ECON403B Project 1 1.22.2018

Mohammed Ibraaz Syed, Minxuan Wang, Yating Zhang

January 22, 2018

GROUP MEMBERS: Mohammed Ibraaz Syed, Minxuan Wang, Yating Zhang

Setting working directory:

```
setwd("C:/Users/Ibraaz/Desktop/ECON403B Homework Folder/")
```

PROJECT 1 PART I: Airline Data

Getting the data:

```
data("USAirlines", package = "AER")
```

PROJECT 1 PART I (a)

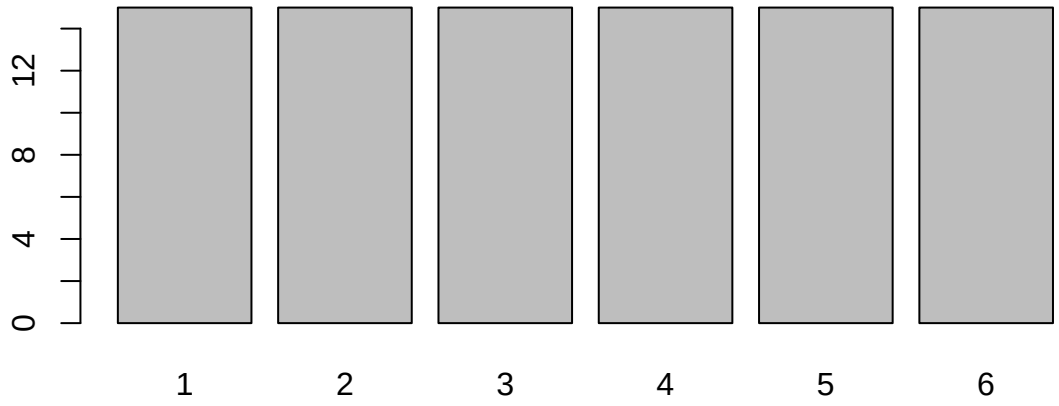
Exploring the data set:

```
str(USAirlines)
```

```
## 'data.frame':    90 obs. of  6 variables:
##  $ firm  : Factor w/ 6 levels "1","2","3","4",..: 1 1 1 1 1 1 1 1 1 1 ...
##   ..- attr(*, "contrasts")= num [1:6, 1:5] 1 0 0 0 0 0 0 1 0 0 ...
##   .. ..- attr(*, "dimnames")=List of 2
##   .. .. ..$ : chr  "1" "2" "3" "4" ...
##   .. .. ..$ : chr  "1" "2" "3" "4" ...
##  $ year  : Factor w/ 15 levels "1970","1971",..: 1 2 3 4 5 6 7 8 9 10 ...
##   ..- attr(*, "contrasts")= num [1:15, 1:14] 1 0 0 0 0 0 0 0 0 0 ...
##   .. ..- attr(*, "dimnames")=List of 2
##   .. .. ..$ : chr  "1970" "1971" "1972" "1973" ...
##   .. .. ..$ : chr  "1" "2" "3" "4" ...
##  $ output: num  0.953 0.987 1.092 1.176 1.16 ...
##  $ cost  : int  1140640 1215690 1309570 1511530 1676730 1823740 2022890 2314760 2639160 3247620 ...
##  $ price : int  106650 110307 110574 121974 196606 265609 263451 316411 384110 569251 ...
##  $ load  : num  0.534 0.532 0.548 0.541 0.591 ...
```

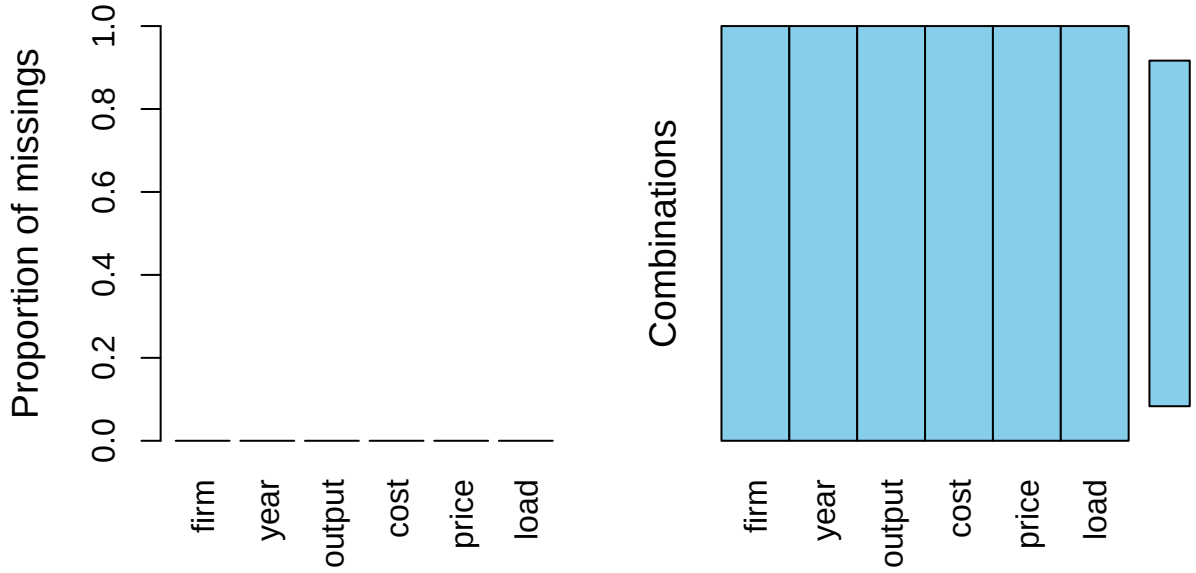
The above information about the data set indicates that there are only 6 firms. We verify that this is the case:

```
plot(USAirlines$firm)
```



We check for missing values:

```
aggr(USAirlines)
```



Since there are only 6 firms in the data set and we have no missing values, we conclude that we already have the complete set of observations.

PROJECT 1 PART I (b)

Summary statistics for variables in the data set:

```
describe(USAirlines)
```

```
##        vars  n       mean         sd    median   trimmed       mad
## firm*     1 90       3.50       1.72      3.50      3.50      2.22
## year*     2 90       8.00       4.34      8.00      8.00      5.93
## output    3 90       0.54       0.53      0.31      0.47      0.34
## cost      4 90 1122523.83 1192074.70 637001.00 889610.75 718743.72
## price     5 90  471683.01  329502.91 357433.50 451890.86 352077.47
## load      6 90       0.56       0.05      0.57      0.56      0.05
##               min        max      range  skew kurtosis        se
## firm*        1.00       6.00       5.00  0.00    -1.31      0.18
## year*        1.00      15.00      14.00  0.00    -1.25      0.46
## output       0.04       1.94       1.90  0.98    -0.35      0.06
## cost     68978.00 4748320.00 4679342.00  1.49     1.18 125655.71
## price   103795.00 1015610.00  911815.00  0.41    -1.49  34732.66
## load         0.43       0.68       0.24 -0.30    -0.36      0.01
```

PROJECT 1 PART I (c)

Defining variables

```
lnC <- log(USAirlines$cost)
lnQ <- log(USAirlines$output)
lnQsq <- lnQ^2
lnP <- log(USAirlines$price)
```

Estimating the cost equation

```
model <- lm(lnC ~ lnQ + lnQsq + lnP + load, data = USAirlines)
summary(model)
```

```
##
## Call:
## lm(formula = lnC ~ lnQ + lnQsq + lnP + load, data = USAirlines)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.24060 -0.06740 -0.01145  0.06233  0.32458
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  9.42058    0.23035  40.896  < 2e-16 ***
## lnQ          0.93543    0.02929  31.941  < 2e-16 ***
## lnQsq        0.02254    0.01122   2.009   0.0477 *
## lnP          0.45767    0.02004  22.838  < 2e-16 ***
## load        -1.53744    0.34232  -4.491 2.21e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1225 on 85 degrees of freedom
## Multiple R-squared:  0.9888, Adjusted R-squared:  0.9883
## F-statistic:  1880 on 4 and 85 DF,  p-value: < 2.2e-16
```

PROJECT 1 PART I (d)

Model with time effects only:

```
model.time <- plm(lnC ~ lnQ + lnQsq + lnP + load, data = USAirlines,
                  index = c("firm", "year"), model = "within", effect = "time")
summary(model.time)
```

```
## Oneway (time) effect Within Model
##
## Call:
## plm(formula = lnC ~ lnQ + lnQsq + lnP + load, data = USAirlines,
##     effect = "time", model = "within", index = c("firm", "year"))
##
## Balanced Panel: n=6, T=15, N=90
##
## Residuals :
##      Min.   1st Qu.    Median   3rd Qu.      Max.
## -0.217183 -0.047131 -0.011365  0.032726  0.307386
##
## Coefficients :
##        Estimate Std. Error t-value  Pr(>|t|)
## lnQ    0.917259   0.029957 30.6195 < 2.2e-16 ***
## lnQsq  0.021418   0.011180  1.9158 0.0594199 .
## lnP   -0.457158   0.357824 -1.2776 0.2055496
## load  -1.807037   0.441154 -4.0962 0.0001099 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    76.734
## Residual Sum of Squares: 1.0347
## R-Squared:      0.98652
## Adj. R-Squared: 0.9831
## F-statistic: 1298.6 on 4 and 71 DF, p-value: < 2.22e-16
```

Model with firm effect only:

```
model.firm <- plm(lnC ~ lnQ + lnQsq + lnP + load, data = USAirlines,
                  index = c("firm", "year"), model = "within", effect = "individual")
summary(model.firm)
```

```
## Oneway (individual) effect Within Model
##
## Call:
## plm(formula = lnC ~ lnQ + lnQsq + lnP + load, data = USAirlines,
##     effect = "individual", model = "within", index = c("firm",
##         "year"))
##
## Balanced Panel: n=6, T=15, N=90
##
## Residuals :
##      Min.   1st Qu.    Median   3rd Qu.      Max.
## -0.161208 -0.029347 -0.010894  0.041436  0.165034
##
## Coefficients :
##         Estimate Std. Error t-value  Pr(>|t|)
## lnQ    1.0452200  0.0547869 19.0779 < 2.2e-16 ***
## lnQsq  0.0261284  0.0096708  2.7018  0.008418 **
## lnP    0.3983273  0.0162683 24.4849 < 2.2e-16 ***
## load  -1.1037983  0.1946694 -5.6701 2.186e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    39.361
## Residual Sum of Squares: 0.26815
## R-Squared:      0.99319
## Adj. R-Squared: 0.99242
## F-statistic: 2915.7 on 4 and 80 DF, p-value: < 2.22e-16
```

Model with both firm and time effects:

```
model.timeANDfirm <- plm(lnC ~ lnQ + lnQsq + lnP + load, data = USAirlines,
                         index = c("firm", "year"), model = "within", effect = "twoways")
summary(model.timeANDfirm)
```

```
## Twoways effects Within Model
##
## Call:
## plm(formula = lnC ~ lnQ + lnQsq + lnP + load, data = USAirlines,
##     effect = "twoways", model = "within", index = c("firm", "year"))
##
## Balanced Panel: n=6, T=15, N=90
##
## Residuals :
##       Min.    1st Qu.     Median    3rd Qu.       Max.
## -0.1245865 -0.0273088 -0.0026827  0.0241649  0.1337247
##
## Coefficients :
##         Estimate Std. Error t-value  Pr(>|t|)
## lnQ    0.8866465  0.0628364 14.1104 < 2.2e-16 ***
## lnQsq  0.0126129  0.0098605  1.2791  0.205330
## lnP    0.1280783  0.1657643  0.7727  0.442486
## load  -0.8854826  0.2605115 -3.3990  0.001151 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    2.0542
## Residual Sum of Squares: 0.17257
## R-Squared:      0.91599
## Adj. R-Squared: 0.88672
## F-statistic: 179.913 on 4 and 66 DF, p-value: < 2.22e-16
```

Comparing the three models:

```
stargazer(model.time, model.firm, model.timeANDfirm,
title="(1) time effects only vs. (2) firm effect only vs. (3) both firm and time effects",
align = TRUE, header = FALSE, type = 'text', digits = 7, order=c("Constant"),
model.names = FALSE, omit.stat=c("f"), column.sep.width ="6pt", column.labels =
c("  Time Effects Only  ", "  Firm Effect Only  ", "  Both Firm and Time Effects  "))
```

```
##
## (1) time effects only vs. (2) firm effect only vs. (3) both firm and time effects
## ================================================================================
##                                    Dependent variable:
##               ------------------------------------------------------------
##                                           lnC
##               Time Effects Only Firm Effect Only Both Firm and Time Effects
##                      (1)              (2)                    (3)
## --------------------------------------------------------------------------------
## lnQ             0.9172593***      1.0452200***          0.8866465***
##                  (0.0299567)      (0.0547869)           (0.0628364)
##
## lnQsq            0.0214178*       0.0261284***           0.0126129
##                  (0.0111796)      (0.0096708)           (0.0098605)
##
## lnP              -0.4571584       0.3983273***           0.1280783
##                  (0.3578238)      (0.0162683)           (0.1657643)
##
## load            -1.8070370***    -1.1037980***         -0.8854826***
##                  (0.4411542)      (0.1946694)           (0.2605115)
##
## --------------------------------------------------------------------------------
## Observations        90               90                    90
## R2                0.9865157        0.9931873             0.9159934
## Adjusted R2       0.9830972        0.9924209             0.8867183
## ================================================================================
## Note:                                            *p<0.1; **p<0.05; ***p<0.01
```

We note that the log of output is statistically significant for all models, and so is the load. However, the square of the log of output is only significant at the 1% level for the model with firm effects only, and is significant at the 10% level for the model with time effects only. The square of the log of output is not significant at even the 10% level for the model with both firm and time effects. Also, the log of price is only significant at the 1% level for the model with firm effects only, and is not significant at even the 10% level for the other two models. The Adjusted R squared value of the model with firm effects is the highest, followed by the model with time effects only. The model with both firm and time effects has the lowest Adjusted R squared value.

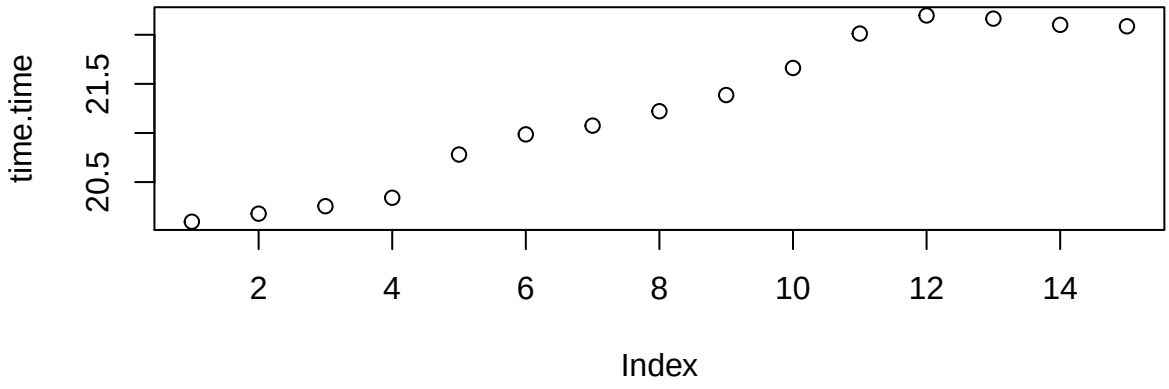
PROJECT 1 PART I (e)

Extracting the fixed effects estimates:

```
time.time <- fixef(model.time)
time.both <- fixef(model.timeANDfirm,effect="time")
```

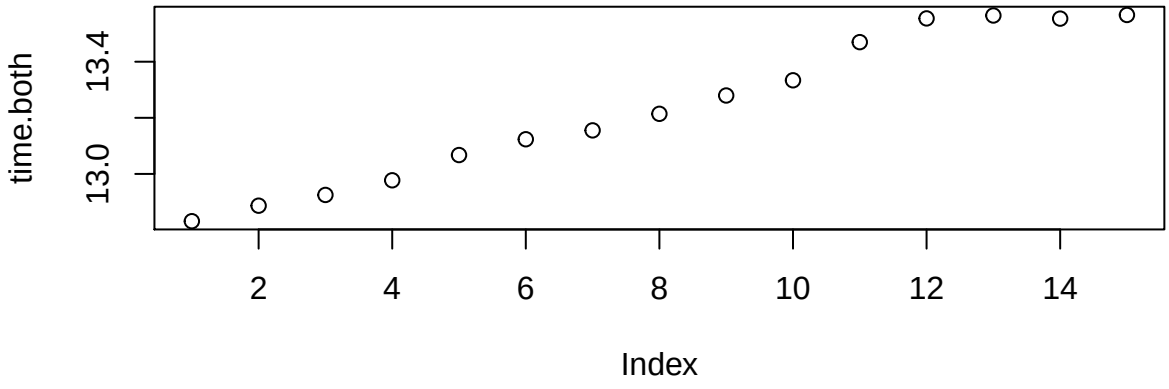
Estimated yearly effect of model with time effects only:

```
plot(time.time)
```



Estimated yearly effect of model with both time and firm effects:

```
plot(time.both)
```



Observe that the yearly effect of the model with only time effects is much larger than that of the model with both time and firm effects.

PROJECT 1 PART I (f)

We note that using the default estimator of the transformation parameter leads to the estimated variance of the time effect being negative, and so we use the "amemiya" estimator from "The estimation of the variances in a variance–components model" (Amemiya, 1971):

```
model.random <- plm(lnC ~ lnQ + lnQsq + lnP + load, data = USAirlines,
index = c("firm", "year"), model = "random", effect = "twoways",
random.method = "amemiya")
summary(model.random)
```

```
## Twoways effects Random Effect Model
##    (Amemiya's transformation)
##
## Call:
## plm(formula = lnC ~ lnQ + lnQsq + lnP + load, data = USAirlines,
##     effect = "twoways", model = "random", random.method = "amemiya",
##     index = c("firm", "year"))
##
## Balanced Panel: n=6, T=15, N=90
##
## Effects:
##                    var  std.dev share
## idiosyncratic 0.002465 0.049652 0.029
## individual    0.013192 0.114856 0.154
## time          0.069785 0.264169 0.817
## theta  : 0.8891 (id) 0.9235 (time) 0.8757 (total)
##
## Residuals :
##       Min.    1st Qu.     Median    3rd Qu.       Max.
## -0.1155769 -0.0307879 -0.0062083  0.0268412  0.1599713
##
## Coefficients :
##               Estimate Std. Error t-value  Pr(>|t|)
## (Intercept) 10.1993413  0.9299317 10.9678 < 2.2e-16 ***
## lnQ          0.8959663  0.0428955 20.8872 < 2.2e-16 ***
## lnQsq        0.0116532  0.0077606  1.5016 0.1369104
## lnP          0.3678163  0.0723244  5.0856 2.156e-06 ***
## load        -0.9100796  0.2393795 -3.8018 0.0002695 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    3.1916
## Residual Sum of Squares: 0.19328
## R-Squared:      0.93944
## Adj. R-Squared: 0.93659
## F-statistic: 329.638 on 4 and 85 DF, p-value: < 2.22e-16
```

PROJECT 1 PART I (g)

Conducting the Hausman Test:

```
phtest(model.timeANDfirm, model.random)
```

```
##
##  Hausman Test
##
## data:  lnC ~ lnQ + lnQsq + lnP + load
## chisq = 2.6984, df = 4, p-value = 0.6095
## alternative hypothesis: one model is inconsistent
```

Due to the high p-value, we fail to reject the null hypothesis, which implies that both models are consistent. If the fixed effect model was appropriate, the random effects estimator would be inconsistent, and so the fixed effects model would've been better. Since neither model is inconsistent, this implies that the random effects model is both consistent and efficient. **Thus, the random effects model is a much more suitable choice for the data at hand.**

PROJECT 1 PART II: Wage Equation

Getting the data:

```
data<-read.csv("wage.csv")
```

PROJECT 1 PART II (a)

Fitting a regular OLS model to the data:

```
Pr1IIa <- lm(LWAGE ~ EXPER + WKS + OCC + IND + SOUTH + SMSA + MS +
FEM + UNION + ED + BLK + YEAR + ID, data = data); summary(Pr1IIa)
```

```
##
## Call:
## lm(formula = LWAGE ~ EXPER + WKS + OCC + IND + SOUTH + SMSA +
##     MS + FEM + UNION + ED + BLK + YEAR + ID, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.0386 -0.1969  0.0061  0.2042  1.9928
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.219e+00  6.276e-02  83.154  < 2e-16 ***
## EXPER        6.440e-03  4.776e-04  13.484  < 2e-16 ***
## WKS          4.693e-03  9.562e-04   4.908 9.56e-07 ***
## OCC         -1.374e-01  1.299e-02 -10.581  < 2e-16 ***
## IND          5.405e-02  1.045e-02   5.172 2.43e-07 ***
## SOUTH       -5.866e-02  1.109e-02  -5.292 1.27e-07 ***
## SMSA         1.694e-01  1.075e-02  15.753  < 2e-16 ***
## MS           1.063e-01  1.824e-02   5.826 6.10e-09 ***
## FEM         -3.340e-01  2.221e-02 -15.036  < 2e-16 ***
## UNION        9.562e-02  1.132e-02   8.446  < 2e-16 ***
## ED           5.392e-02  2.314e-03  23.303  < 2e-16 ***
## BLK         -1.601e-01  1.951e-02  -8.207 3.00e-16 ***
## YEAR         9.058e-02  2.445e-03  37.043  < 2e-16 ***
## ID          -2.076e-04  2.861e-05  -7.255 4.78e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3092 on 4151 degrees of freedom
## Multiple R-squared:  0.5526, Adjusted R-squared:  0.5512
## F-statistic: 394.4 on 13 and 4151 DF,  p-value: < 2.2e-16
```

Conducting the Breusch-Pagan test against heteroskedasticity:

```
bptest(Pr1IIa)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  Pr1IIa
## BP = 142.56, df = 13, p-value < 2.2e-16
```

Due to the low p-value, we reject the null hypothesis and conclude that there are signs of heteroskedasticity.

PROJECT 1 PART II (b)

White standard errors:

```
coeftest(Pr1IIa,vcov=hccm(Pr1IIa))
```

```
##
## t test of coefficients:
##
##                Estimate  Std. Error  t value  Pr(>|t|)
## (Intercept)  5.2187e+00  6.7943e-02  76.8102 < 2.2e-16 ***
## EXPER        6.4400e-03  5.2947e-04  12.1631 < 2.2e-16 ***
## WKS          4.6929e-03  1.0474e-03   4.4803 7.654e-06 ***
## OCC         -1.3741e-01  1.3163e-02 -10.4391 < 2.2e-16 ***
## IND          5.4050e-02  1.0569e-02   5.1143 3.292e-07 ***
## SOUTH       -5.8664e-02  1.1556e-02  -5.0766 4.012e-07 ***
## SMSA         1.6942e-01  1.0997e-02  15.4053 < 2.2e-16 ***
## MS           1.0626e-01  1.9968e-02   5.3215 1.083e-07 ***
## FEM         -3.3396e-01  2.1714e-02 -15.3800 < 2.2e-16 ***
## UNION        9.5622e-02  1.0868e-02   8.7982 < 2.2e-16 ***
## ED           5.3917e-02  2.4047e-03  22.4219 < 2.2e-16 ***
## BLK         -1.6009e-01  1.9221e-02  -8.3289 < 2.2e-16 ***
## YEAR         9.0576e-02  2.3812e-03  38.0374 < 2.2e-16 ***
## ID          -2.0756e-04  2.8077e-05  -7.3928 1.730e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Double checking White standard errors:

```
coeftest(Pr1IIa, vcovHC)
```

```
##
## t test of coefficients:
##
##                Estimate  Std. Error  t value  Pr(>|t|)
## (Intercept)  5.2187e+00  6.7943e-02  76.8102 < 2.2e-16 ***
## EXPER        6.4400e-03  5.2947e-04  12.1631 < 2.2e-16 ***
## WKS          4.6929e-03  1.0474e-03   4.4803 7.654e-06 ***
## OCC         -1.3741e-01  1.3163e-02 -10.4391 < 2.2e-16 ***
## IND          5.4050e-02  1.0569e-02   5.1143 3.292e-07 ***
## SOUTH       -5.8664e-02  1.1556e-02  -5.0766 4.012e-07 ***
## SMSA         1.6942e-01  1.0997e-02  15.4053 < 2.2e-16 ***
## MS           1.0626e-01  1.9968e-02   5.3215 1.083e-07 ***
## FEM         -3.3396e-01  2.1714e-02 -15.3800 < 2.2e-16 ***
## UNION        9.5622e-02  1.0868e-02   8.7982 < 2.2e-16 ***
## ED           5.3917e-02  2.4047e-03  22.4219 < 2.2e-16 ***
## BLK         -1.6009e-01  1.9221e-02  -8.3289 < 2.2e-16 ***
## YEAR         9.0576e-02  2.3812e-03  38.0374 < 2.2e-16 ***
## ID          -2.0756e-04  2.8077e-05  -7.3928 1.730e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Robust panel standard errors:

We note that the vcovPL() function from the sandwich package returns the Clustered Covariance Matrix Estimation for panel data:

```
coeftest(Pr1IIa, vcovPL(Pr1IIa, cluster = data[, c("ID", "YEAR")]))
```

```
##
## t test of coefficients:
##
##                Estimate  Std. Error  t value  Pr(>|t|)
## (Intercept)  5.2187e+00  6.0467e-02  86.3075 < 2.2e-16 ***
## EXPER        6.4400e-03  7.0764e-04   9.1008 < 2.2e-16 ***
## WKS          4.6929e-03  9.8223e-04   4.7778 1.834e-06 ***
## OCC         -1.3741e-01  5.9867e-03 -22.9528 < 2.2e-16 ***
## IND          5.4050e-02  1.4275e-02   3.7864  0.000155 ***
## SOUTH       -5.8664e-02  1.2839e-03 -45.6926 < 2.2e-16 ***
## SMSA         1.6942e-01  4.4710e-03  37.8928 < 2.2e-16 ***
## MS           1.0626e-01  1.2181e-02   8.7232 < 2.2e-16 ***
## FEM         -3.3396e-01  2.1040e-02 -15.8722 < 2.2e-16 ***
## UNION        9.5622e-02  1.0322e-02   9.2639 < 2.2e-16 ***
## ED           5.3917e-02  2.2704e-03  23.7477 < 2.2e-16 ***
## BLK         -1.6009e-01  7.8535e-03 -20.3842 < 2.2e-16 ***
## YEAR         9.0576e-02  1.7756e-03  51.0118 < 2.2e-16 ***
## ID          -2.0756e-04  1.9394e-05 -10.7028 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
coeftest(Pr1IIa, vcovPL(Pr1IIa, cluster = data[, c("ID")]))
```

```
##
## t test of coefficients:
##
##                Estimate  Std. Error  t value  Pr(>|t|)
## (Intercept)  5.2187e+00  6.0467e-02  86.3075 < 2.2e-16 ***
## EXPER        6.4400e-03  7.0764e-04   9.1008 < 2.2e-16 ***
## WKS          4.6929e-03  9.8223e-04   4.7778 1.834e-06 ***
## OCC         -1.3741e-01  5.9867e-03 -22.9528 < 2.2e-16 ***
## IND          5.4050e-02  1.4275e-02   3.7864  0.000155 ***
## SOUTH       -5.8664e-02  1.2839e-03 -45.6926 < 2.2e-16 ***
## SMSA         1.6942e-01  4.4710e-03  37.8928 < 2.2e-16 ***
## MS           1.0626e-01  1.2181e-02   8.7232 < 2.2e-16 ***
## FEM         -3.3396e-01  2.1040e-02 -15.8722 < 2.2e-16 ***
## UNION        9.5622e-02  1.0322e-02   9.2639 < 2.2e-16 ***
## ED           5.3917e-02  2.2704e-03  23.7477 < 2.2e-16 ***
## BLK         -1.6009e-01  7.8535e-03 -20.3842 < 2.2e-16 ***
## YEAR         9.0576e-02  1.7756e-03  51.0118 < 2.2e-16 ***
## ID          -2.0756e-04  1.9394e-05 -10.7028 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Comparing White standard errors and robust panel standard errors:

```
as.data.frame(sqrt(diag(hccm(Pr1IIa)))
            - sqrt(diag(vcovPL(Pr1IIa, cluster = data[, c("ID")]))))
```

```
##             sqrt(diag(hccm(Pr1IIa))) - sqrt(diag(vcovPL(Pr1IIa, cluster = data[, c("ID")])))
## (Intercept)                                                                      7.476553e-03
## EXPER                                                                           -1.781632e-04
## WKS                                                                              6.521091e-05
## OCC                                                                              7.176405e-03
## IND                                                                             -3.706442e-03
## SOUTH                                                                            1.027195e-02
## SMSA                                                                             6.526338e-03
## MS                                                                               7.786638e-03
## FEM                                                                              6.734551e-04
## UNION                                                                            5.463843e-04
## ED                                                                               1.342454e-04
## BLK                                                                              1.136723e-02
## YEAR                                                                             6.056482e-04
## ID                                                                               8.683259e-06
```

Due to the robust panel standard errors being generally smaller than the White standard errors, we can estimate that there is negative correlation within the residuals when we cluster the data. This implies that our model is possibly misspecified and we need better predictors. If we had better predictors, the negative correlation would decrease, leading to a better model.

NO IDEA WHAT'S UP WITH THIS:

```
coeftest(Pr1IIa, vcovPL(Pr1IIa))
```

```
##
## t test of coefficients:
##
##                Estimate  Std. Error t value  Pr(>|t|)
## (Intercept)  5.2187e+00  1.1372e-01 45.8901 < 2.2e-16 ***
## EXPER        6.4400e-03  1.0795e-03  5.9657 2.639e-09 ***
## WKS          4.6929e-03  1.4769e-03  3.1776 0.0014961 **
## OCC         -1.3741e-01  2.3571e-02 -5.8298 5.973e-09 ***
## IND          5.4050e-02  2.0721e-02  2.6085 0.0091268 **
## SOUTH       -5.8664e-02  2.3323e-02 -2.5152 0.0119329 *
## SMSA         1.6942e-01  2.1546e-02  7.8629 4.749e-15 ***
## MS           1.0626e-01  3.7571e-02  2.8282 0.0047031 **
## FEM         -3.3396e-01  4.1388e-02 -8.0688 9.207e-16 ***
## UNION        9.5622e-02  2.0754e-02  4.6074 4.200e-06 ***
## ED           5.3917e-02  4.7584e-03 11.3311 < 2.2e-16 ***
## BLK         -1.6009e-01  4.0320e-02 -3.9704 7.296e-05 ***
## YEAR         9.0576e-02  2.8065e-03 32.2733 < 2.2e-16 ***
## ID          -2.0756e-04  5.8104e-05 -3.5723 0.0003579 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

PROJECT 1 PART II (c)

Model with individual effect only:

```
model.individual <- plm(LWAGE ~ EXPER + WKS + OCC + IND + SOUTH +
                        SMSA + MS + FEM + UNION + ED + BLK,
                        data = data, index = c("ID", "YEAR"),
                        model = "within", effect = "individual")
summary(model.individual)
```

```
## Oneway (individual) effect Within Model
##
## Call:
## plm(formula = LWAGE ~ EXPER + WKS + OCC + IND + SOUTH + SMSA +
##     MS + FEM + UNION + ED + BLK, data = data, effect = "individual",
##     model = "within", index = c("ID", "YEAR"))
##
## Balanced Panel: n=595, T=7, N=4165
##
## Residuals :
##       Min.    1st Qu.     Median    3rd Qu.       Max.
## -1.7984511 -0.0535263  0.0042525  0.0628480  1.9452352
##
## Coefficients :
##          Estimate  Std. Error t-value Pr(>|t|)
## EXPER  0.09657698  0.00119085 81.0992  < 2e-16 ***
## WKS    0.00114223  0.00060316  1.8937  0.05834 .
## OCC   -0.02486403  0.01388776 -1.7904  0.07348 .
## IND    0.02075656  0.01556962  1.3331  0.18257
## SOUTH -0.00319792  0.03457562 -0.0925  0.92631
## SMSA  -0.04372702  0.01958444 -2.2327  0.02563 *
## MS    -0.03025961  0.01913663 -1.5812  0.11391
## UNION  0.03415826  0.01504220  2.2708  0.02322 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    240.65
## Residual Sum of Squares: 83.624
## R-Squared:      0.65251
## Adj. R-Squared: 0.59378
## F-statistic: 836.082 on 8 and 3562 DF, p-value: < 2.22e-16
```

Model with individual and time effects:

```
model.individualANDtime <- plm(LWAGE ~ EXPER + WKS + OCC + IND + SOUTH +
                               SMSA + MS + FEM + UNION + ED + BLK,
                               data = data, index = c("ID", "YEAR"),
                               model = "within", effect = "twoways")
summary(model.individualANDtime)
```

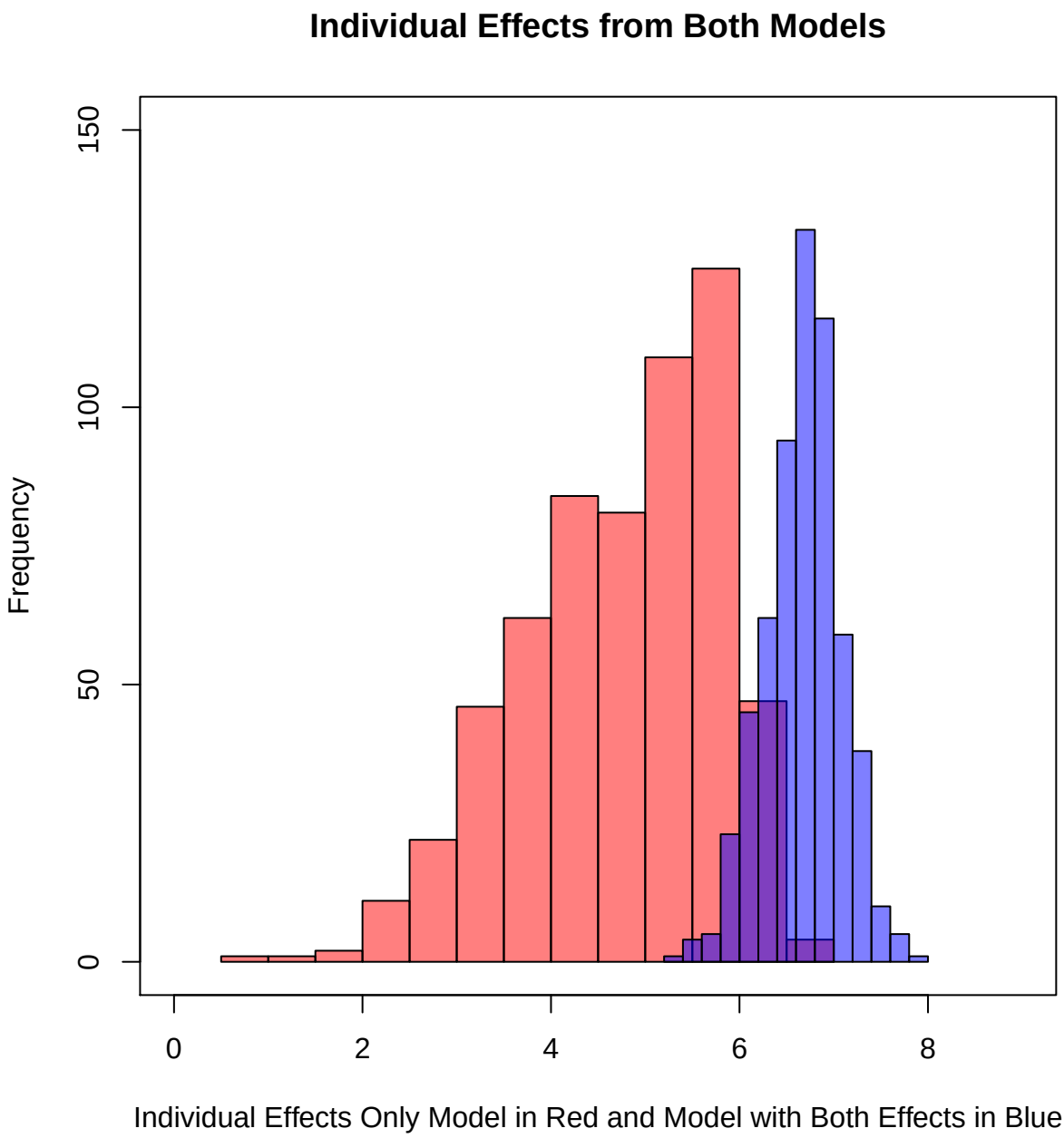
```
## Twoways effects Within Model
##
## Call:
## plm(formula = LWAGE ~ EXPER + WKS + OCC + IND + SOUTH + SMSA +
##     MS + FEM + UNION + ED + BLK, data = data, effect = "twoways",
##     model = "within", index = c("ID", "YEAR"))
##
## Balanced Panel: n=595, T=7, N=4165
##
## Residuals :
##       Min.    1st Qu.     Median    3rd Qu.       Max.
## -1.8149692 -0.0549163  0.0038679  0.0629619  1.9263486
##
## Coefficients :
##          Estimate  Std. Error t-value Pr(>|t|)
## WKS    0.00094862  0.00060235  1.5749  0.11538
## OCC   -0.02213145  0.01384344 -1.5987  0.10998
## IND    0.02235811  0.01551107  1.4414  0.14955
## SOUTH  0.00228936  0.03443928  0.0665  0.94700
## SMSA  -0.04318164  0.01951543 -2.2127  0.02698 *
## MS    -0.02899049  0.01905818 -1.5212  0.12831
## UNION  0.03067453  0.01498978  2.0464  0.04079 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    83.178
## Residual Sum of Squares: 82.751
## R-Squared:      0.0051317
## Adj. R-Squared: -0.16464
## F-statistic: 2.62109 on 7 and 3557 DF, p-value: 0.010634
```

Extracting the Fixed Effects estimates:

```
individual.individual <- fixef(model.individual)
individual.individualANDtime <- fixef(model.individualANDtime, effect="individual")
```

Plotting the Individual Effects from both models:

```
hist(individual.individual, col=rgb(1,0,0,0.5), xlim = c(0,9), ylim = c(0,150),
main = "Individual Effects from Both Models",
xlab = "Individual Effects Only Model in Red and Model with Both Effects in Blue")
hist(individual.individualANDtime, col=rgb(0,0,1,0.5), add=T)
box()
```



We observe that the model with both individual and time effects has larger individual effects than the model with only individual effects. Also, the individual effects of the model with both effects has less variance.

PROJECT 1 PART II (d)

Fitting a random effect model:

```
Pr1IId.model.random <- plm(LWAGE ~ EXPER + WKS + OCC + IND + SOUTH +
                           SMSA + MS + FEM + UNION + ED + BLK,
                           data = data, index = c("ID", "YEAR"),
                           model = "random")
summary(Pr1IId.model.random)
```

```
## Oneway (individual) effect Random Effect Model
##    (Swamy-Arora's transformation)
##
## Call:
## plm(formula = LWAGE ~ EXPER + WKS + OCC + IND + SOUTH + SMSA +
##     MS + FEM + UNION + ED + BLK, data = data, model = "random",
##     index = c("ID", "YEAR"))
##
## Balanced Panel: n=595, T=7, N=4165
##
## Effects:
##                   var std.dev share
## idiosyncratic 0.02350 0.15329 0.245
## individual    0.07242 0.26910 0.755
## theta:  0.7895
##
## Residuals :
##        Min.     1st Qu.      Median     3rd Qu.        Max.
## -2.04950451 -0.11860965 -0.00020286  0.12339957  2.07663474
##
## Coefficients :
##                Estimate  Std. Error t-value  Pr(>|t|)
## (Intercept)  4.45782097  0.09870197 45.1645 < 2.2e-16 ***
## EXPER        0.04856641  0.00105856 45.8797 < 2.2e-16 ***
## WKS          0.00163673  0.00078407  2.0875 0.0369059 *
## OCC         -0.05649044  0.01692945 -3.3368 0.0008549 ***
## IND          0.00724880  0.01759477  0.4120 0.6803710
## SOUTH       -0.01348744  0.02719083 -0.4960 0.6199001
## SMSA        -0.02239776  0.02040357 -1.0977 0.2723828
## MS          -0.07348646  0.02340646 -3.1396 0.0017038 **
## FEM         -0.33352407  0.05276998 -6.3203 2.885e-10 ***
## UNION        0.06821790  0.01738599  3.9237 8.860e-05 ***
## ED           0.10201525  0.00591165 17.2566 < 2.2e-16 ***
## BLK         -0.21589065  0.05976067 -3.6126 0.0003067 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    269.28
## Residual Sum of Squares: 169.92
## R-Squared:      0.369
## Adj. R-Squared: 0.36733
## F-statistic: 220.781 on 11 and 4153 DF, p-value: < 2.22e-16
```

Conducting the Hausman test:

```r
phtest(model.individualANDtime, Pr1IId.model.random)
```

```
##
##  Hausman Test
##
## data:  LWAGE ~ EXPER + WKS + OCC + IND + SOUTH + SMSA + MS + FEM + UNION +  ...
## chisq = 50.847, df = 7, p-value = 9.846e-09
## alternative hypothesis: one model is inconsistent
```

Due to the very low p-value, we reject the null hypothesis, which implies that the random effect model is inconsistent. **Thus the fixed effect model is more suitable.**

PROJECT 1 PART III: US Consumption

Getting the data:

```
data("USConsump1993", package = "AER")
```

PROJECT 1 PART III (a)

Calculating Investment in each period

```
# Defining Income
Pr1Income <- USConsump1993[,1]
# Defining Expenditure
Pr1Expenditure <- USConsump1993[,2]
# Calculating Investment as difference between Income and Expenditure
Pr1Investment <- USConsump1993[,1] - USConsump1993[,2]
```

Project 1 PART III (b)

Calculating the summary statistics of each variable:

```
describe(USConsump1993)
```

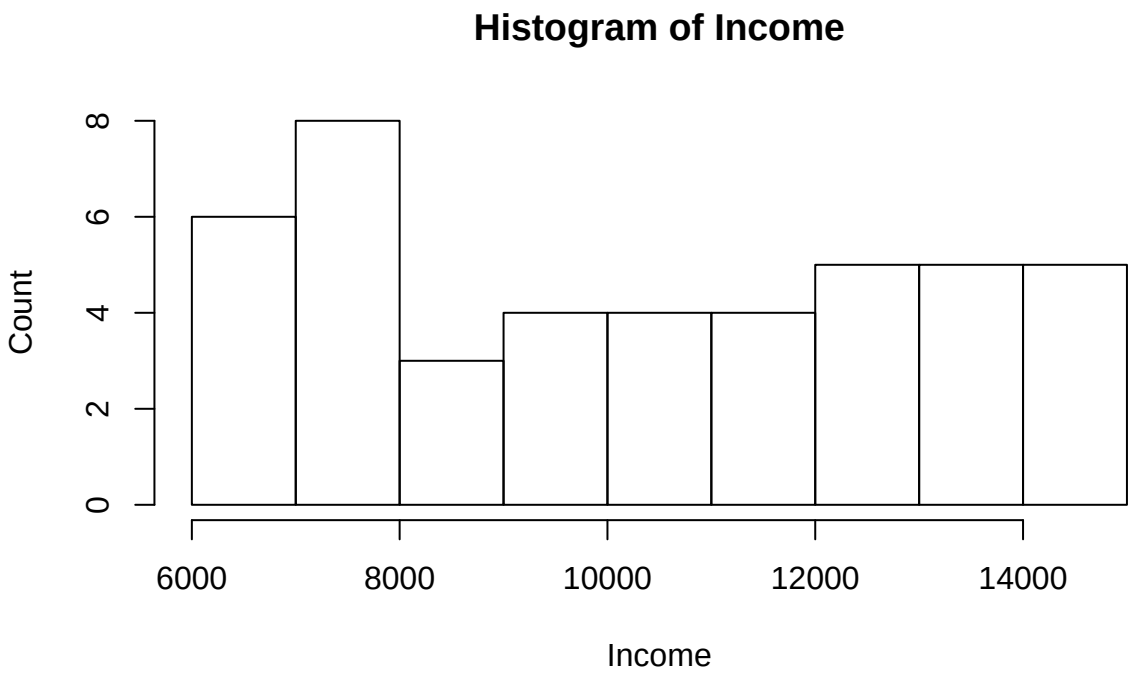
```
##             vars  n     mean      sd  median  trimmed     mad  min   max
## income         1 44 10174.86 2708.52 10262.5 10144.17 4037.12 6284 14341
## expenditure    2 44  9250.55 2484.62  9223.5  9187.97 3473.73 5820 13391
##             range skew kurtosis     se
## income       8057 0.05    -1.49 408.32
## expenditure  7571 0.16    -1.39 374.57
```

Estimating Underlying Distributions

Histogram and Density Curve for income

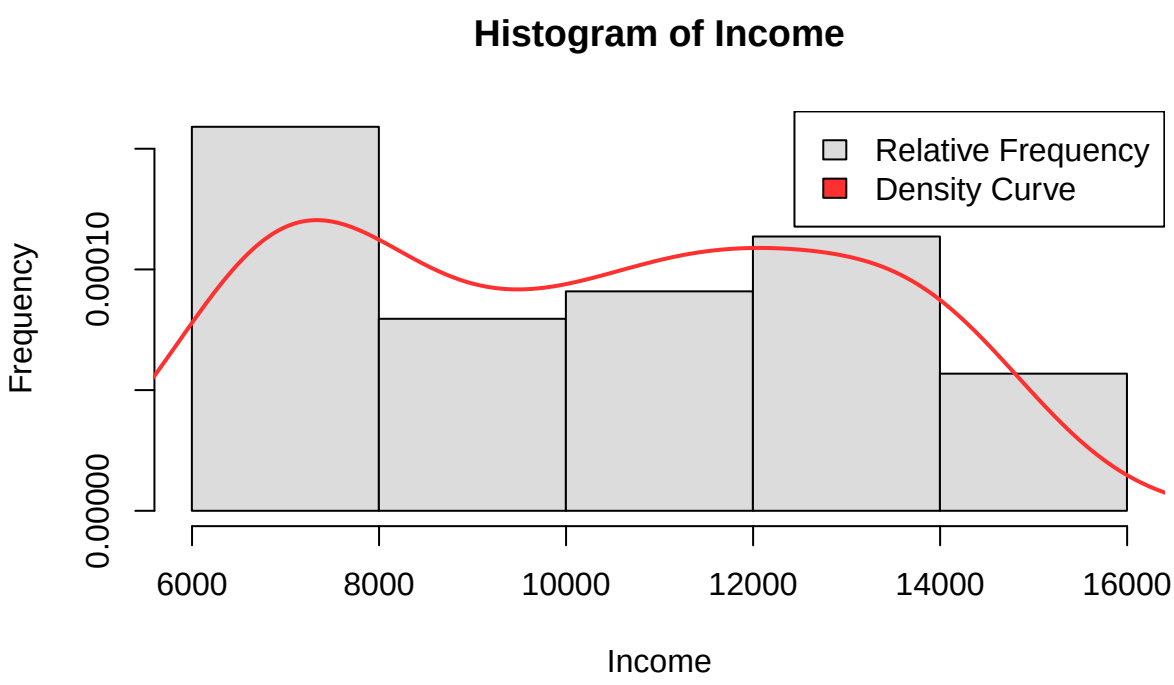
Histogram

```
hist(USConsump1993[,1], xlab= "Income", ylab= "Count",
     main= "Histogram of Income")
```



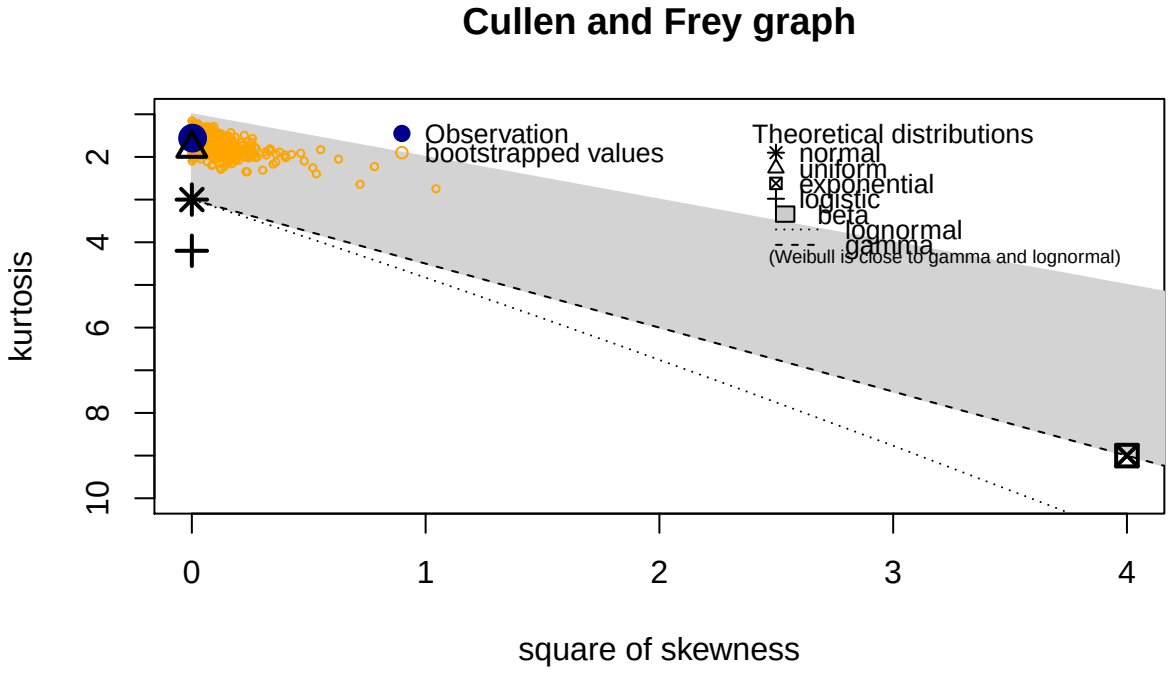
Histogram and Density Curve

```
truehist(USConsump1993[,1],col="gainsboro", ylab="Frequency",
         xlab= "Income",
         main= "Histogram of Income")
lines(density((USConsump1993[,1])), lwd=2,col="firebrick1")
legend("topright", c("Relative Frequency", "Density Curve"),
       fill=c("gainsboro", "firebrick1"))
```



Note that from the summary statistics, we know that both Income and Expenditure have 44 observations. We plot the Cullen and Frey graph to give us an idea of what distributions to try out:

```
descdist(USConsump1993[,1][1:44], boot = 1000)
```



```
## summary statistics
## ------
## min:  6284   max:  14341
## median:  10262.5
## mean:  10174.86
## estimated sd:  2708.519
## estimated skewness:  0.05614605
## estimated kurtosis:  1.560164
```

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

Due to the square of skewness values being close to zero, we will also test normal and logistic distributions just to make sure that those are indeed not good fits.

We proceed to fit a uniform distribution, a normal distribution, and a logistic distribution.

Testing fits for distributions

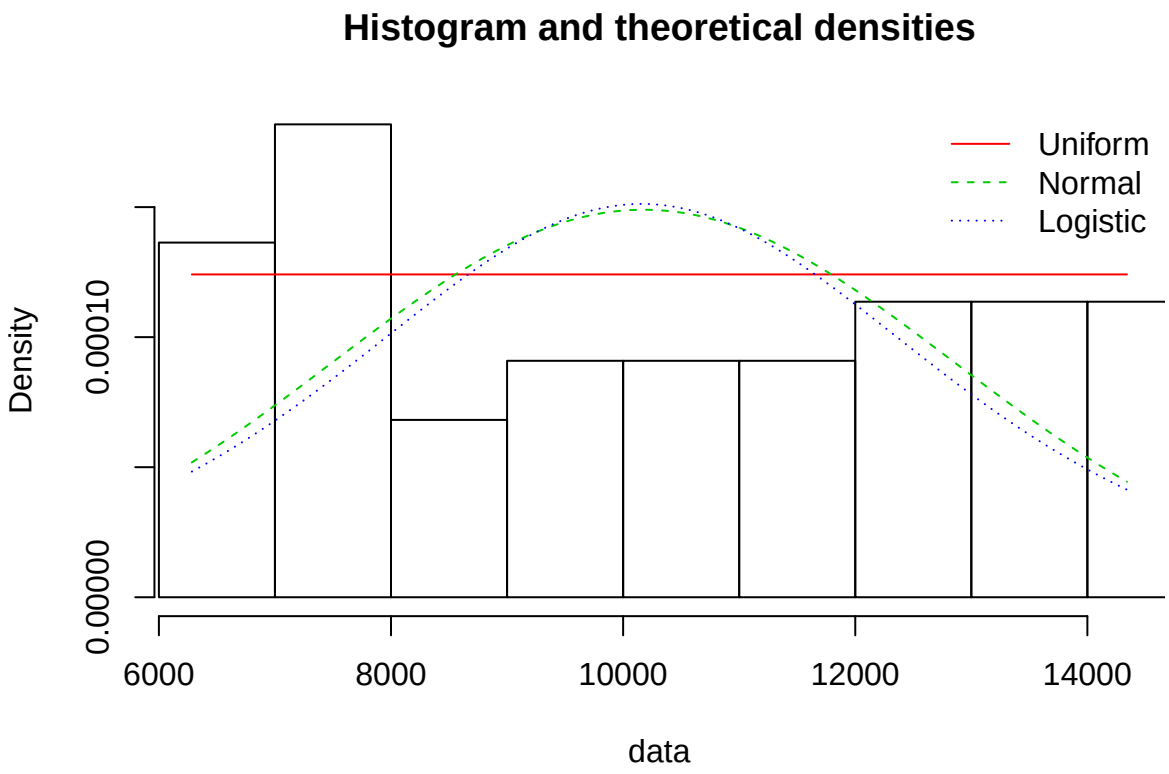
```
# Testing for a uniform distribution
incomeunif <- fitdist(USConsump1993[,1][1:44], "unif")
# Testing for a normal distribution
incomenorm <- fitdist(USConsump1993[,1][1:44], "norm")
# Testing fit for a logistic distribution
incomelogis <- fitdist(USConsump1993[,1][1:44], "logis")
```

Setting Legend

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

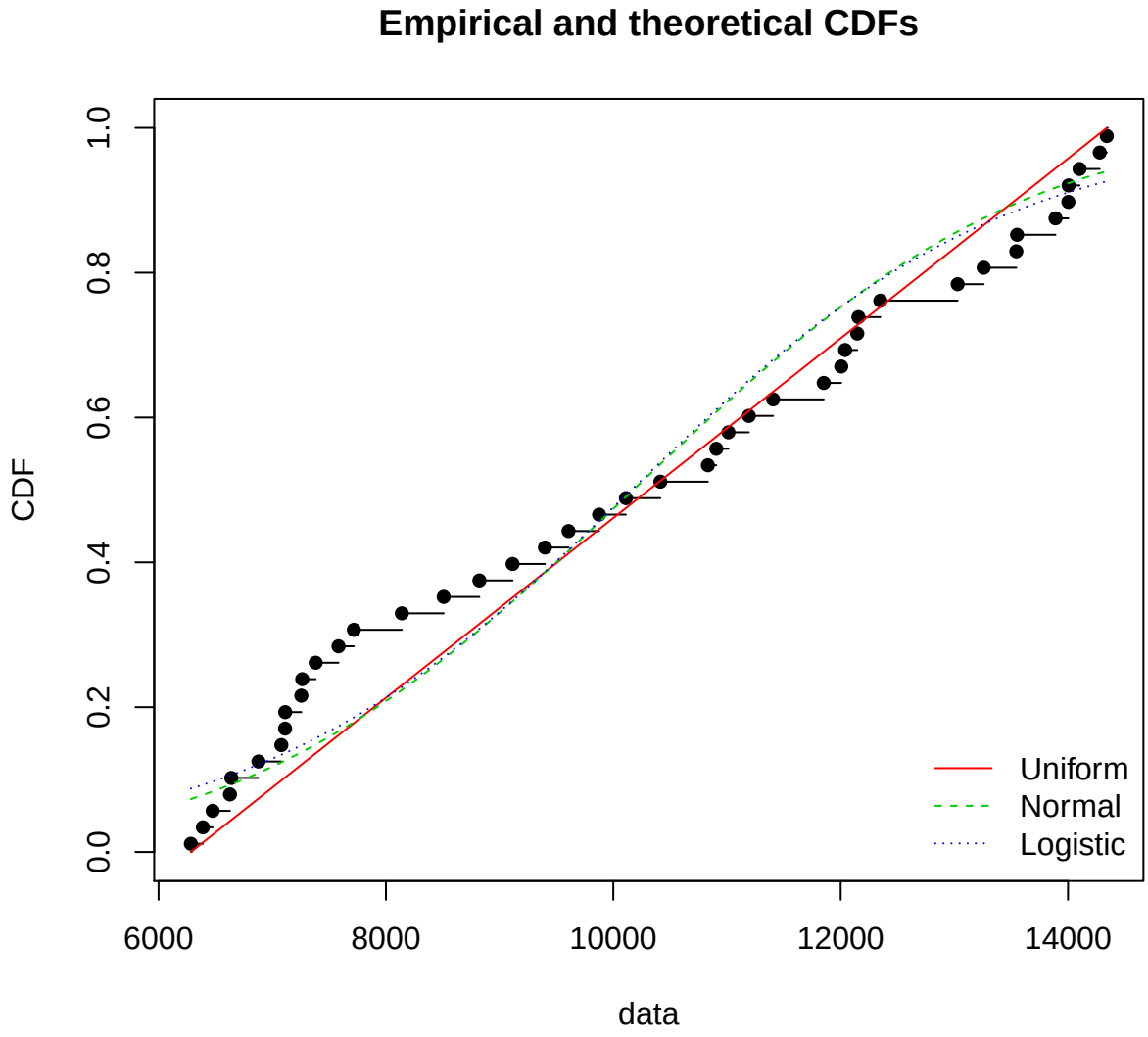
Comparing Histogram and Theoretical Densities

```
denscomp(list(incomeunif, incomenorm, incomelogis), legendtext = plot.legend)
```



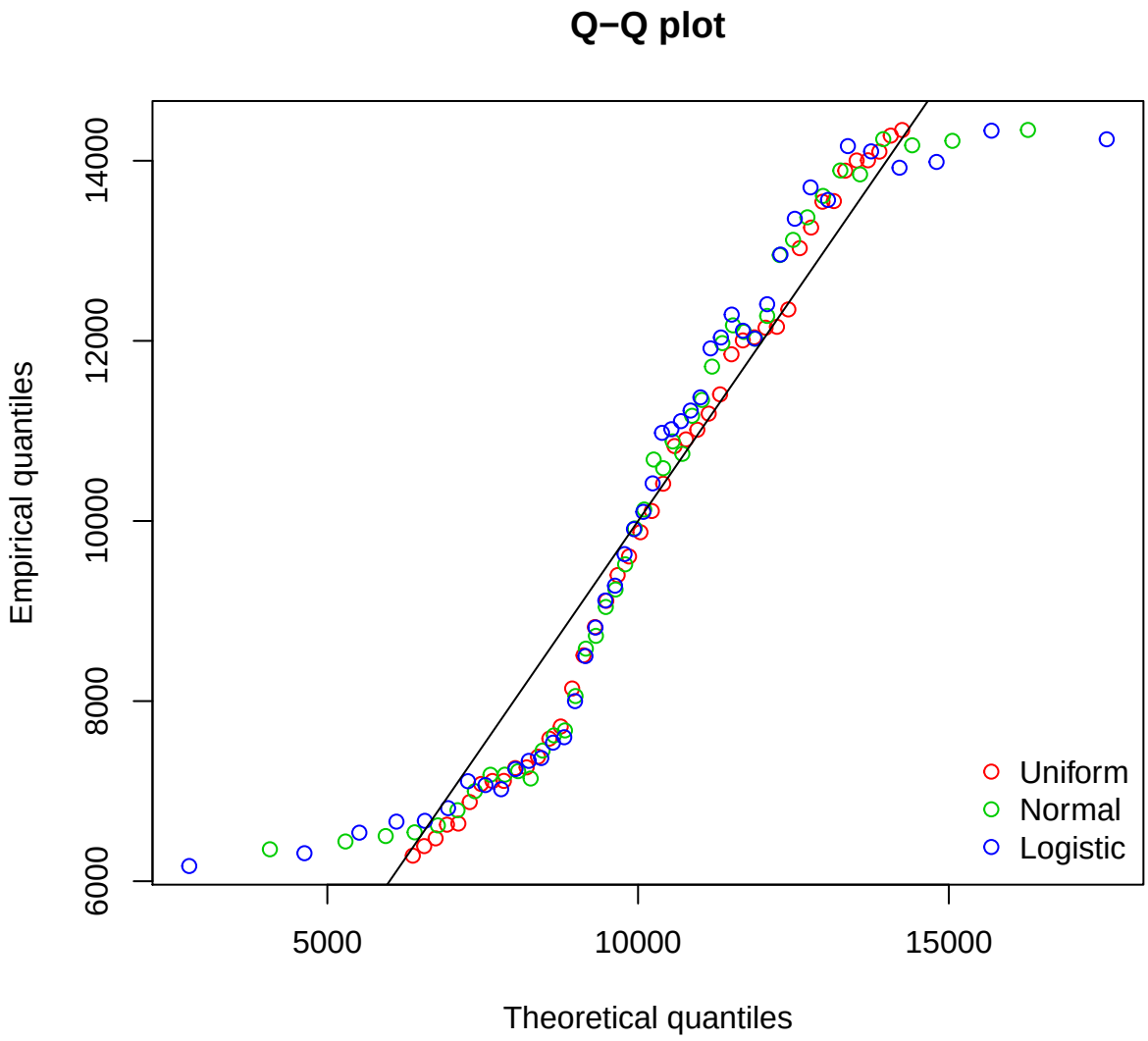
Observe that the uniform distribution seems to be a better fir than the normal and logistic distributions based on the theoretical densities of the distributions.

```
cdfcomp(list(incomeunif, incomenorm, incomelogis), legendtext = plot.legend)
```



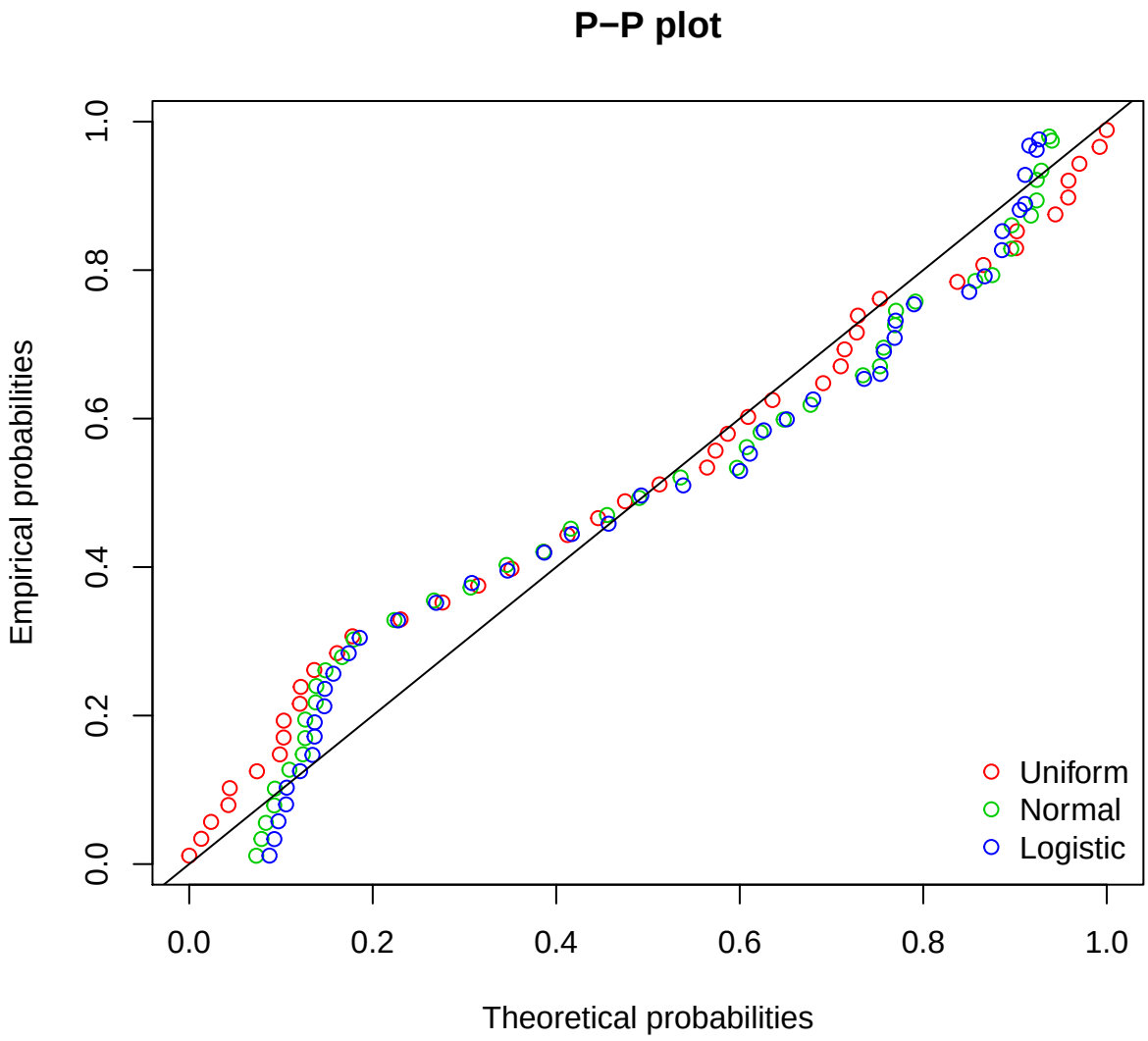
Observe that the uniform distribution appears much more appropriate than the other two distributions.

```
qqcomp(list(incomeunif, incomenorm, incomelogis), legendtext = plot.legend)
```



As far as the empirical quantiles compared to the theoretical quantiles, the uniform distribution is much better than the normal and logistic distributions.

```
ppcomp(list(incomeunif, incomenorm, incomelogis), legendtext = plot.legend)
```



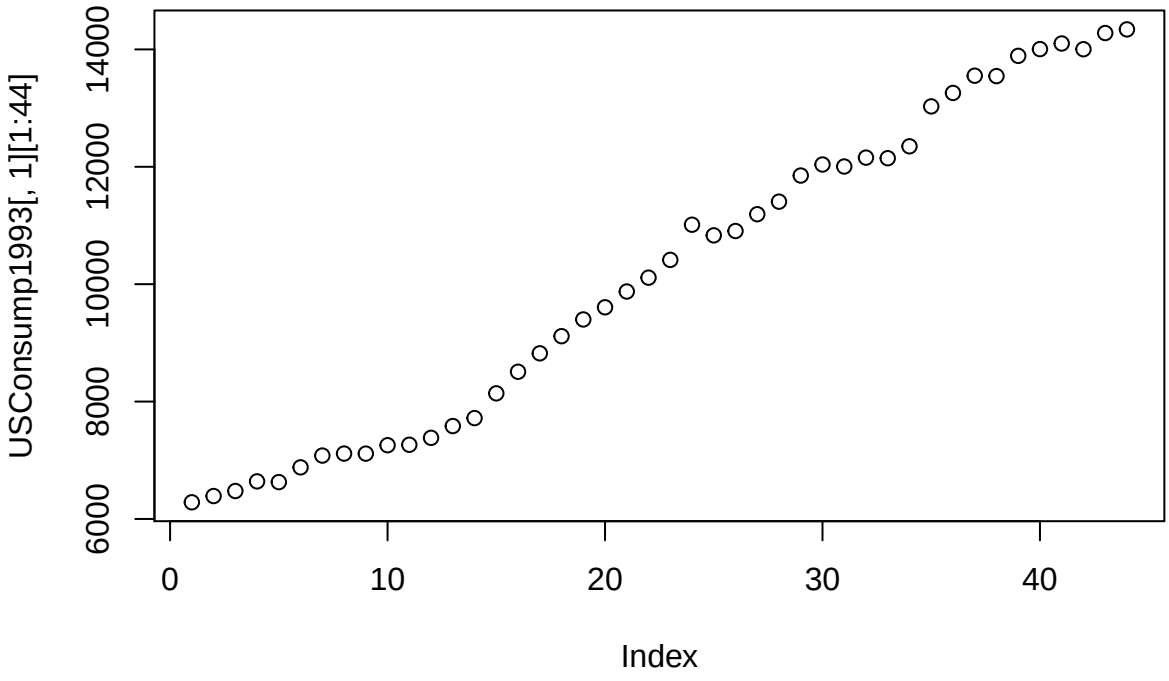
The P-P plot gives mixed results about which distribution is the best.

Conclusion about Income

We conclude that the Income variable is best approximated by a uniform distribution.

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

```
plot(USConsump1993[,1][1:44])
```

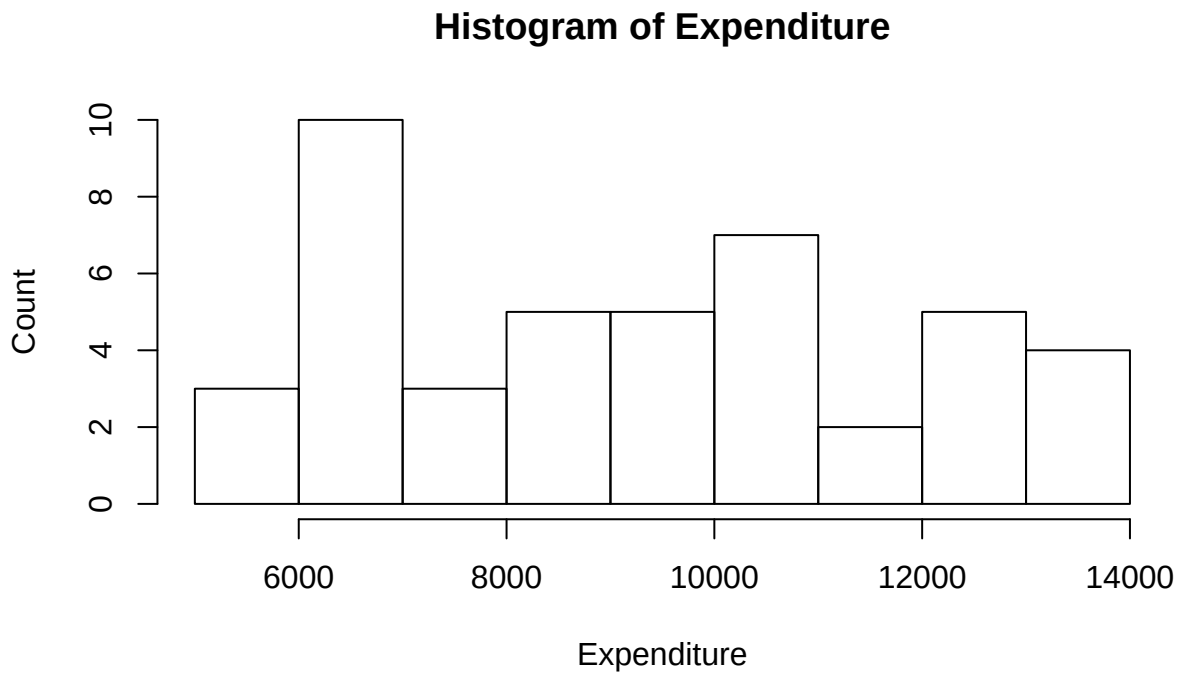


It does indeed appear that Income follows a relatively uniform distribution.

Histogram and Density Curve for Expenditure

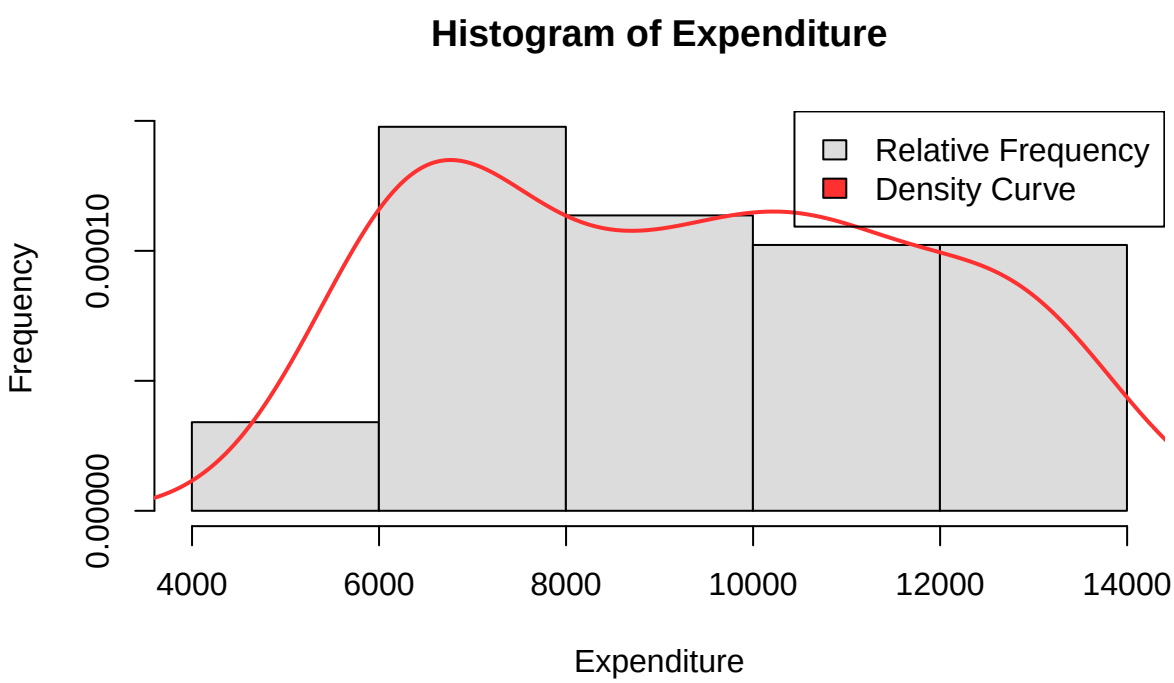
Histogram

```
hist(USConsump1993[,2], xlab= "Expenditure", ylab= "Count",
     main= "Histogram of Expenditure")
```



Histogram and Density Curve

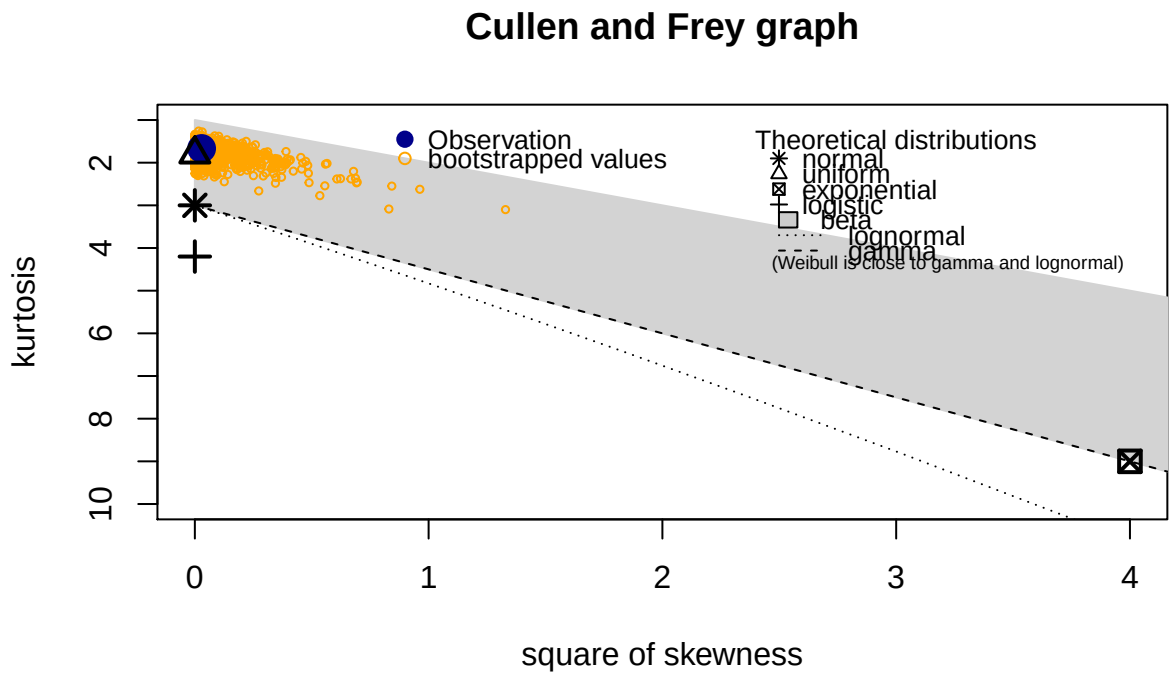
```
truehist(USConsump1993[,2],col="gainsboro", ylab="Frequency",
         xlab= "Expenditure",
         main= "Histogram of Expenditure")
lines(density((USConsump1993[,2])), lwd=2,col="firebrick1")
legend("topright", c("Relative Frequency", "Density Curve"),
       fill=c("gainsboro", "firebrick1"))
```


Histogram of Expenditure
Relative Frequency
Density Curve
Frequency
0.00010
0.00000
4000
6000
8000
10000
12000
14000
Expenditure

Note that from the summary statistics, we know that both Income and Expenditure have 44 observations.

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

```
descdist(USConsump1993[,2][1:44], boot = 1000)
```




```
## summary statistics
## ------
## min:  5820   max:  13391
## median:  9223.5
## mean:  9250.545
## estimated sd:  2484.624
## estimated skewness:  0.1702747
## estimated kurtosis:  1.667652
```

It appears that Expenditure is very likely to follow a uniform distribution.

Due to the square of skewness values being close to zero, we will also test normal and logistic distributions just to be sure those are indeed not good fits.

We proceed to fit a uniform distribution, a normal distribution, and a logistic distribution.

Testing fits for distributions

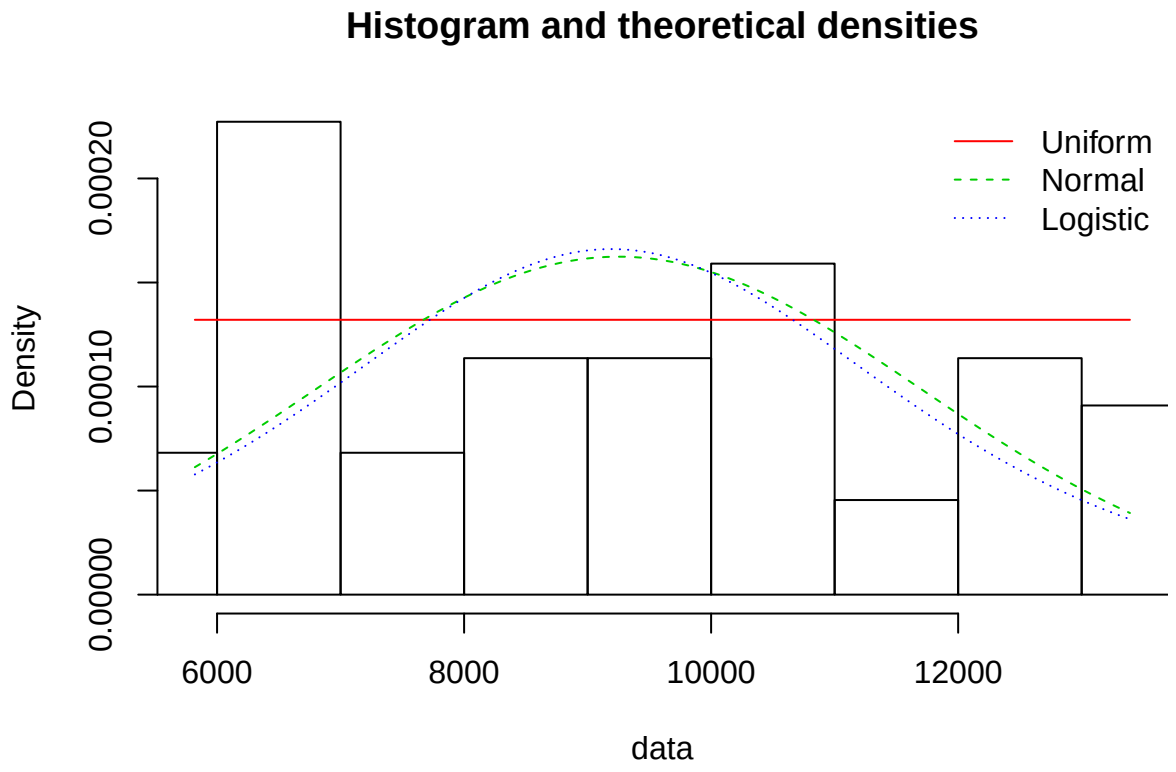
```
# Testing for a uniform distribution
expenditureunif <- fitdist(USConsump1993[,2][1:44], "unif")
# Testing for a normal distribution
expenditurenorm <- fitdist(USConsump1993[,2][1:44], "norm")
# Testing fit for a logistic distribution
expenditurelogis <- fitdist(USConsump1993[,2][1:44], "logis")
```

Setting Legend

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

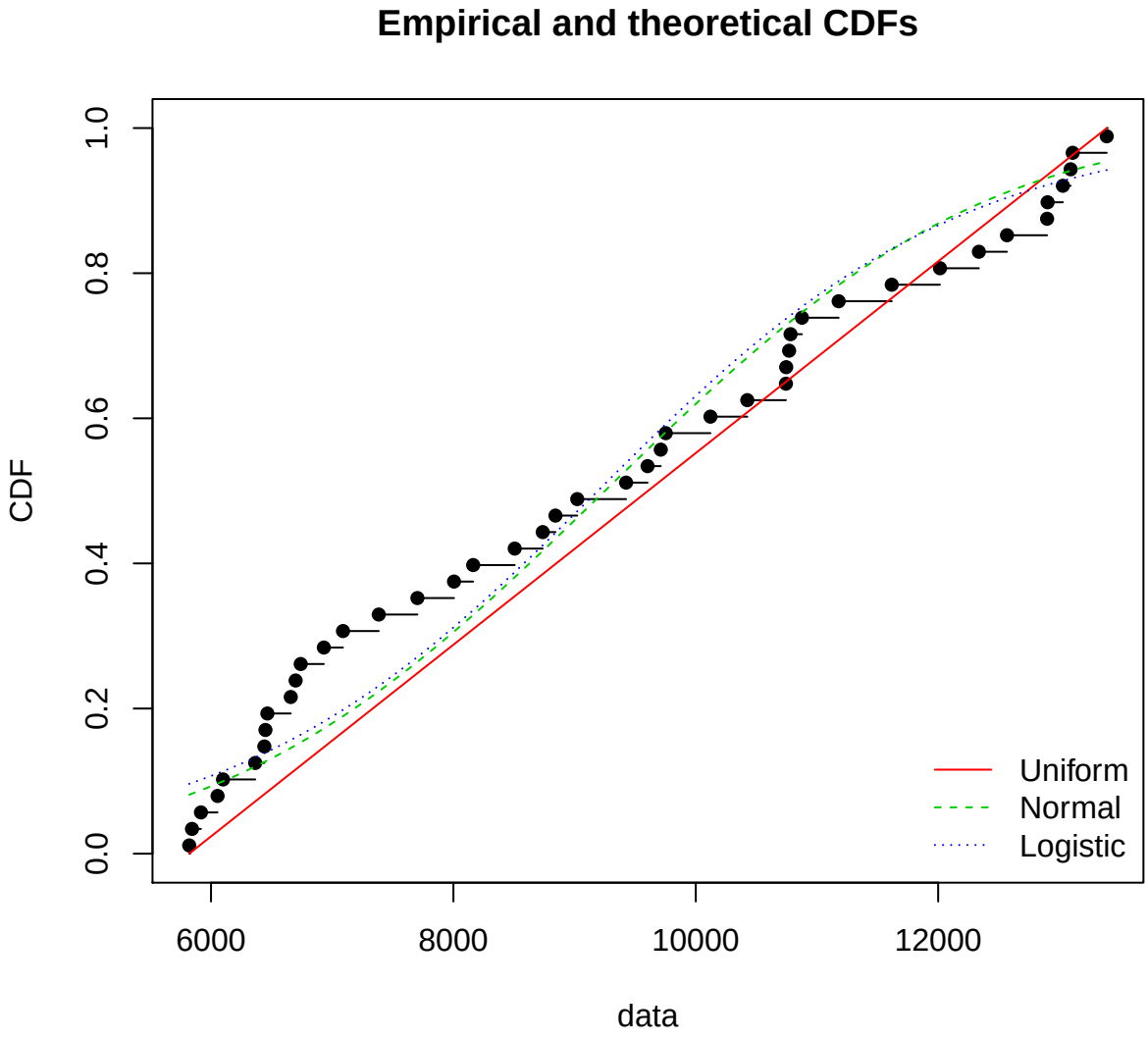
Comparing Histogram and Theoretical Densities

```
denscomp(list(expenditureunif, expenditurenorm, expenditurelogis), legendtext = plot.legend)
```



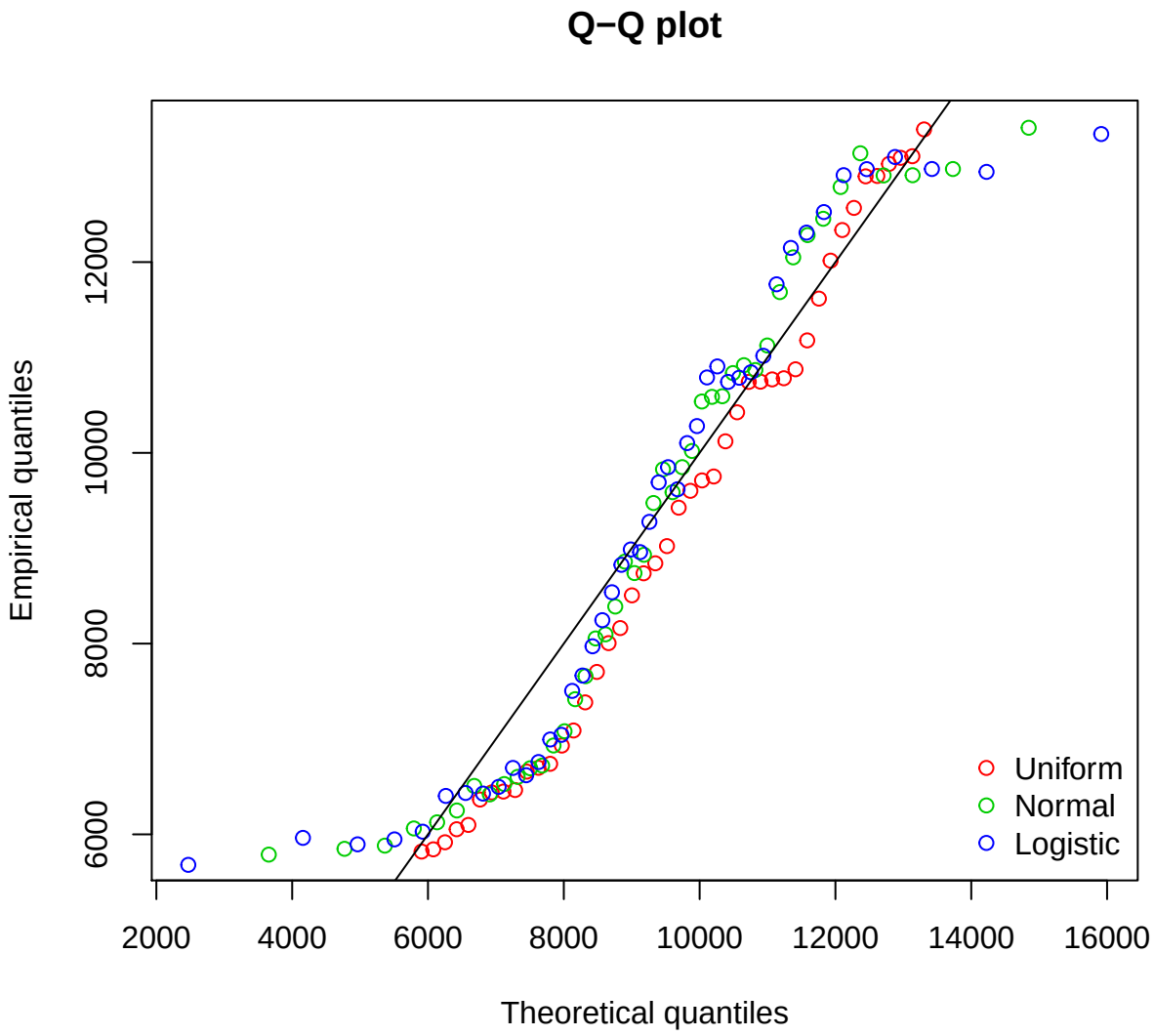
Observe that the uniform distribution seems to be the best fit based on the theoretical densities of the distributions.

```
cdfcomp(list(expenditureunif, expenditurenorm, expenditurelogis), legendtext = plot.legend)
```



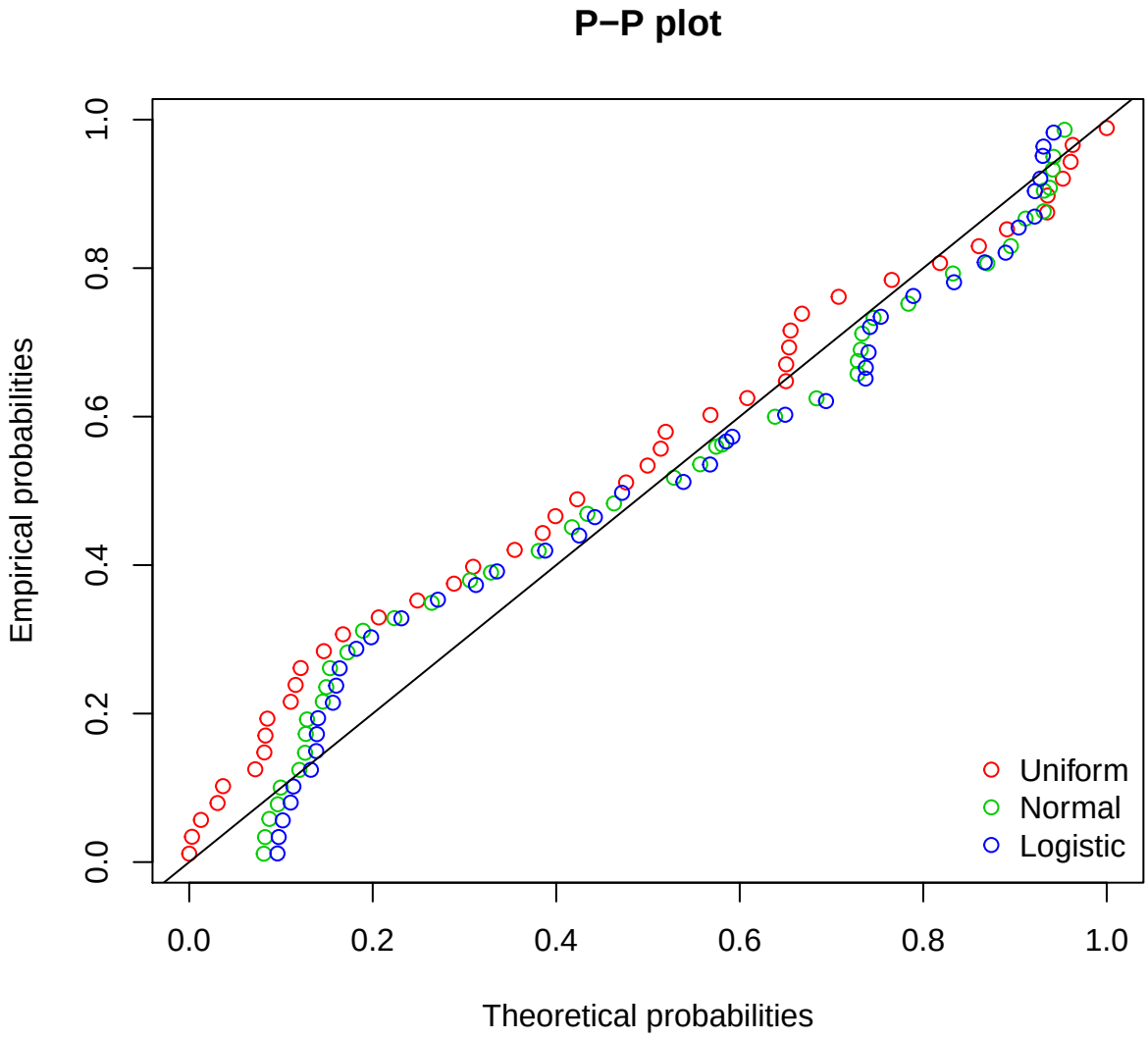
Observe that comparing the empirical and theoretical CDFs gives us mixed results as to which distribution is best.

```
qqcomp(list(expenditureunif, expenditurenorm, expenditurelogis), legendtext = plot.legend)
```



As far as the empirical quantiles compared to the theoretical quantiles, the uniform distribution is much better than the normal and logistic distributions.

```
ppcomp(list(expenditureunif, expenditurenorm, expenditurelogis), legendtext = plot.legend)
```



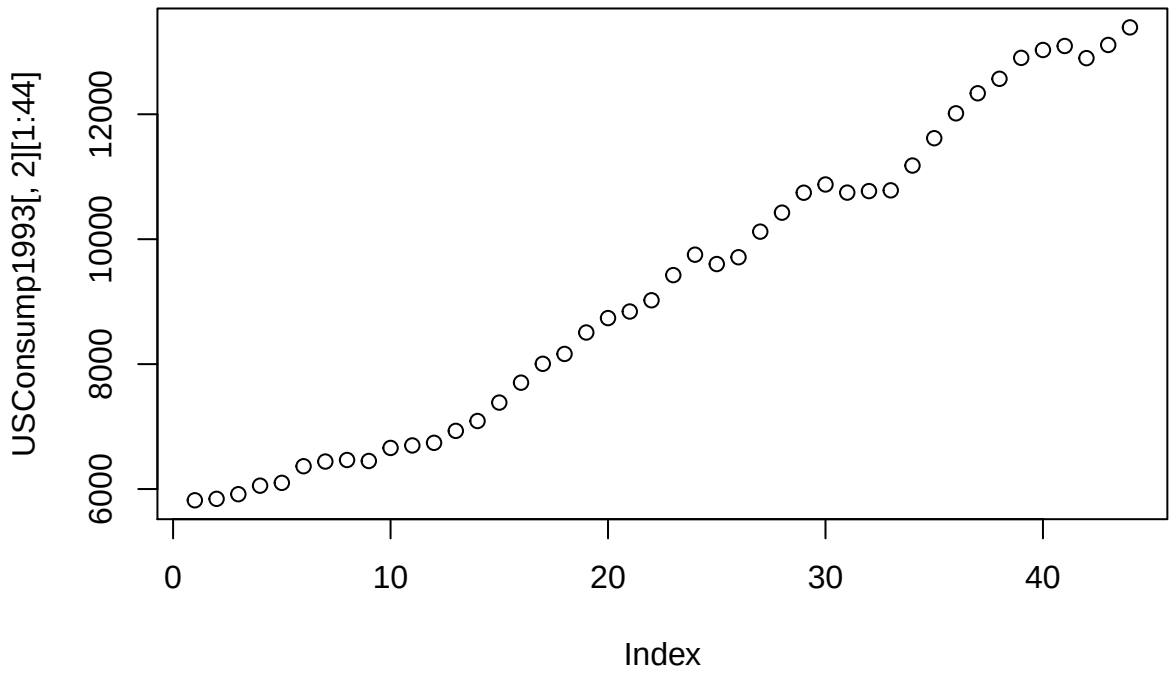
The P-P plot gives mixed results about which distribution is the best.

Conclusion about Expenditure

We conclude that the Expenditure variable is best approximated by a uniform distribution.

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

```
plot(USConsump1993[,2][1:44])
```



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

PROJECT 1 PART III (c)

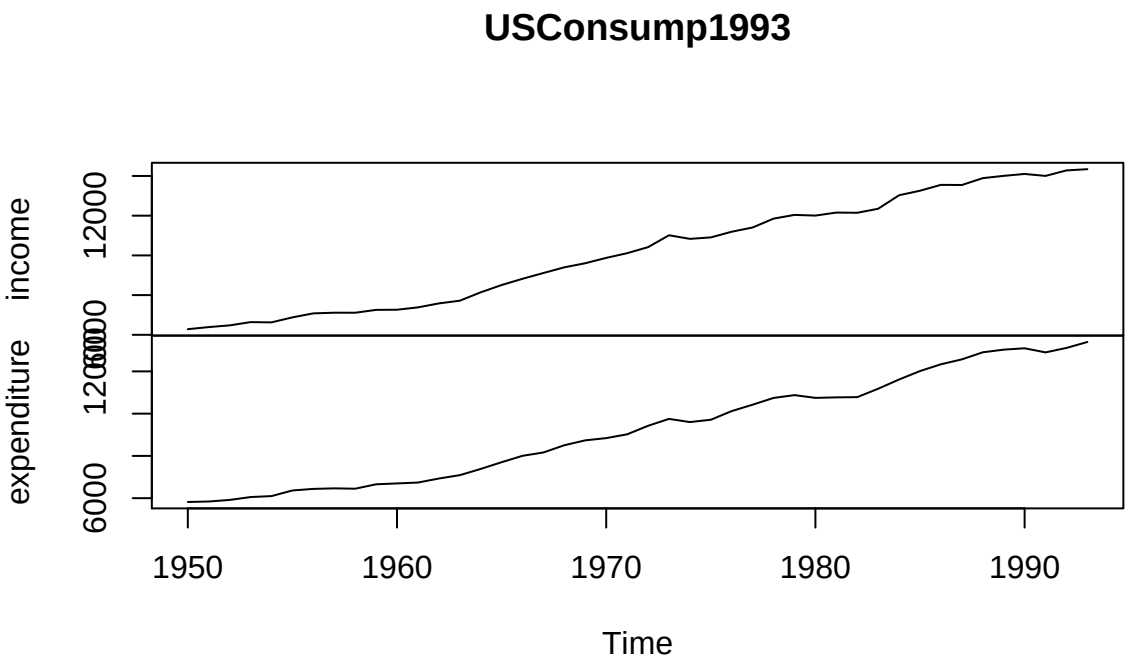
Regressing Income on Expenditure using a regular OLS model:

```
Pr1IIIc <- lm(Pr1Income ~ Pr1Expenditure)
summary(Pr1IIIc)
```

```
##
## Call:
## lm(formula = Pr1Income ~ Pr1Expenditure)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -338.99  -99.72  -11.83   80.75  327.86
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    109.57289   98.36319   1.114    0.272
## Pr1Expenditure   1.08808    0.01028 105.874   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 167.4 on 42 degrees of freedom
## Multiple R-squared:  0.9963, Adjusted R-squared:  0.9962
## F-statistic: 1.121e+04 on 1 and 42 DF,  p-value: < 2.2e-16
```

Note that the plots of the two variables are the following:

```
plot(USConsump1993)
```



The coefficient for expenditure being 1.08808 is consistent with what we see in the above plot.

PROJECT 1 PART III (d)

We were notified on January 20 by the professor that "you can skip question parts 3d and 4c."

PROJECT 1 PART IV: Women's Education

Getting the data (since the original data set did not have labels, we confirmed during office hours that we could acquire and use a version of the data set - that did include labels - from an alternate source):

```
Pr1IVdatafile <- "fertil1.dta"
Pr1IV <- read.dta(Pr1IVdatafile)
```

PROJECT 1 PART IV (a)

Using OLS to estimate model relating number of children ever born to a woman to years of education, age, region, race, and type of environment reared in. A quadratic in age and year dummy variables are included.

```
Pr1IVa <- lm(kids ~ educ + age + east + northcen + west
                  + black + farm + othrural + town
                  + smcity + agesq + y74 + y76 + y78
                  + y80 + y82 + y84 , data = Pr1IV)
summary(Pr1IVa)
```

```
##
## Call:
## lm(formula = kids ~ educ + age + east + northcen + west + black +
##     farm + othrural + town + smcity + agesq + y74 + y76 + y78 +
##     y80 + y82 + y84, data = Pr1IV)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.9878 -1.0086 -0.0767  0.9331  4.6548
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -7.742457   3.051767  -2.537 0.011315 *
## educ        -0.128427   0.018349  -6.999 4.44e-12 ***
## age          0.532135   0.138386   3.845 0.000127 ***
## east         0.217324   0.132788   1.637 0.101992
## northcen     0.363114   0.120897   3.004 0.002729 **
## west         0.197603   0.166913   1.184 0.236719
## black        1.075658   0.173536   6.198 8.02e-10 ***
## farm        -0.052557   0.147190  -0.357 0.721105
## othrural    -0.162854   0.175442  -0.928 0.353481
## town         0.084353   0.124531   0.677 0.498314
## smcity       0.211879   0.160296   1.322 0.186507
## agesq       -0.005804   0.001564  -3.710 0.000217 ***
## y74          0.268183   0.172716   1.553 0.120771
## y76         -0.097379   0.179046  -0.544 0.586633
## y78         -0.068666   0.181684  -0.378 0.705544
## y80         -0.071305   0.182771  -0.390 0.696511
## y82         -0.522484   0.172436  -3.030 0.002502 **
## y84         -0.545166   0.174516  -3.124 0.001831 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.555 on 1111 degrees of freedom
## Multiple R-squared:  0.1295, Adjusted R-squared:  0.1162
## F-statistic: 9.723 on 17 and 1111 DF,  p-value: < 2.2e-16
```

Estimated relationship between fertility and education:

The coefficient for the years of education variable in the model is -0.128427, which indicates that for each additional year of education, a woman has 0.128427 less children.

```
1/0.128427
```

```
## [1] 7.786525
```

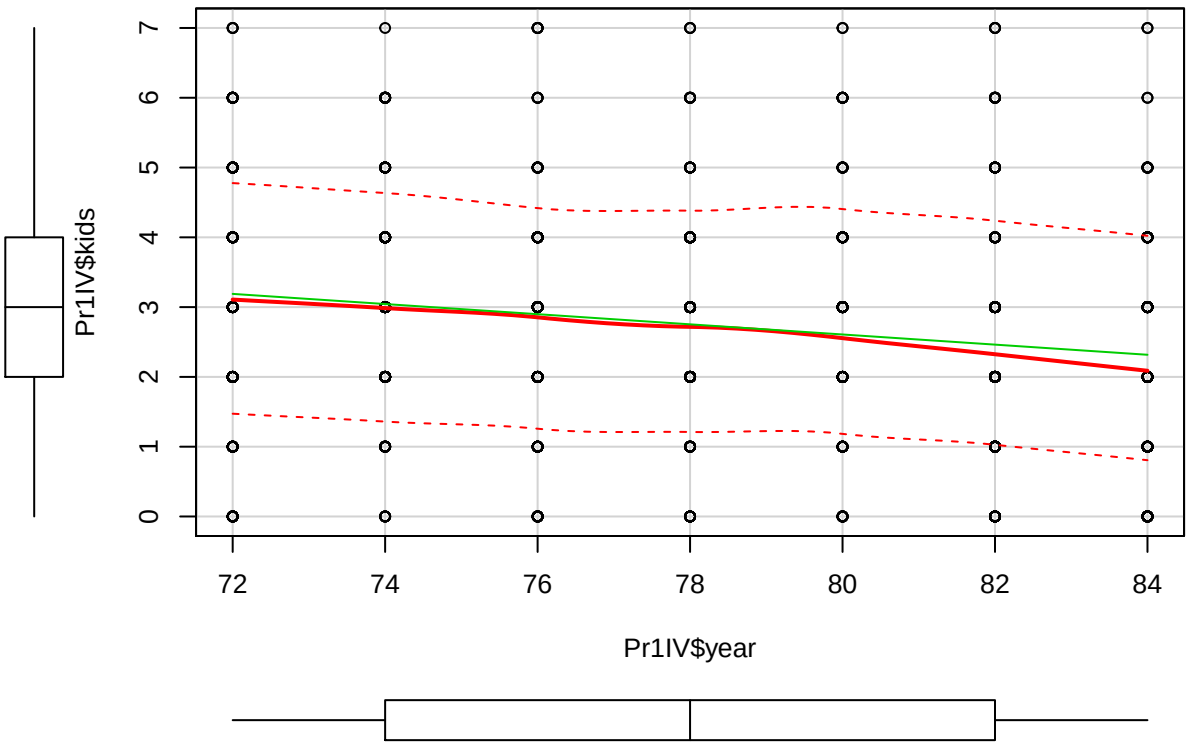
Note that $1/0.128427 = 7.786525$, so for around every 8 years of education that a woman has, she is estimated to have 1 less child.

Notable secular change in fertility over the time period:

We note from the year dummy variables that the coefficients of y74, y76, y78, y80, y82, and y84 are 0.268183, -0.097379, -0.068666, -0.071305, -0.522484, and -0.545166, respectively. So it appears that fertility has been decreasing over the time period considered. However, since y74, y76, y78, and y80 have p-values above 0.10, with y76, y78, and y80 having particularly high p-values (each over 0.50), we can not draw strong conclusions from the coefficients of these particular variables. Nonetheless, since y82 and y84 do indeed have very low p-values, we conclude that there is a long-term negative trend in fertility over the time period.

Just to verify our conclusion, we plot a loess smoother (red) and a regression line (green) for the kids and year variables.

```
scatterplot(Pr1IV$year, Pr1IV$kids, smoother = loessLine)
```



Our plot verifies our conclusion from our OLS model that there is a long-term negative trend in fertility over the time period.

PROJECT 1 PART IV (b)

We need to re-estimate the model above, using mother's education and father's education as instruments for education.

We first verify that educ and meduc, and educ and feduc, are indeed correlated:

```
Pr1IV.MeducCORR <- lm(educ ~ meduc, data = Pr1IV)
summary(Pr1IV.MeducCORR)
```

```
##
## Call:
## lm(formula = educ ~ meduc, data = Pr1IV)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.5714  -1.5714  -0.3433   1.1917   9.1128
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  9.88724    0.17269   57.25   <2e-16 ***
## meduc        0.30701    0.01731   17.73   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.336 on 1127 degrees of freedom
## Multiple R-squared:  0.2182, Adjusted R-squared:  0.2175
## F-statistic: 314.5 on 1 and 1127 DF,  p-value: < 2.2e-16
```

Due to the low p-value, and the non-zero coefficient for meduc, we note that the correlation between education and father's education is nonzero.

```
Pr1IV.FeducCORR <- lm(educ ~ feduc, data = Pr1IV)
summary(Pr1IV.FeducCORR)
```

```
##
## Call:
## lm(formula = educ ~ feduc, data = Pr1IV)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.0799  -1.5043  -0.0799   1.0714   9.7687
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  9.23126    0.20489   45.05   <2e-16 ***
## feduc        0.35609    0.01984   17.94   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.33 on 1127 degrees of freedom
## Multiple R-squared:  0.2222, Adjusted R-squared:  0.2215
## F-statistic:   322 on 1 and 1127 DF,  p-value: < 2.2e-16
```

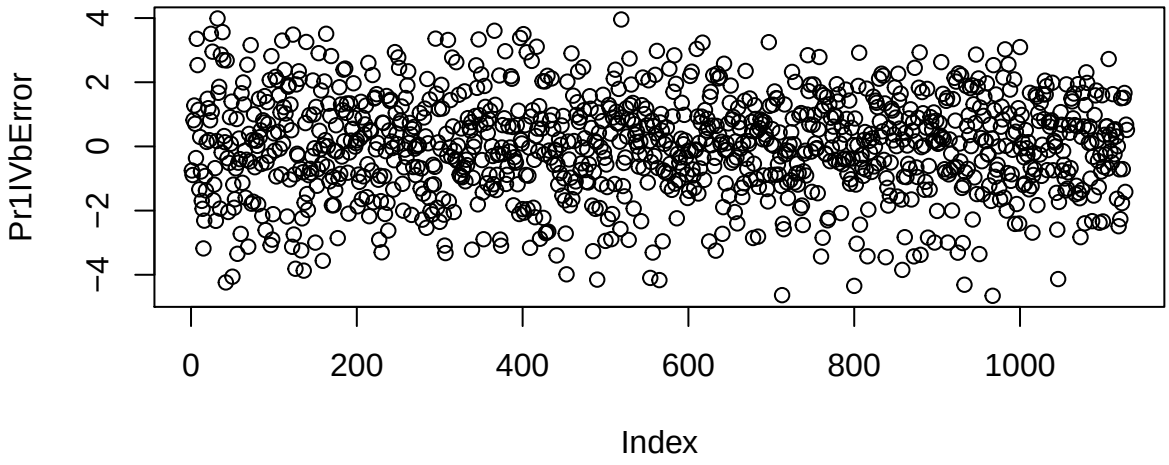
Due to the low p-value, and the non-zero coefficient for feduc, we note that the correlation between education and father's education is nonzero.

Thus, we conclude the relevance of both variables.

Now we proceed to check for exogeneity, in that mother's education and father's education are not correlated with the error term from the original regression.

Estimating errors from the original regression:

```
# Calculating Errors
Pr1IVbError <- fitted(Pr1IVa) - Pr1IV$kids
# Plotting Errors
plot(Pr1IVbError)
```




```
# Histogram of Errors
hist(Pr1IVbError)
```

We note that there does not seem to be a discernible pattern in the errors.

We proceed to check that mother's and father's education is uncorrelated with the errors from our regression.

```
meducCHECK <- lm(Pr1IVbError ~ Pr1IV$meduc)
summary(meducCHECK)
```

```
##
## Call:
## lm(formula = Pr1IVbError ~ Pr1IV$meduc)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.6707 -0.9393  0.0712  1.0203  3.9885
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.047173   0.114139  -0.413    0.679
## Pr1IV$meduc  0.005166   0.011442   0.451    0.652
##
## Residual standard error: 1.544 on 1127 degrees of freedom
## Multiple R-squared:  0.0001808,  Adjusted R-squared:  -0.0007063
## F-statistic: 0.2038 on 1 and 1127 DF,  p-value: 0.6517
```

Due to the high p-value, we conclude that mother's education is not correlated with the error from the original regression.

```
feducCHECK <- lm(Pr1IVbError ~ Pr1IV$feduc)
summary(feducCHECK)
```

```
##
## Call:
## lm(formula = Pr1IVbError ~ Pr1IV$feduc)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.6570 -0.9491  0.0680  1.0080  4.0088
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.07501    0.13576  -0.552    0.581
## Pr1IV$feduc  0.00772    0.01315   0.587    0.557
##
## Residual standard error: 1.544 on 1127 degrees of freedom
## Multiple R-squared:  0.0003058,  Adjusted R-squared:  -0.0005813
## F-statistic: 0.3447 on 1 and 1127 DF,  p-value: 0.5572
```

Due to the high p-value, we conclude that father's education is not correlated with the error from the original regression.

Therefore, we conclude that mother's education and father's education are valid instruments for education.

We re-estimate our model, using mother's education and father's education as instruments for education:

```
Pr1IVb <- lm(kids ~ educ + age + east + northcen + west
                  + black + farm + othrural + town
                  + smcity + agesq + y74 + y76 + y78
                  + y80 + y82 + y84 + meduc + feduc, data = Pr1IV)
```

Summary of model:

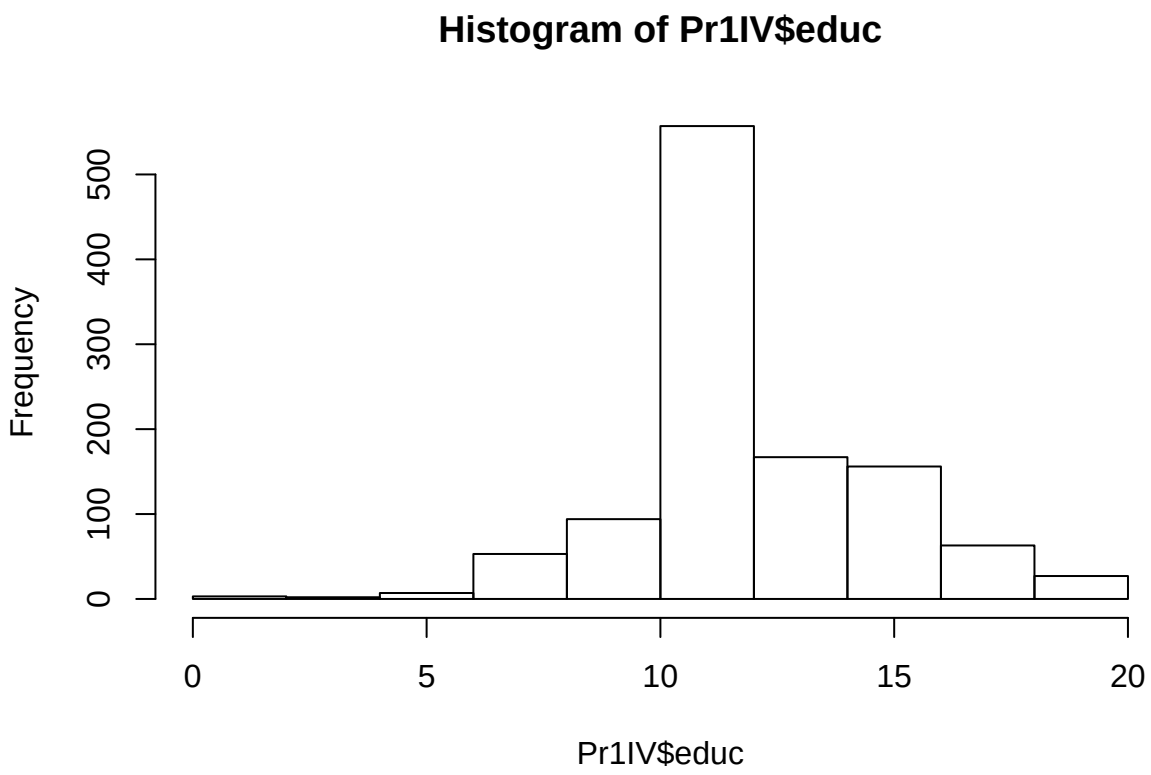
```
summary(Pr1IVb)
```

```
##
## Call:
## lm(formula = kids ~ educ + age + east + northcen + west + black +
##     farm + othrural + town + smcity + agesq + y74 + y76 + y78 +
##     y80 + y82 + y84 + meduc + feduc, data = Pr1IV)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.9879 -1.0026 -0.0625  0.9315  4.6631
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -7.652949   3.056940  -2.503 0.012441 *
## educ        -0.121602   0.020779  -5.852 6.38e-09 ***
## age          0.530054   0.138535   3.826 0.000137 ***
## east         0.220216   0.133027   1.655 0.098121 .
## northcen     0.372397   0.121699   3.060 0.002267 **
## west         0.206237   0.167711   1.230 0.219065
## black        1.061702   0.174819   6.073 1.72e-09 ***
## farm        -0.063920   0.148624  -0.430 0.667220
## othrural    -0.176670   0.176998  -0.998 0.318423
## town         0.081368   0.125159   0.650 0.515752
## smcity       0.210263   0.160436   1.311 0.190274
## agesq       -0.005790   0.001566  -3.698 0.000228 ***
## y74          0.271413   0.172950   1.569 0.116859
## y76         -0.098247   0.179174  -0.548 0.583576
## y78         -0.063298   0.181964  -0.348 0.728012
## y80         -0.065244   0.183103  -0.356 0.721665
## y82         -0.513913   0.172965  -2.971 0.003030 **
## y84         -0.534214   0.175300  -3.047 0.002363 **
## meduc       -0.003280   0.015778  -0.208 0.835377
## feduc       -0.008821   0.018145  -0.486 0.626947
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.556 on 1109 degrees of freedom
## Multiple R-squared:  0.1299, Adjusted R-squared:  0.115
## F-statistic: 8.715 on 19 and 1109 DF,  p-value: < 2.2e-16
```

Observe that in the original model, the coefficient for educ was -0.128427 and had a p-value less than 0.001. In the model with meduc and feduc as instrumental variables, the coefficient for educ is -0.121602 and the p-value is less than 0.001.

Observe from the below histogram of the variable educ that the minimum is 0 and the maximum is 20:

```
hist(Pr1IV$educ)
```



We note that since the variable educ has a range between 0 and 20, the difference between the coefficient for educ in the original model (-0.128427), and the coefficient for educ in the model with meduc and feduc as instrumental variables (-0.121602), is negligible.