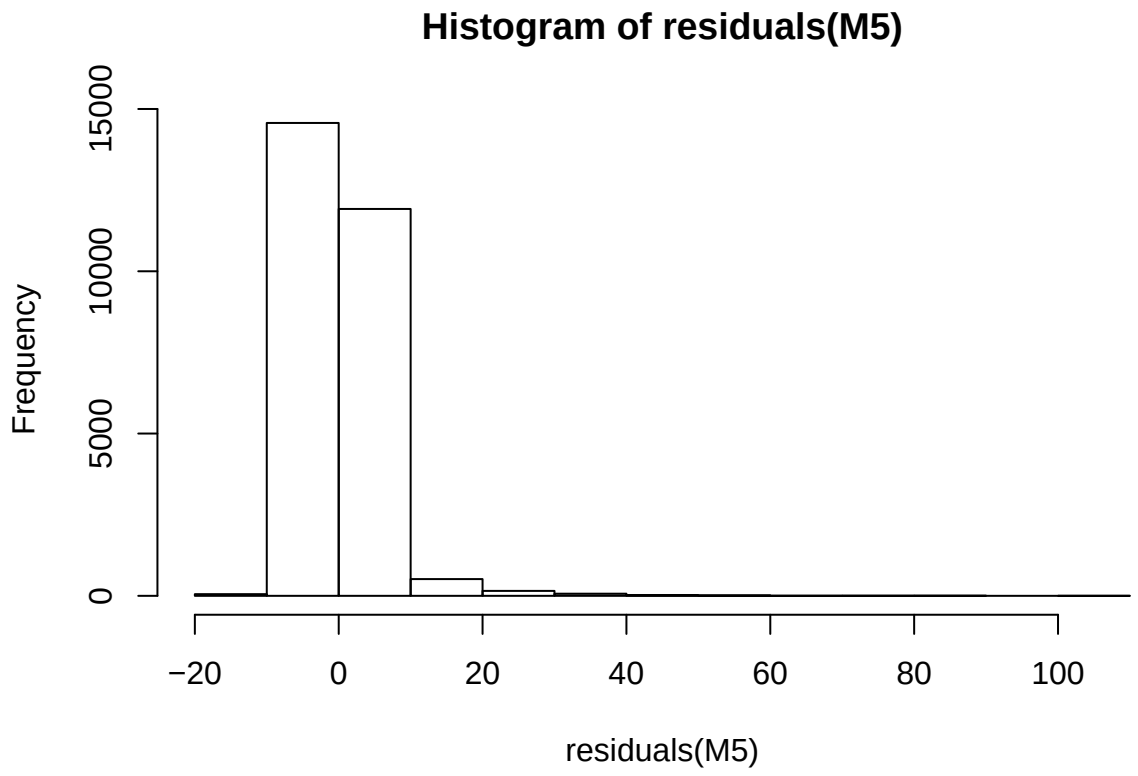


The mean of the residuals is greater than zero. There does appear to be a pattern in the residual values. For instance, between fitted values of around -1 to 3.5, the residuals follow a linear pattern. Also, between fitted values of 3.5 to 15, there appear to be vertical bars of clustered residual values.

Part III (b)

Histogram of Residuals

```
hist(residuals(M5))
```



Jarque-Bera Test

```
jarque.bera.test(residuals(M5))
```

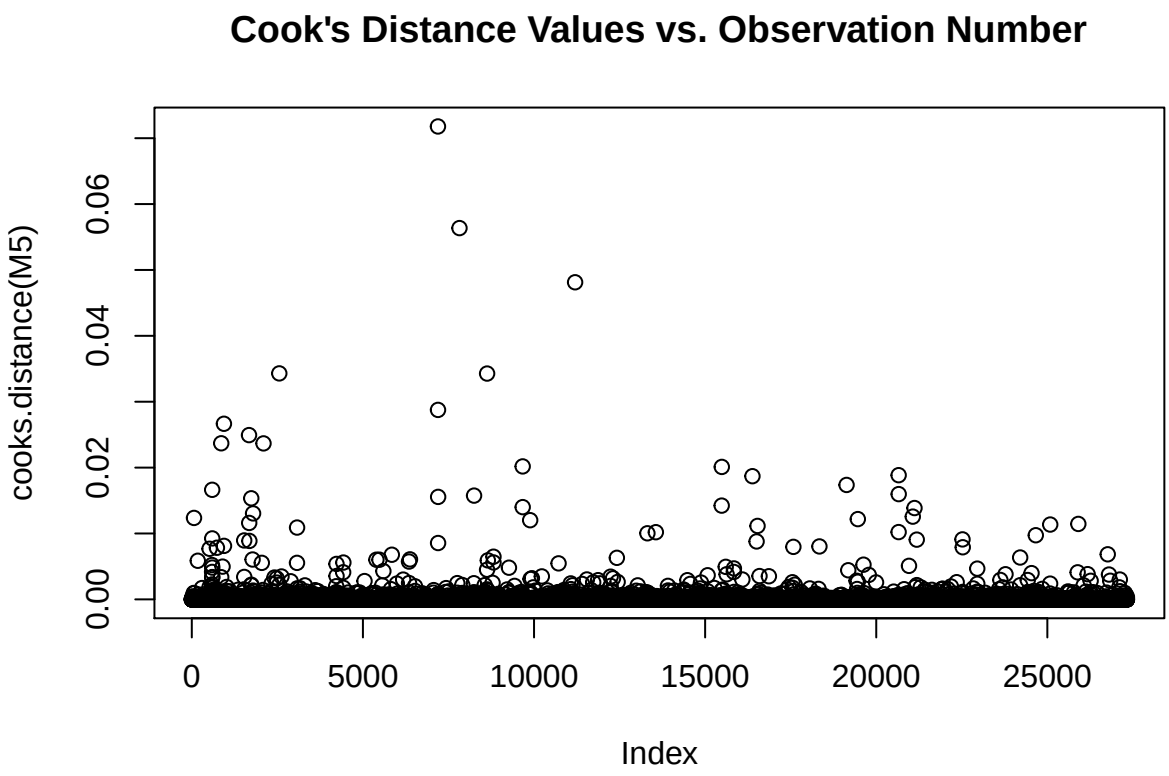
```
##
##  Jarque Bera Test
##
## data:  residuals(M5)
## X-squared = 4706500, df = 2, p-value < 2.2e-16
```

Due to the low p-value, we reject the null hypothesis and conclude that the residuals of M_5 do not follow a normal distribution.

Part III (c)

Plotting Cook's Distance Values

```
plot(cooks.distance(M5), main = "Cook's Distance Values vs. Observation Number")
```



Considering Outliers:

```
Pr3IIIcOutliers <- Pr3new[which(cooks.distance(M5)>(4*mean(cooks.distance(M5)))),]
str(Pr3IIIcOutliers[,0])
```

```
## 'data.frame':    804 obs. of  0 variables
```

Observe that there are **804** observations that can be considered outliers.
. Note that we consider any observation with a Cook's Distance value 4 times or greater the mean on the Cook's Distance values to be an outlier.

Note also that since we have 27,326 observations, 804 outliers account for less than **3%** of all the observations in our data.

Since less than 2% of the observations are outliers, we propose removing them from our data set. We proceed to do just that:

```
Pr3IIInooutliers<-Pr3new[-which(cooks.distance(M5)>(4*mean((cooks.distance(M5))))),]
```

We now proceed to reestimate our model using just the data without outliers:

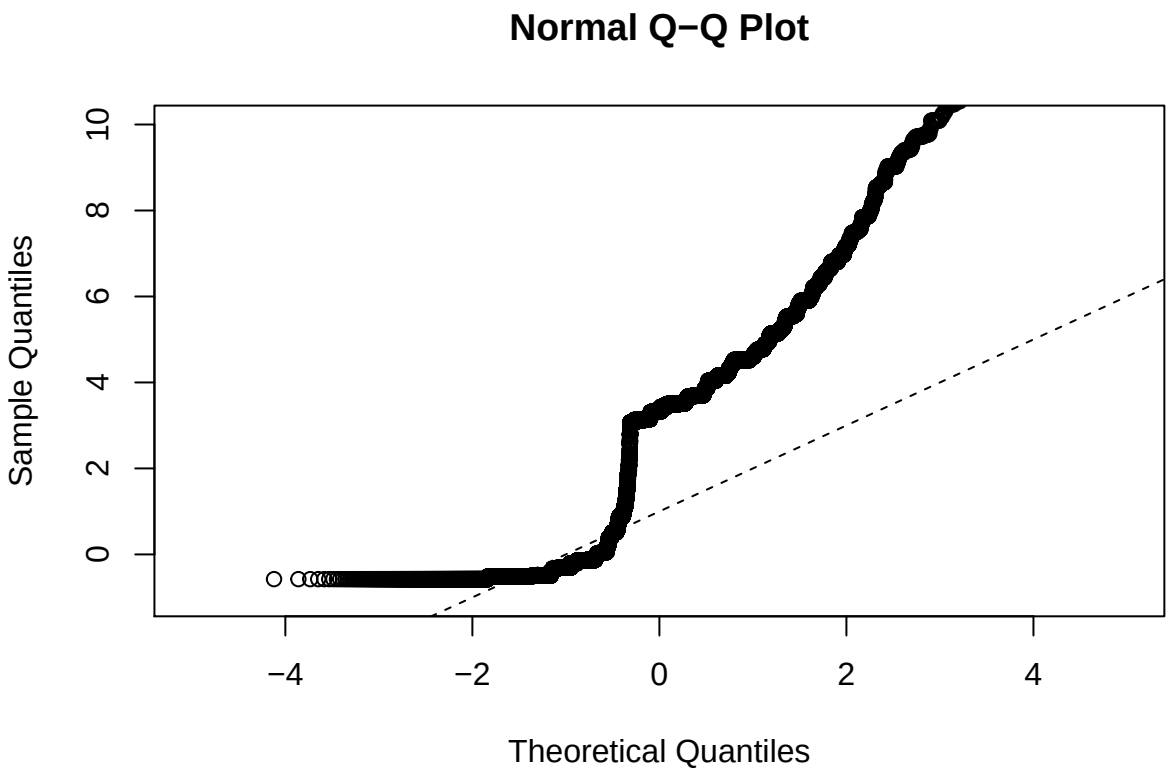
```
M6 <- lm(DOCVIS ~ FEMALE + HANDPER + I(HANDPER*HANDPER) + DOCTOR
       + HOSPITAL  + NEWHSAT + I(NEWHSAT*NEWHSAT), data = Pr3IIInooutliers)
summary(M6)
```

```
##
## Call:
## lm(formula = DOCVIS ~ FEMALE + HANDPER + I(HANDPER * HANDPER) +
##     DOCTOR + HOSPITAL + NEWHSAT + I(NEWHSAT * NEWHSAT), data = Pr3IIInooutliers)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -9.090 -1.655 -0.130  0.513 59.858
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)           5.015e+00  1.298e-01  38.643  < 2e-16 ***
## FEMALE                3.661e-01  3.699e-02   9.900  < 2e-16 ***
## HANDPER               2.507e-02  3.466e-03   7.233 4.85e-13 ***
## I(HANDPER * HANDPER) -2.056e-04  4.548e-05  -4.521 6.18e-06 ***
## DOCTOR                3.643e+00  3.963e-02  91.940  < 2e-16 ***
## HOSPITAL              1.065e+00  6.796e-02  15.672  < 2e-16 ***
## NEWHSAT              -1.250e+00  3.919e-02 -31.894  < 2e-16 ***
## I(NEWHSAT * NEWHSAT)  6.988e-02  3.046e-03  22.941  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.952 on 26514 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.383,  Adjusted R-squared:  0.3828
## F-statistic:  2351 on 7 and 26514 DF,  p-value: < 2.2e-16
```

We observe that all the explanatory variables in our model are statistically significant. In addition, the Adjusted R-squared value of the new M_6 is 0.3767, which is substantially higher than the 0.2842 Adjusted R-squared value of M_5.

Part III (d)

```
qqnorm(fitted(M6), xlim = c(-5,5), ylim = c(-1,10))
abline(a = 1, b = 1, lty = 2)
```

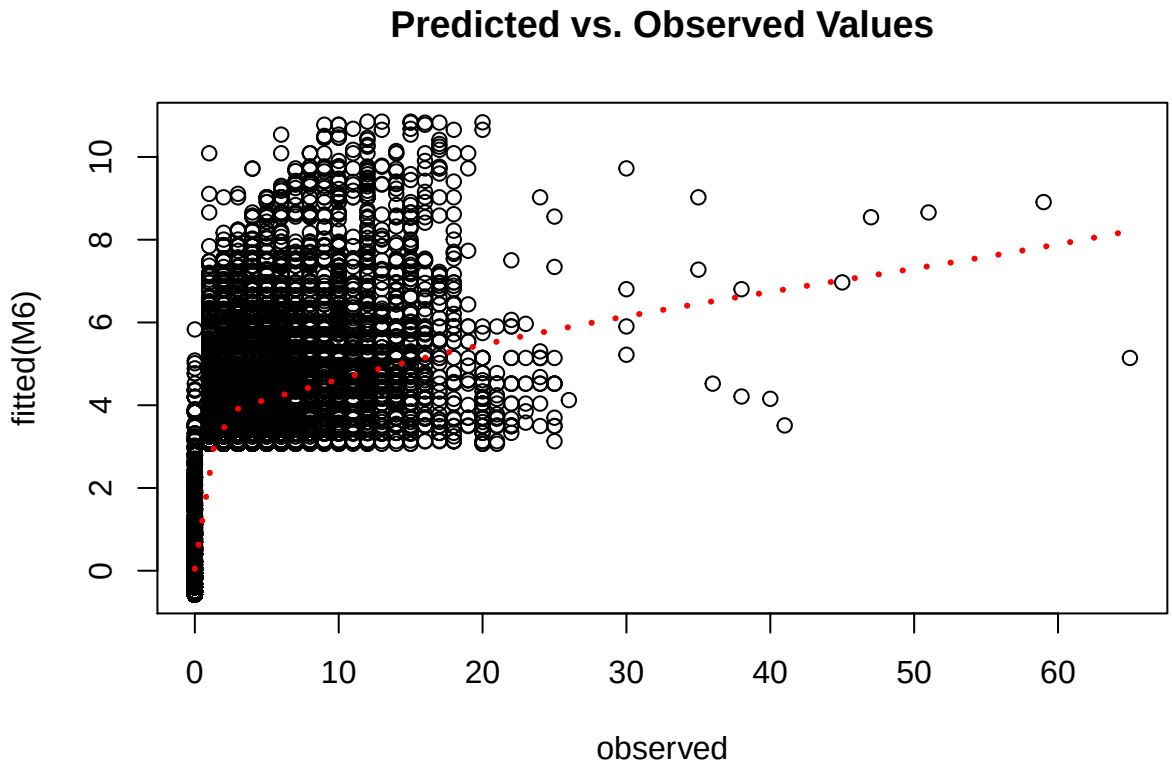


Observe that the quantiles from M_6 do not match the theoretical quantiles. This implies that the fitted data does not follow a normal distribution.

Part III (e)

We proceed to plot observed vs. fitted values but note that there is an NA value in the original data set (we were instructed to impute all missing values in our predictors only - not the entire data set). Thus we plot the fitted values of M_6 only to the non-missing data in the data set:

```
scatter.smooth(Pr3IIInooutliers$DOCVIS[!is.na(Pr3IIInooutliers$DOCVIS)],
               fitted(M6), lpars=list(col="red",lwd=3,lty=3),
               xlab = "observed", main = "Predicted vs. Observed Values")
```



We note that the predicted values correspond very well to observed values between 0 and about 3 or 4. As the magnitude of the observed values increases, the fitted values begin to diverge from the observed values.

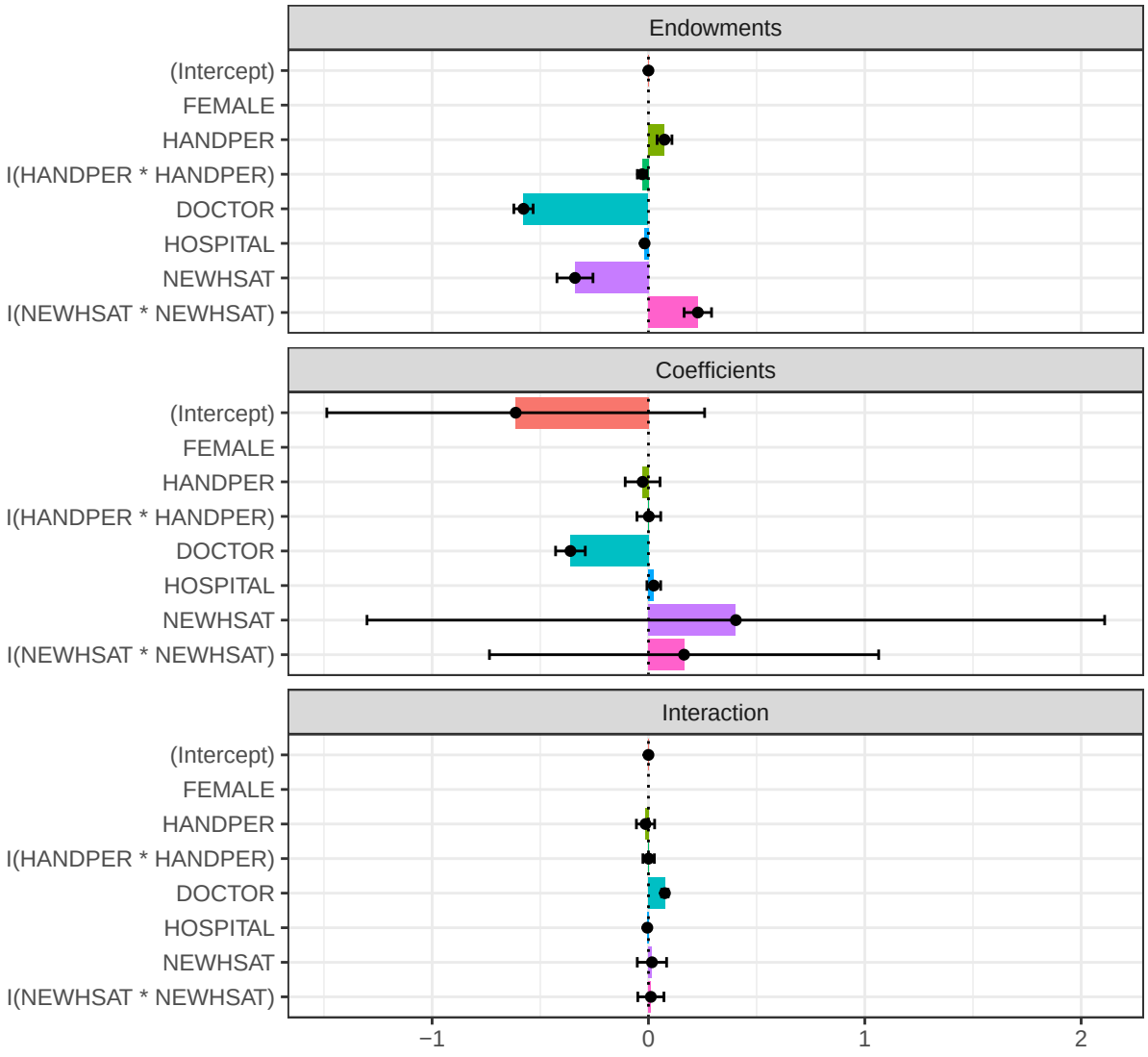
Part III (f)

Oaxaca Decomposition

```
oaxaca.results.1 <- oaxaca(DOCVIS ~ FEMALE + HANDPER + I(HANDPER*HANDPER) + DOCTOR
      + HOSPITAL  + NEWHSAT + I(NEWHSAT*NEWHSAT) | FEMALE, data = Pr3IIInooutliers)
```

```
## oaxaca: oaxaca() performing analysis. Please wait.
##
## Bootstrapping standard errors:
## 1 / 100 (1%)
## 10 / 100 (10%)
## 20 / 100 (20%)
## 30 / 100 (30%)
## 40 / 100 (40%)
## 50 / 100 (50%)
## 60 / 100 (60%)
## 70 / 100 (70%)
## 80 / 100 (80%)
## 90 / 100 (90%)
## 100 / 100 (100%)
```

```
plot(oaxaca.results.1)
```



Note that the *Endowments* are the explained differences based on the explanatory variables, and the *Coefficients* are the unexplained differences that are not explained by differences in the explanatory variables.

We note a few interesting observations. First of all, we observe that the **DOCTOR** variable results in an endowment of around -0.6 and that **NEWHSAT** results in an endowment of around -0.25. So these differences are the result of explanatory variables across groups.

Then, considering the unexplained variables, the coefficients, we note that females have a negative intercept and also have a negative estimate for the **DOCTOR** variable. However, due to the size of the confidence interval for the intercept, it is difficult to properly extract meaning out of the value.

Also, we note that the variables **NEWHSAT** and **NEWHSAT^2** have unexplained positive estimates, but that their estimates are difficult to interpret as they have incredibly wide confidence intervals.

In general, outside of the **DOCTOR** variable, none of the other unexplained differences are both reasonably statistially interpretable and significant.

Part III (g)

```
Pr3IIIfprediction<-data.frame(FEMALE = 0, YEAR = 1991, HANDPER = 0,
                              DOCTOR = 0, HOSPITAL = 1, NEWHSAT = 8)
predict(M6,Pr3IIIfprediction,interval="confidence")
```

```
##         fit       lwr       upr
## 1 0.5520883 0.4060911 0.6980856
```

We predict that an individual who is not handicapped, is quite satisfied with his health, who has not visited the doctor in the last three months, but who has gone to the hospital in the last year - would be unlikely to visit the doctor. Our prediction supports this conclusion.

The estimate is 0.5520883 and the 95% confidence interval is (0.4060911,0.6980856).

IV. Hypotheses Tests

Part IV (a) Differences in Differences

i. Determine whether or not the policy worked for women.

Our null hypothesis will be that womens' doctor visits *did not change* during or after 1987. For this we will test the mean value of doctor visits for females. Our alternate hypothesis will be that womens' doctor visits *increased* during or after 1987.

```
Pr3IViA <- Pr3IIInooutliers[Pr3IIInooutliers$FEMALE == "1"
                          & Pr3IIInooutliers$YEAR >= "1987",]$DOCVIS
Pr3IViB <- Pr3IIInooutliers[Pr3IIInooutliers$FEMALE == "1"
                          & Pr3IIInooutliers$YEAR < "1987",]$DOCVIS
t.test(Pr3IViA, Pr3IViB, alternative = "greater")
```

```
##
##  Welch Two Sample t-test
##
## data:  Pr3IViA and Pr3IViB
## t = -0.18707, df = 11259, p-value = 0.5742
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  -0.1341699        Inf
## sample estimates:
## mean of x mean of y
##  3.153532  3.167232
```

We observe that the p-value is around 0.5688, which is relatively high. Thus, we fail to reject the null hypothesis.

ii. Determine whether or not the policy worked for unemployed.

Our null hypothesis will be that there will be *no difference* between how much unemployed individuals visited the doctor before 1987 and how much they visited the doctor during or after 1987. For this, we will test the mean value of doctor visits for unemployed individuals. Our alternate hypothesis will be that the doctor visits of unemployed individuals *increased* during or after 1987.

```
Pr3IViiA <- Pr3IIInooutliers[Pr3IIInooutliers$WORKING == "0"
                          & Pr3IIInooutliers$YEAR >= "1987",]$DOCVIS
Pr3IViiB <- Pr3IIInooutliers[Pr3IIInooutliers$WORKING == "0"
                          & Pr3IIInooutliers$YEAR < "1987",]$DOCVIS
t.test(Pr3IViiA, Pr3IViiB, alternative = "greater")
```

```
##
##  Welch Two Sample t-test
##
## data:  Pr3IViiA and Pr3IViiB
## t = 0.33137, df = 8253.2, p-value = 0.3702
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  -0.1197385        Inf
## sample estimates:
## mean of x mean of y
##  3.355683  3.325479
```

We observe that the p-value is around 0.4487, which is relatively high. Thus, we fail to reject the null hypothesis.

Part IV (b)

We test the hypothesis that the number of doctor visits a patient has over a 3 month period is greater for women than for men.

```
Pr3IVb1 <- Pr3IIInooutliers[Pr3IIInooutliers$FEMALE == "1",]$DOCVIS
Pr3IVb2 <- Pr3IIInooutliers[Pr3IIInooutliers$FEMALE == "0",]$DOCVIS
t.test(Pr3IVb1, Pr3IVb2, alternative = "greater")
```

```
##
##  Welch Two Sample t-test
##
## data:  Pr3IVb1 and Pr3IVb2
## t = 21.306, df = 24812, p-value < 2.2e-16
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  0.9075158       Inf
## sample estimates:
## mean of x mean of y
##  3.159276  2.175834
```

Due to the low p-value, we reject the null hypothesis and conclude that the number of doctor visits a patient has over a 3 month period is greater for women than for men.

Part IV (c)

We test the hypothesis that the number of doctor visits a patient has over a 3 month period is different for women with children than for women without children.

```
Pr3IVc1 <- Pr3IIInooutliers[Pr3IIInooutliers$FEMALE == "1"
                          & Pr3IIInooutliers$HHKIDS == "1",]$DOCVIS
Pr3IVc2 <- Pr3IIInooutliers[Pr3IIInooutliers$FEMALE == "1"
                          & Pr3IIInooutliers$HHKIDS == "0",]$DOCVIS
t.test(Pr3IVc1, Pr3IVc2, alternative = "two.sided")
```

```
##
##  Welch Two Sample t-test
##
## data:  Pr3IVc1 and Pr3IVc2
## t = -11.162, df = 11784, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.9323732 -0.6538202
## sample estimates:
## mean of x mean of y
##  2.681638  3.474734
```

Due to the low p-value, we reject the null hypothesis and conclude that the number of doctor visits a patient has over a 3 month period is greater for women with children than for women without children.