Econ 104L: Group

Project #3: Korean Welfare

Predicitng Income as a mesure of welfare or something

Ye Wang, Omer Abdelrahim, Shane Barry

Contents

1 Part 1 **1**

1.1 What We Want to Answer With This Model on Korean Welfare 1

2 Part 2 **2**

2.1 Descriptive Analysis of the Variables 2

3 Part 3 **16**

3.1 Pooled Model 16

3.2 Fixed Effects Model 18

3.3 Random Effects Model 26

3.4 Comparing the Models 27

3.5 Conclusion for Korean Welfare Model 28

```
Kwelfare<-read.csv("/Users/omerabdelrahim/Downloads/Korea Income and Welfare.csv")
attach(Kwelfare)
```

1 Part 1

1.1 What We Want to Answer With This Model on Korean Welfare

A multiyear study of Korean Welfare recipients is analyzed. In this case nincome will be the dependent variable and we desire to identify the most relevant variables that affect nincome and what may or may not be the determining factors of income that would place an individual on welfare.

The data itself is not balanced, but we wish to capture dynamics such as family size, age, education, region and year in order to explain differences in nincome among welfare recipients. We'd like to say that possibly such factors like age might determine a higher likelihood of receiving welfare, or heavily affecting nincome, but that is something that lies outside the viability of this model. It is unfortunately unable to determine questions of causality, and as such we find ourselves unable to make statements on the regressors affecting such and such resulting in a higher or lower income.

Yet it is still interesting to see what could be the possible effects that the regressors may have on nincome, such as one having a larger family resulting in a higher or lower income and as a result a higher necessity for welfare or vice a versa.

Ultimately we seek to answer what effects the various regressors have on nincome, and possibly what similarities exist among those who are recepients of welfare

2 Part 2

2.1 Descriptive Analysis of the Variables

Dataset: Korea Income and Welfare.csv

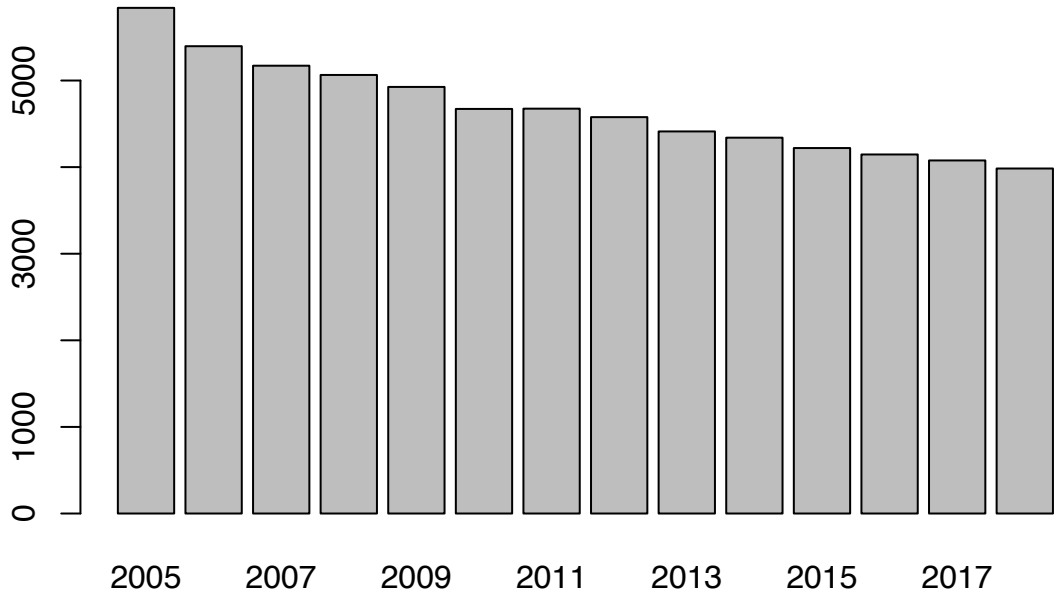
Contains information on a multiyear study from 2005 to 2018 with 65,499 observations of about ~10,000 individuals

13 variables in the data ;

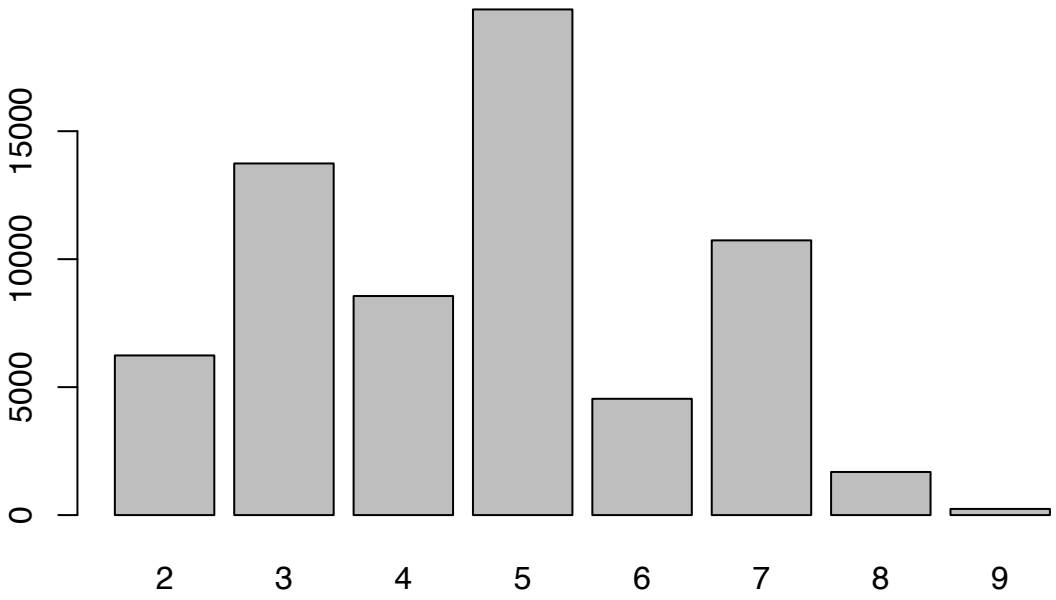
```
1. id

2. year : study conducted

3. wave : from the 1st wave in 2005 to the 14th wave in 2018

4. region:
  1) Seoul
  2) Kyeong-gi
  3) Kyoung-nam
  4) Kyoung-buk
  5) Chung-nam
  6) Gang-won &.Chung-buk
  7) Jeolla & Jeju

5. income: yearly income in M KRW(Million Korean Won. 1100 KRW = 1 USD)

6. family_member: no. of family members

7. gender:
  1) male
  2) female

8. year_born: The year that individul was born

9. education_level:
  1) no education(under 7 yrs-old)
  2) no education(7 & over 7 yrs-old)
  3) elementary
  4) middle school
  5) high school
  6) college
  7) university degree
  8) MA
  9) doctoral degree
```

```
10. marriage: marital status.
  1) not applicable (under 18)
  2) married
  3) separated by death
  4) separated
  5) not married yet
  6) others
11. religion:
  1) have religion
  2) do not have

12. yb: age

13. nincome: Income normalized on a 0-100 scale.
             Lowest Value: -232,174
             Highest Value: 468,209
```

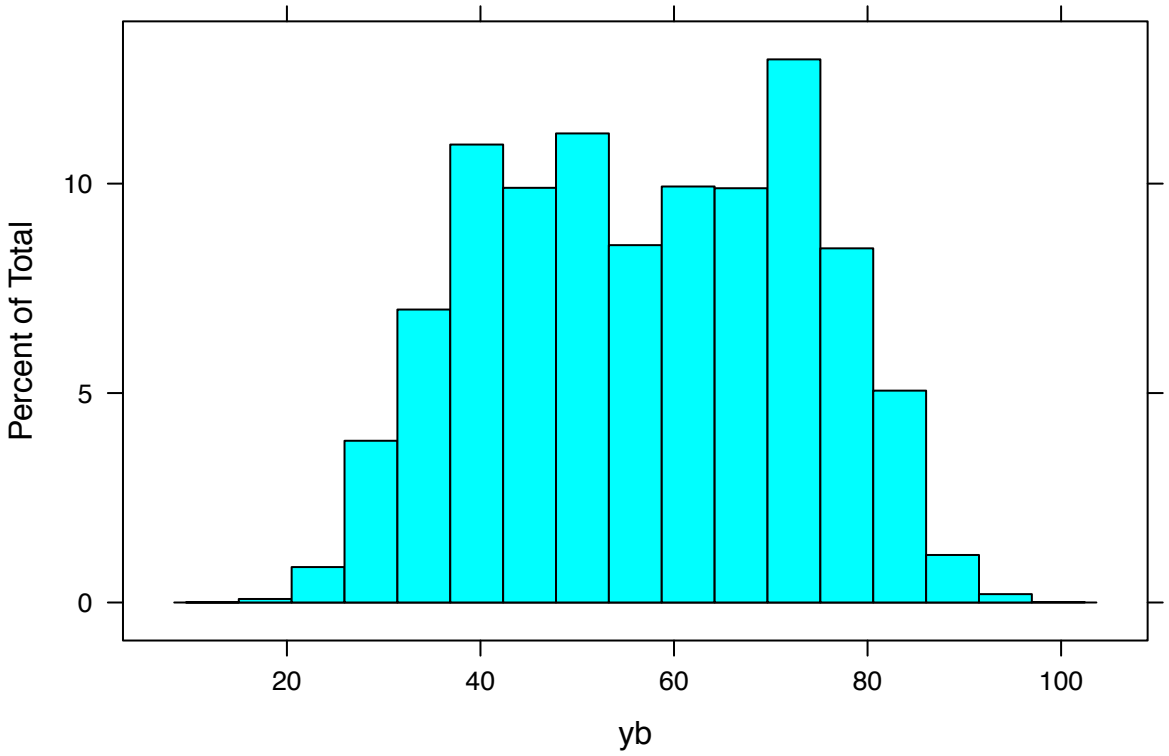
```
barplot(table(year))
```



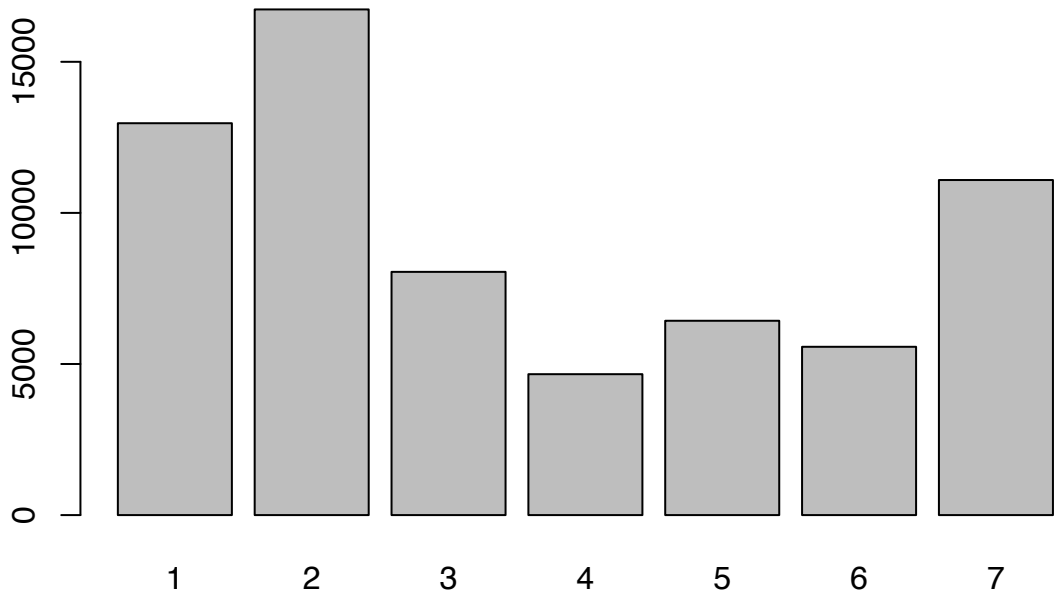

```
barplot(table(education_level))
```



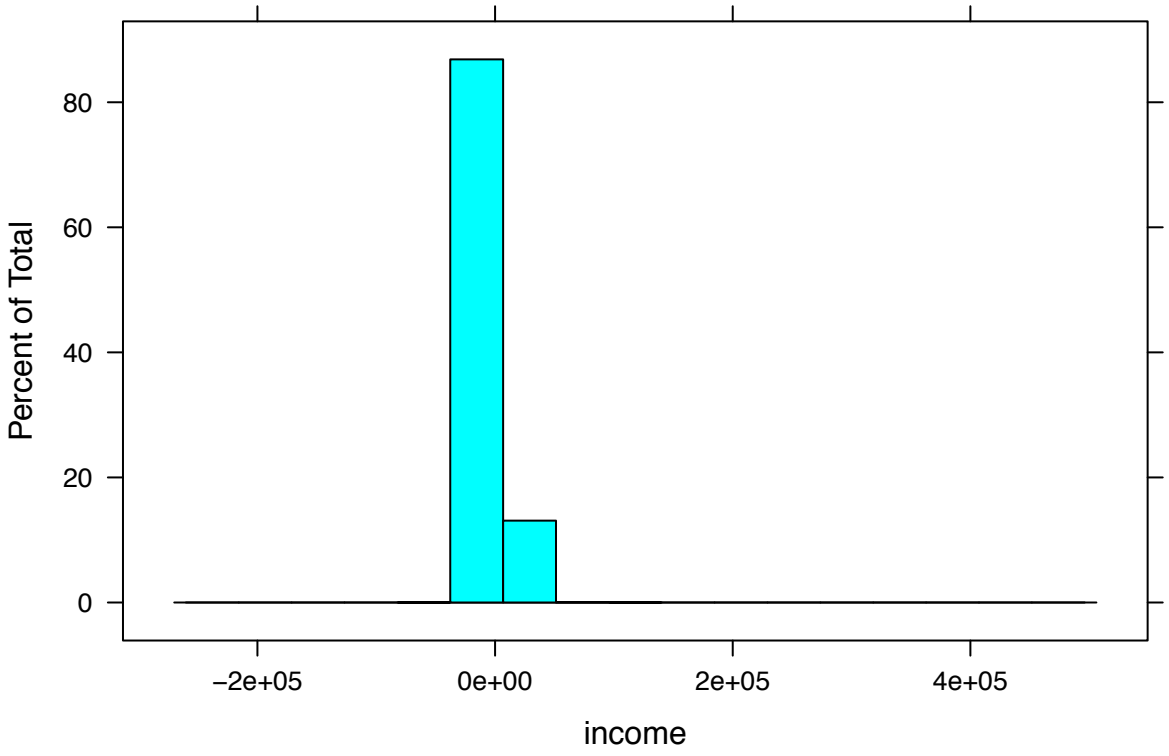

```
histogram(yb)
```



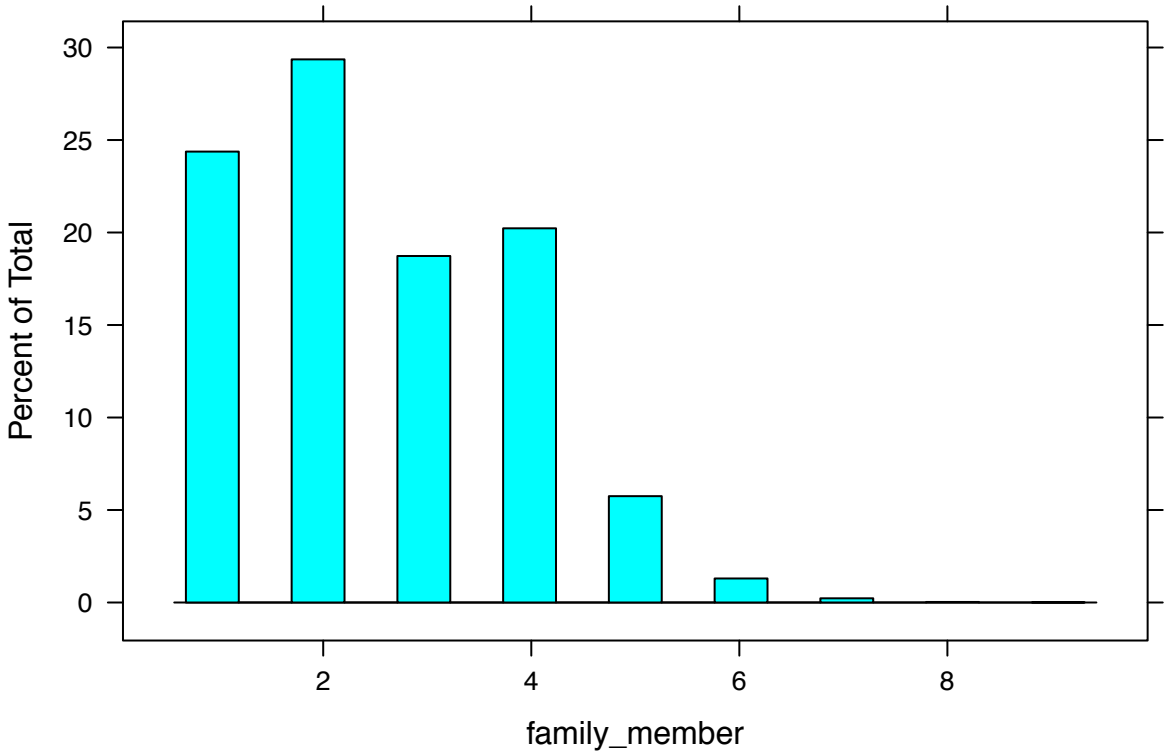

```
barplot(table(region))
```



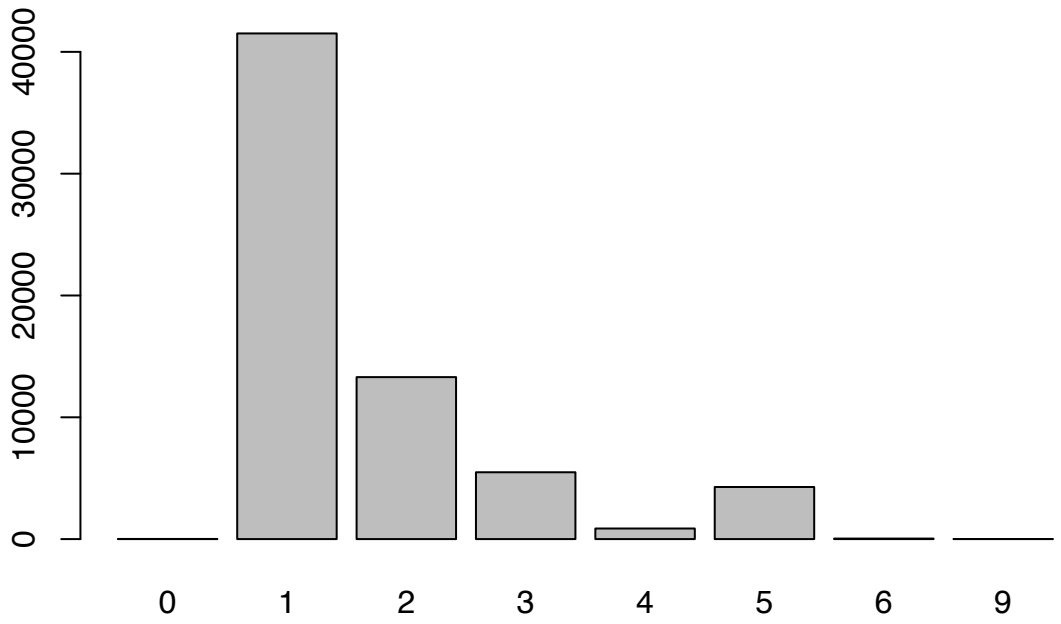

```
histogram(income)
```




```
histogram(family_member)
```




```
barplot(table(marriage))
```



As can be seen by the histogram of the year the majority of the recipients were tracked in the first year and then year after year the amount of observations fell year after year. This maybe due to the subjects no keeping up with the program within the first year, or the fact that the recipients were no longer eligible for healthcare.

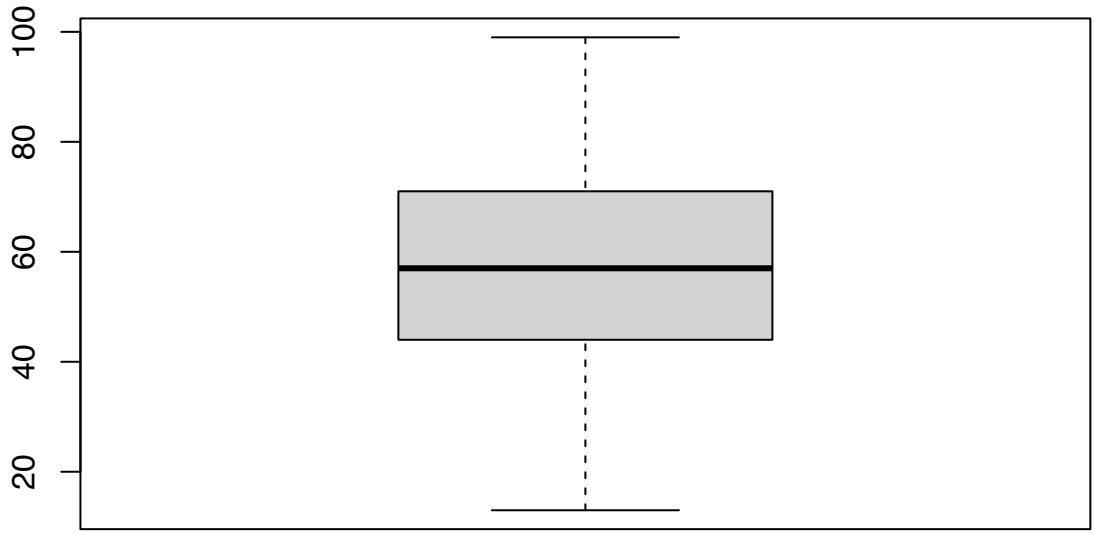
Most welfare recipients are between 40 and 75 years of age. I expected the majority of people with welfare to be young, but middle aged people with families, as well as the elderly would also be crime candidates for a welfare program.

Region shows the distribution of those receiving welfare across Korea, with most people hailing from Kyeong-gi

Income is grouped around zero, with heavy outliers of -200,000 and 400,000. Most people don't go above 1,000, with few above it.

Those who are single, have a family of 2-4 being the representative of the majority of those who receive welfare. This is also compounded by the individuals who are not married, indicating that there is only a family of 1, or that they support a family, such as elders or young children without being married themselves.

```
boxplot(yb)
```

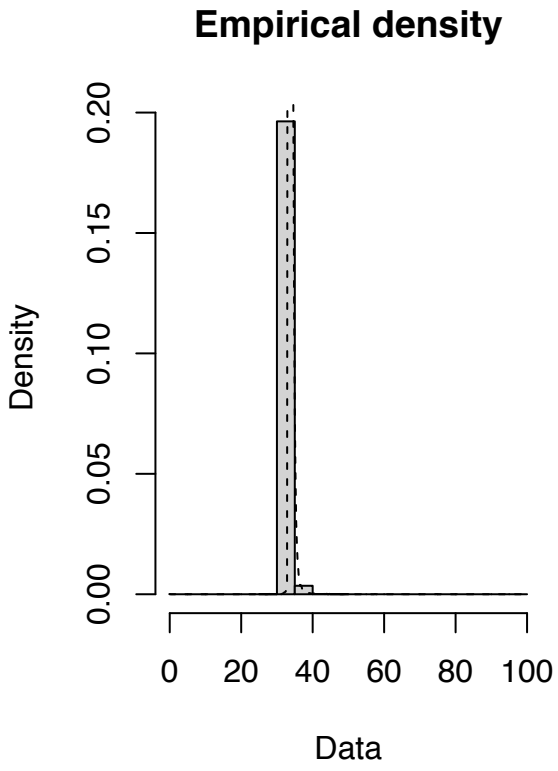


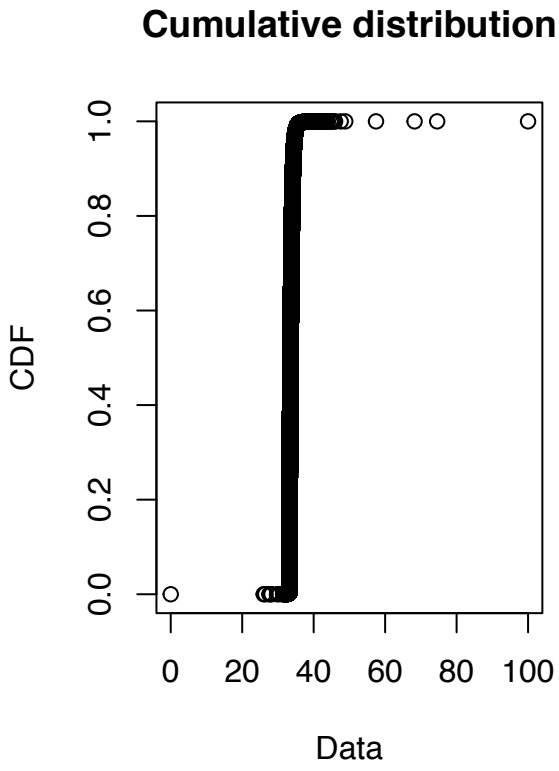

```
fivenum(yb)
```

```
## [1] 13 44 57 71 99
```

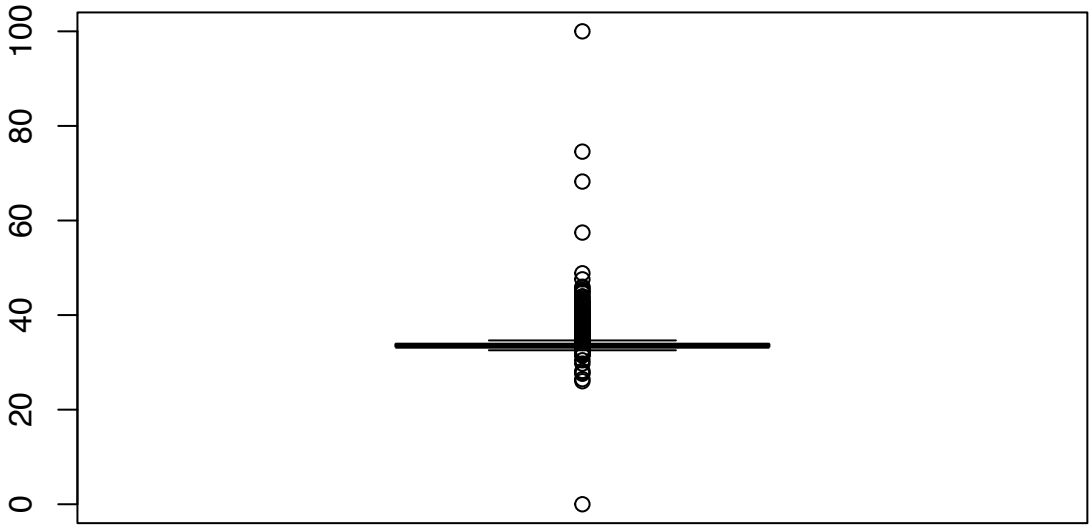
The mean age of individuals who are on welfare is 57, with the middle 50 being 44-71, showing that welfare in the country is skewed towards the middle aged and elderly.

```
plotdist(nincome,  histo = TRUE, demp = TRUE)
```





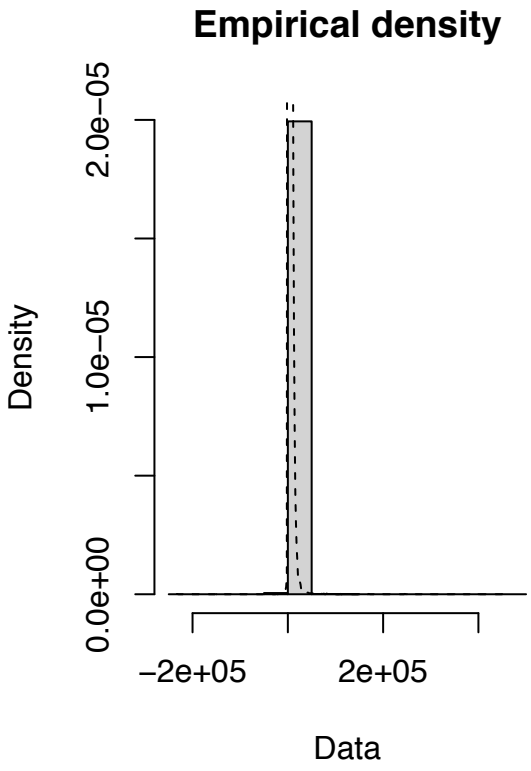

```
boxplot(nincome)
```

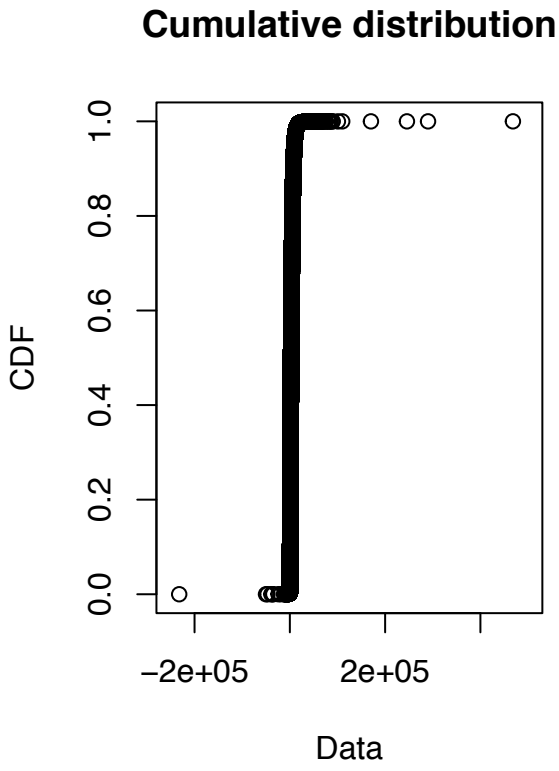



```
fivenum(nincome)
```

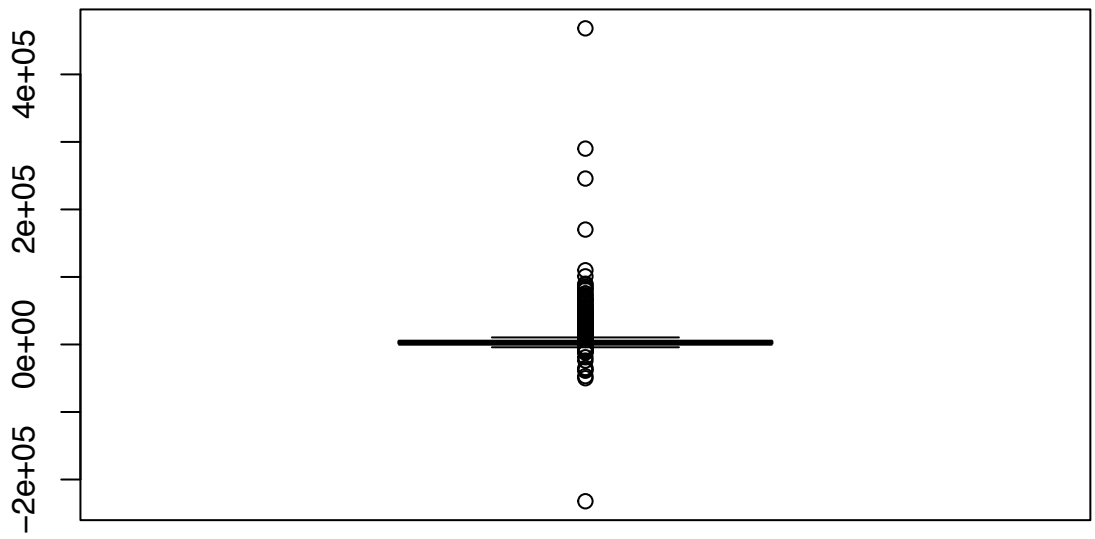
```
## [1]   0.00000  33.32734  33.52965  33.85319 100.00000
```

```
plotdist(income, histo = TRUE, demp = TRUE)
```





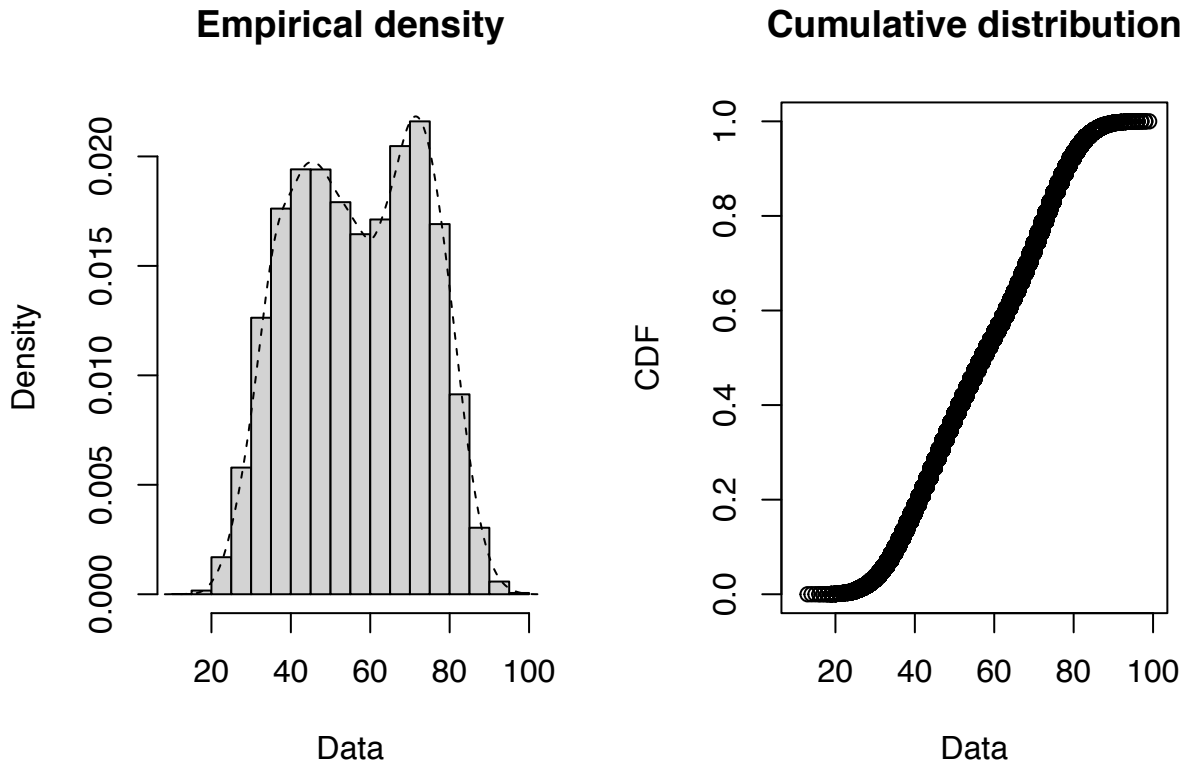

```
boxplot(income)
```



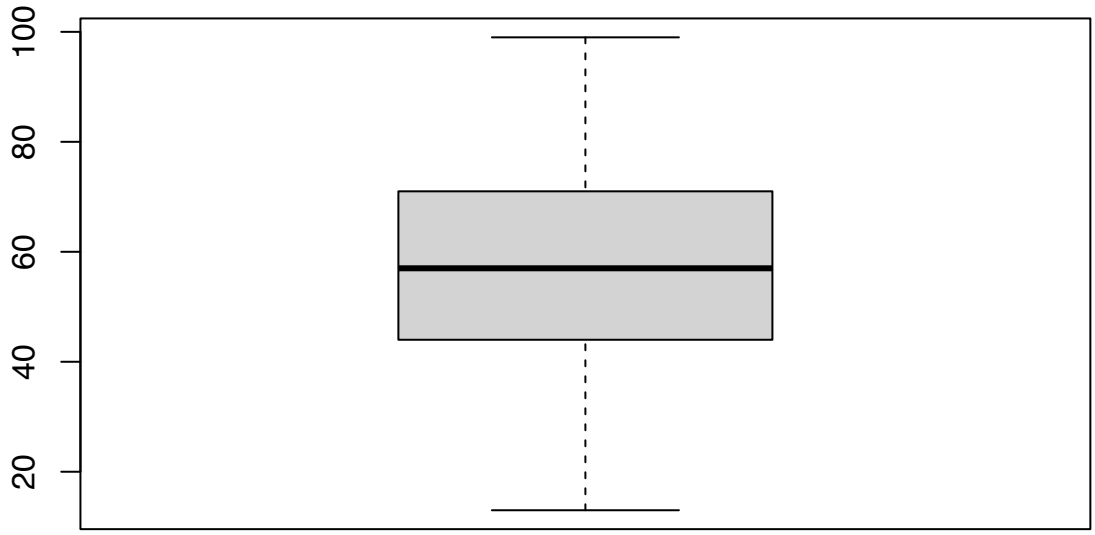

```
fivenum(income)
```

```
## [1] -232174    1245    2662    4928  468209
```

```
plotdist(yb, histo=TRUE, demp=TRUE)
```


Empirical density
Density
0.000
0.005
0.010
0.015
0.020
20
40
60
80
100
Data
Cumulative distribution
CDF
0.0
0.2
0.4
0.6
0.8
1.0
20
40
60
80
100
Data

```
boxplot(yb)
```




```
fivenum(yb)
```

```
## [1] 13 44 57 71 99
```

nincome and income tells us the same exact things. It doesn't really give a very accurate display of the data fram as there are incredibly massive outliers at both ends of the range. The only possible explanation for such outliers can include the fact those who experience large increases or devreases in income is a result of things such as gambling or the lotto among other things. For those who reported receiving welfare, especially at the top end of the range could have been jobless during the year, or experienced some form of cashflow via multiple outside factors.

yb, as has been said before, has an approximaltely normal distribution with small tails.

3 Part 3

3.1 Pooled Model

```
# Converting data to panel and creating a pooled model
Kwelfare.pd<- pdata.frame(Kwelfare, index=c("id", "year"))
head(Kwelfare.pd)
```

```
##               id year wave region income family_member gender year_born
```

```
## 10101-2005 10101 2005    1      1    614             1      2      1936
## 10101-2011 10101 2011    7      1    896             1      2      1936
## 10101-2012 10101 2012    8      1   1310             1      2      1936
## 10101-2013 10101 2013    9      1   2208             1      2      1936
## 10101-2014 10101 2014   10      1    864             1      2      1936
## 10101-2015 10101 2015   11      1   1171             1      2      1936
##            education_level marriage religion  nincome yb
## 10101-2005               2        2        2 33.23724 69
## 10101-2011               2        2        2 33.27751 75
## 10101-2012               2        2        2 33.33662 76
## 10101-2013               2        2        2 33.46483 77
## 10101-2014               2        2        2 33.27294 78
## 10101-2015               2        2        1 33.31677 79
```

```
Kwelfare.pooled <- plm(nincome~education_level+yb+region+family_member,
  model="pooling", data = Kwelfare.pd)
summary(Kwelfare.pooled)
```

```
## Pooling Model
##
## Call:
## plm(formula = nincome ~ education_level + yb + region + family_member,
##     data = Kwelfare.pd, model = "pooling")
##
## Unbalanced Panel: n = 6724, T = 1-14, N = 65499
##
## Residuals:
##      Min.   1st Qu.    Median   3rd Qu.      Max.
## -34.15576  -0.19952  -0.04129   0.10258  65.74217
##
## Coefficients:
##                    Estimate  Std. Error   t-value  Pr(>|t|)
## (Intercept)     32.83867616  0.01764718 1860.8451 < 2.2e-16 ***
## education_level  0.10079842  0.00169579   59.4403 < 2.2e-16 ***
## yb               0.00035603  0.00017761    2.0046   0.04501 *
## region          -0.00762035  0.00102630   -7.4251 1.141e-13 ***
## family_member    0.14091998  0.00190615   73.9289 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    26115
## Residual Sum of Squares: 20604
## R-Squared:      0.21104
## Adj. R-Squared: 0.211
## F-statistic: 4379.87 on 4 and 65494 DF, p-value: < 2.22e-16
```

```
# Estimating the Pooled OLS w/ Cluster-robust standard errors
coeftest(Kwelfare.pooled, vcov=vcovHC(Kwelfare.pooled,
                    type="HC0",cluster="group"))
```

```
##
## t test of coefficients:
##
```

```
##                    Estimate  Std. Error  t value  Pr(>|t|)
## (Intercept)     32.83867616  0.04366045 752.1379 < 2.2e-16 ***
## education_level  0.10079842  0.00435537  23.1435 < 2.2e-16 ***
## yb               0.00035603  0.00039742   0.8959    0.3703
## region          -0.00762035  0.00181642  -4.1953 2.729e-05 ***
## family_member    0.14091998  0.00441183  31.9414 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Of the various predictors of income, education_level and the amount of family members proved to be the mot powerful when comparing the recipients of welfare. This is surprising as one would think that recipients of welfare would not necessarily have as much of a difference between income values. Family is shown to be a boon to income, meanwhile region has a small effect on decresing income, probably having to do something with rural vs. urban economy and we can expect those in a rural setting to make less money and thus more likely are on welfare.

Age is the only major discrepancy between the two regressions, with it being slightly signficanct in the non cluster regression, and it being totally insignficant in the cluster regression. With how the ages are distributed in an almost flat way from 40-75, it can be expected that age does not really affect income in any meaningful way, as the majority have to be below a certain income threshold.

The regression might also be skewed towards family postively affecting income as those who have a large amount of family members don't really find themselves on the list. For what reason can only be guessed, but the limited sample size might point towards the fact that those in larger families don't really receive welfare.

3.2 Fixed Effects Model

We will see later that the Fixed Effects Within model is the best model for running this panel data regression

```
#Fixed Effects Dummy Variable approach
# Subset for the first 10 individuals
Kwelfare.pd10 <- pdata.frame(Kwelfare[Kwelfare$id%in%10101:100101,])
head(Kwelfare.pd10)
```

```
##               id year wave region income family_member gender year_born
## 10101-2005 10101 2005    1      1    614             1      2      1936
## 10101-2011 10101 2011    7      1    896             1      2      1936
## 10101-2012 10101 2012    8      1   1310             1      2      1936
## 10101-2013 10101 2013    9      1   2208             1      2      1936
## 10101-2014 10101 2014   10      1    864             1      2      1936
## 10101-2015 10101 2015   11      1   1171             1      2      1936
##            education_level marriage religion  nincome yb
## 10101-2005               2        2        2 33.23724 69
## 10101-2011               2        2        2 33.27751 75
## 10101-2012               2        2        2 33.33662 76
## 10101-2013               2        2        2 33.46483 77
## 10101-2014               2        2        2 33.27294 78
## 10101-2015               2        2        1 33.31677 79
```

```
#Estimating Dummy Variable Model
Kwelfare.fixed10 <- lm(nincome~education_level+yb+region+family_member+factor(id)-1,
                  data=Kwelfare.pd10)
summary(Kwelfare.fixed10)
```

```
##
## Call:
## lm(formula = nincome ~ education_level + yb + region + family_member +
##     factor(id) - 1, data = Kwelfare.pd10)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.249880 -0.038828 -0.001166  0.043404  0.252592
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## education_level    0.110296   0.098110   1.124 0.263818
## yb                 0.015380   0.002107   7.300 9.57e-11 ***
## region            -0.034930   0.044994  -0.776 0.439519
## family_member      0.109125   0.030606   3.566 0.000576 ***
## factor(id)10101   31.842997   0.270424 117.752  < 2e-16 ***
## factor(id)20101   31.818285   0.427837  74.370  < 2e-16 ***
## factor(id)30101   31.871023   0.332825  95.759  < 2e-16 ***
## factor(id)40101   31.573216   0.715069  44.154  < 2e-16 ***
## factor(id)50101   32.077442   0.509416  62.969  < 2e-16 ***
## factor(id)60101   31.739371   0.600422  52.862  < 2e-16 ***
## factor(id)70101   31.738836   0.274690 115.544  < 2e-16 ***
## factor(id)80101   31.786483   0.331512  95.883  < 2e-16 ***
## factor(id)90101   31.711843   0.437817  72.432  < 2e-16 ***
## factor(id)100101 31.968472   0.488617  65.426  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.07621 on 93 degrees of freedom
## Multiple R-squared:      1,  Adjusted R-squared:      1
## F-statistic: 1.468e+06 on 14 and 93 DF,  p-value: < 2.2e-16
```

Within the Fixed Effects model we can see that each of the intercepts are statistically significant as well as age and the amount of family members. Strangely enough education level and and region cease to be statistically significant, but maybe because those are identifier variables (i.e. they show individual differences rather than just being something like a counting stat) their values are captured by the various intercepts for each individual. Of course this is a regression that only shows the first 10 individuals of a sample that is very large and contains over 10,000 individuals observed, and thus we can't expect such a small slice to be totally accurate when it comes to predicting income among those who are recipients of welfare

```
# Fixed Effects Within Model
# Subset for the first 10 individuals
Kwelfare.pd10 <- pdata.frame(Kwelfare[Kwelfare$id %in% 10101:100101,])
head(Kwelfare.pd10)
```

```
##               id year wave region income family_member gender year_born
## 10101-2005 10101 2005    1      1    614             1      2      1936
## 10101-2011 10101 2011    7      1    896             1      2      1936
## 10101-2012 10101 2012    8      1   1310             1      2      1936
## 10101-2013 10101 2013    9      1   2208             1      2      1936
## 10101-2014 10101 2014   10      1    864             1      2      1936
## 10101-2015 10101 2015   11      1   1171             1      2      1936
##            education_level marriage religion  nincome yb
```

```
## 10101-2005               2        2        2 33.23724 69
## 10101-2011               2        2        2 33.27751 75
## 10101-2012               2        2        2 33.33662 76
## 10101-2013               2        2        2 33.46483 77
## 10101-2014               2        2        2 33.27294 78
## 10101-2015               2        2        1 33.31677 79
```

```
# Estimate the within model
Kwelfare.within10 <- plm(nincome~education_level+yb+region+family_member,
                  data=Kwelfare.pd10,
                  model="within",)
summary(Kwelfare.within10)
```

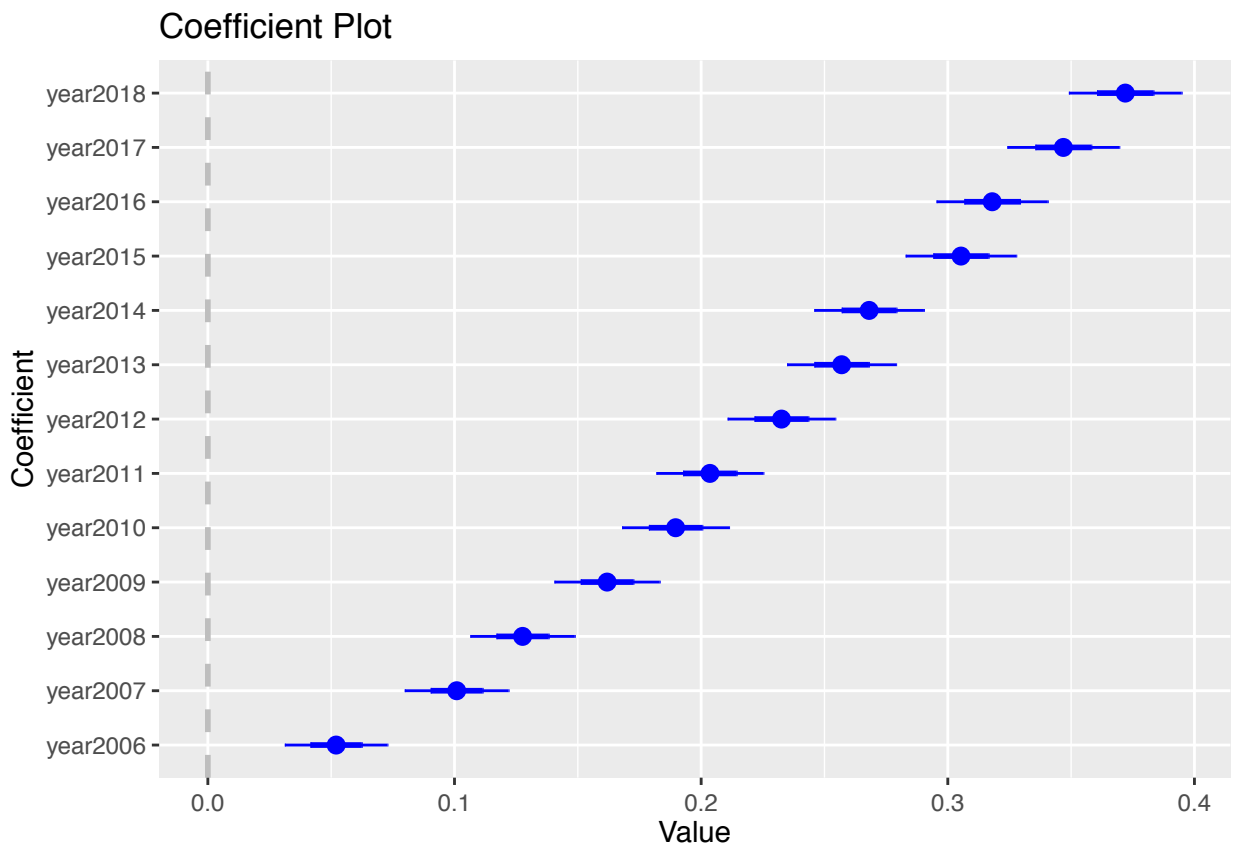
```
## Oneway (individual) effect Within Model
##
## Call:
## plm(formula = nincome ~ education_level + yb + region + family_member,
##     data = Kwelfare.pd10, model = "within")
##
## Unbalanced Panel: n = 10, T = 1-14, N = 107
##
## Residuals:
##       Min.    1st Qu.     Median    3rd Qu.       Max.
## -0.2498803 -0.0388281 -0.0011655  0.0434039  0.2525922
##
## Coefficients:
##                   Estimate Std. Error t-value  Pr(>|t|)
## education_level  0.110296   0.098110  1.1242 0.2638181
## yb               0.015380   0.002107  7.2997 9.565e-11 ***
## region          -0.034931   0.044994 -0.7763 0.4395192
## family_member    0.109126   0.030606  3.5655 0.0005758 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    0.90206
## Residual Sum of Squares: 0.54014
## R-Squared:      0.40121
## Adj. R-Squared: 0.31751
## F-statistic: 15.5785 on 4 and 93 DF, p-value: 8.6433e-10
```

```
fixef(Kwelfare.within10)
```

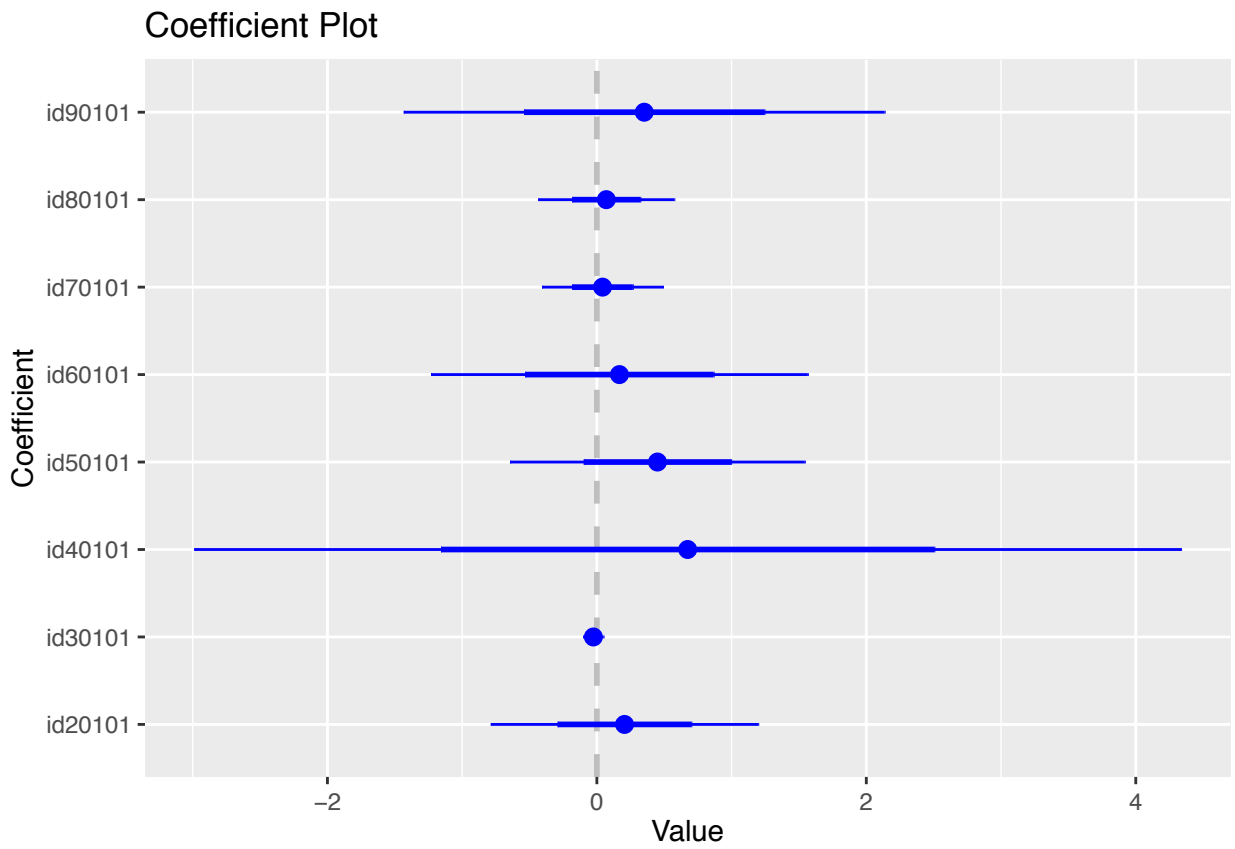
```
##  10101  20101  30101  40101  50101  60101  70101  80101  90101 100101
## 31.843 31.818 31.871 31.573 32.077 31.739 31.739 31.786 31.712 31.968
```

Again within the Within model we see similar measurements as the dummy variable approeach with two signficant variables that are the same and two insignificant variables that are the same. We can also check the individual affects using the fixef function in order to predict the intercept for each indiviudal. Again, this is only for the first 10 individuals in the sample, we will revisit this model using the entire dataset, of course with no summary.

```
#Coefficient plot, now including year
#Year
FE.Kwelfare<-lm(nincome~education_level+yb+region+family_member+year,
                data=Kwelfare.pd,)
coefplot(FE.Kwelfare, predictors="year")
```

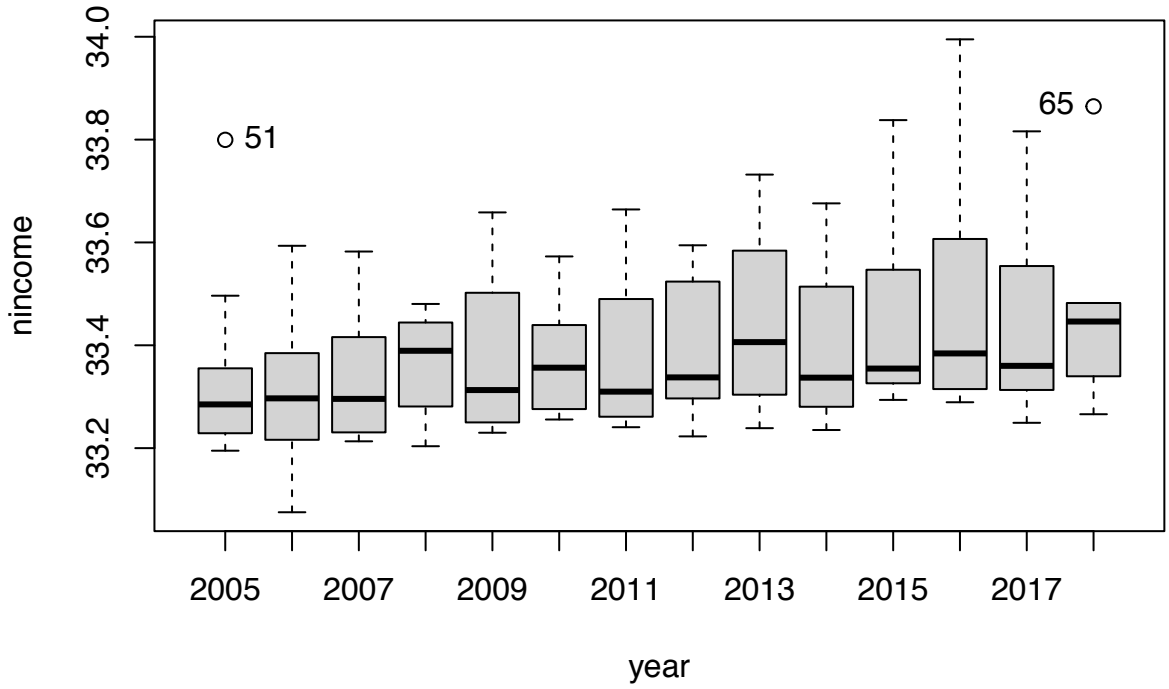



```
#ID up to the first 8, excluding 1 and 10
FE.Kwelfare<-lm(nincome~education_level+yb+region+family_member+year+id,
                data=Kwelfare.pd10,)
coefplot(FE.Kwelfare, predictors="id")
```



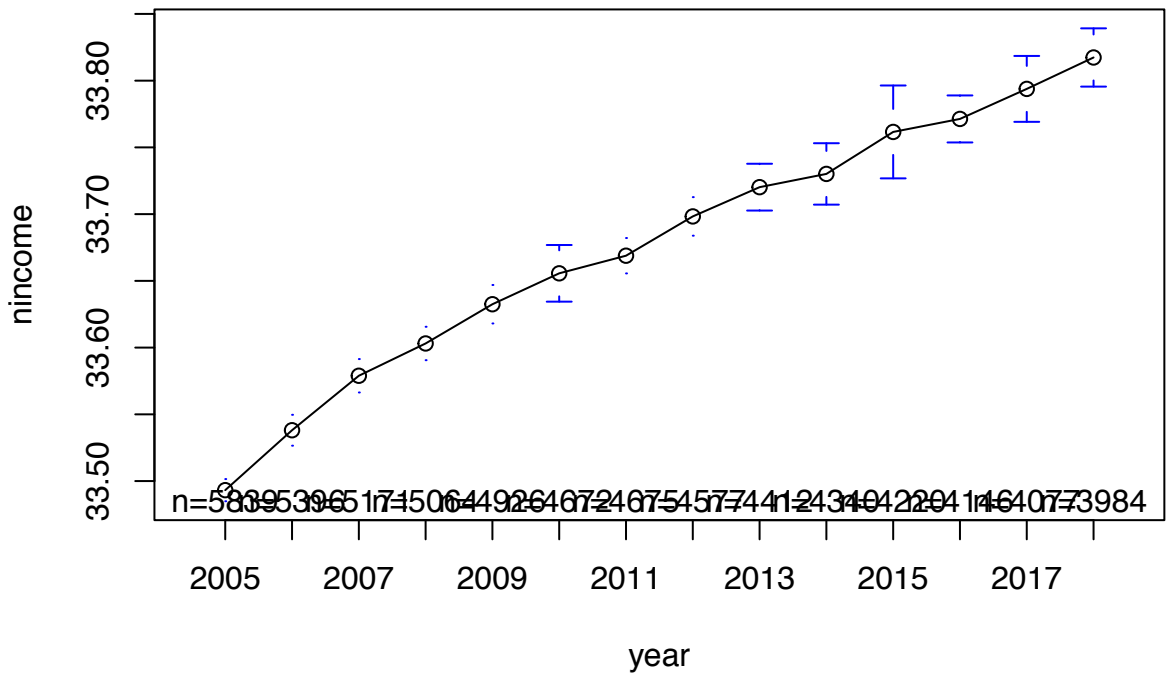
As is shown by the first coefficient plot, there is not hetergeneity over time as the year affects the values of income in a significant way. All the values are above 0, and do not cross zero in any way. This is to be expected, as income and productivity tends to improve year over year, and while it may not be noticeable on the ground level, over time we can see incomes grow. Eligibility for income assistance probably also changes from year to year and as a result those with higher income are eligible for more assistance.

```
# Heterogeneity across time:
#first 10 observations
scatterplot(nincome ~year|id, data=Kwelfare.pd10)
```




```
## [1] "51" "65"
```

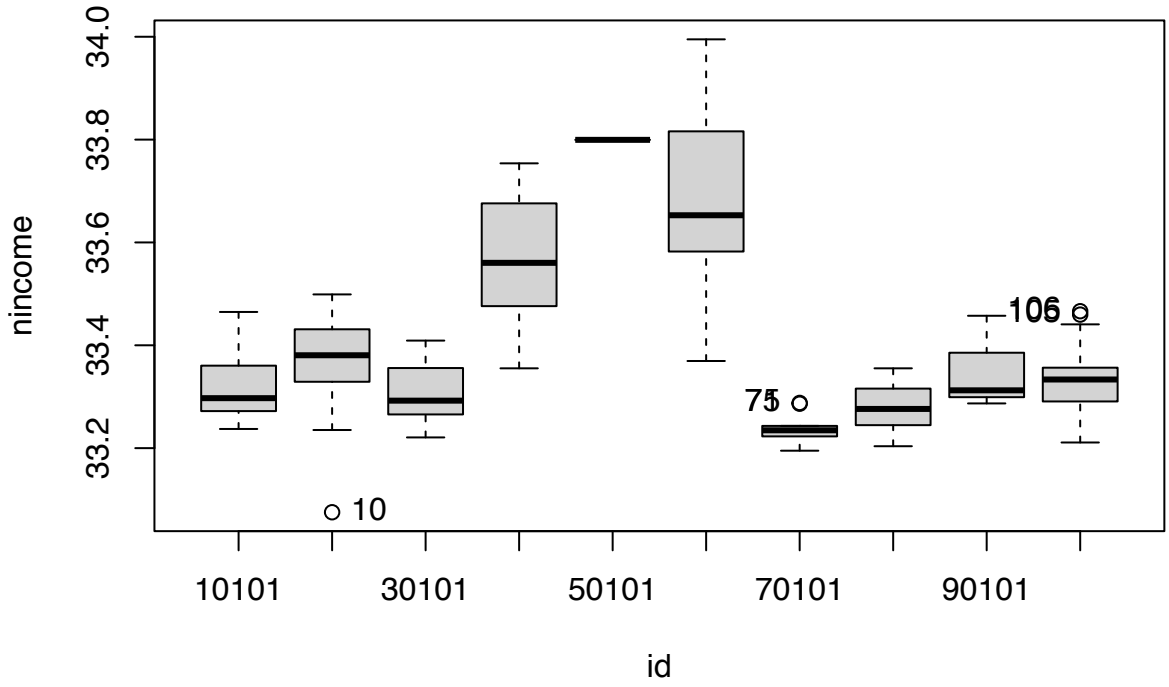
```
#Whole data set
plotmeans(nincome ~year, data = Kwelfare)
```



Income increases year over year as has been established prior to this, as the mean trends upwards overtime. It's more clear in the mean plotting than the scatterplot, because the scatterplot is limited to the first 10 individuals. As a result we can't expect this scatterplot to necessarily be the most accurate representation of the reset of he simple. We also have to consider the the panel is unbalanced as a result there is too much variance in relation to the actual model, and as such we might be able to disregard the scatterplot.

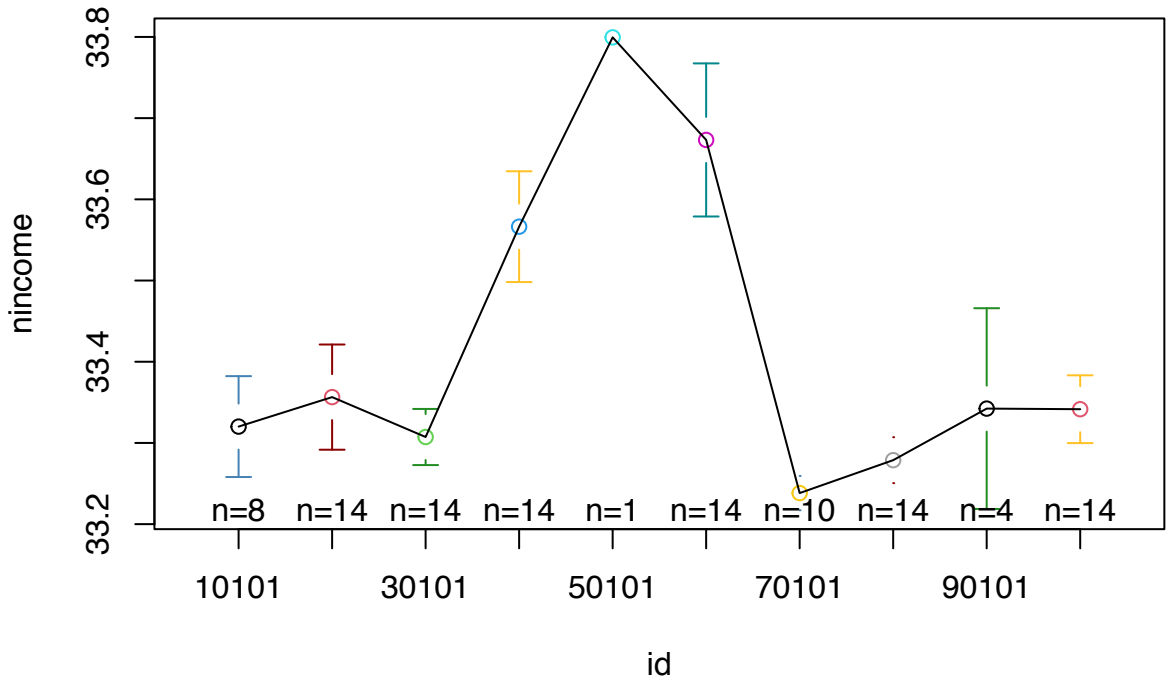
Mean while the mean income moving up over time is evident over the course of the whole sample, and this there is heterogeneity across the income sample.

```
# Heterogeneity across people
scatterplot(nincome ~id|year, data=Kwelfare.pd10)
```




```
## [1] "10"  "71"  "75"  "105" "106"
```

```
plotmeans(nincome ~id, data = Kwelfare.pd10,col =
palette( c( "steelblue", "darkred", "forestgreen", "goldenrod1", "gray67", "turquoise4" )),
barcol =
palette( c( "steelblue", "darkred", "forestgreen", "goldenrod1", "gray67", "turquoise4" ))
, )
```



Not much information is gleaned from checking heterogeneity across people because the differences inherit in individuals varies widely, and as such a 10 person sample from a group of 10,000 people is incredibly limited. Even within this small sample, there is large heterogeneity across individuals, just as there is heterogeneity across income year over year.

3.3 Random Effects Model

```
Kwelfare.random <- plm(nincome~education_level+yb+region+family_member,
                  data=Kwelfare.pd,
                  model="random",)
summary(Kwelfare.random)
```

```
## Oneway (individual) effect Random Effect Model
##    (Swamy-Arora's transformation)
##
## Call:
## plm(formula = nincome ~ education_level + yb + region + family_member,
##     data = Kwelfare.pd, model = "random")
##
## Unbalanced Panel: n = 6724, T = 1-14, N = 65499
##
## Effects:
##                    var std.dev share
## idiosyncratic 0.22823 0.47773  0.74
```

```
## individual    0.08006 0.28295  0.26
## theta:
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.1396  0.5619  0.5887  0.5531  0.5887  0.5887
##
## Residuals:
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -32.130  -0.132  -0.031   0.001   0.075  62.207
##
## Coefficients:
##                    Estimate  Std. Error   z-value  Pr(>|z|)
## (Intercept)     32.48004321  0.02719167 1194.4850 < 2.2e-16 ***
## education_level  0.12258088  0.00284672   43.0603 < 2.2e-16 ***
## yb               0.00520630  0.00027002   19.2811 < 2.2e-16 ***
## region          -0.00748442  0.00181768   -4.1176 3.829e-05 ***
## family_member    0.13159908  0.00259024   50.8057 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    494850
## Residual Sum of Squares: 15276
## R-Squared:      0.96914
## Adj. R-Squared: 0.96913
## Chisq: 5822.65 on 4 DF, p-value: < 2.22e-16
```

With the random model we can see that all values return to a level of high statistical significance. There are no dummy variables, or an embodied intercept per person. Again, education and the amount family members are the most significant factors in predicting a higher level of income among individuals receiving welfare.

3.4 Comparing the Models

```
#Comparing the Within model to the Pooled Model
#Regular within model construction
Kwelfare.within <- plm(nincome~education_level+yb+region+family_member,
                  data=Kwelfare.pd,
                  model="within")
summary(Kwelfare.within)
```

```
## Oneway (individual) effect Within Model
##
## Call:
## plm(formula = nincome ~ education_level + yb + region + family_member,
##     data = Kwelfare.pd, model = "within")
##
## Unbalanced Panel: n = 6724, T = 1-14, N = 65499
##
## Residuals:
##       Min.    1st Qu.     Median    3rd Qu.       Max.
## -30.821736  -0.087303  -0.004982   0.073426  59.747011
##
```

```
## Coefficients:
##                     Estimate  Std. Error t-value Pr(>|t|)
## education_level  0.12021289  0.00577949  20.800   <2e-16 ***
## yb               0.01714002  0.00044425  38.582   <2e-16 ***
## region          -0.00440348  0.00435558  -1.011    0.312
## family_member    0.12399886  0.00344329  36.012   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    14032
## Residual Sum of Squares: 13413
## R-Squared:      0.044083
## Adj. R-Squared: -0.065332
## F-statistic: 677.574 on 4 and 58771 DF, p-value: < 2.22e-16
```

```
#pFtest
pFtest(Kwelfare.within, Kwelfare.pooled)
```

```
##
##  F test for individual effects
##
## data:  nincome ~ education_level + yb + region + family_member
## F = 4.6863, df1 = 6723, df2 = 58771, p-value < 2.2e-16
## alternative hypothesis: significant effects
```

Finally we get around to a regular fixed effects model regression and we find that the region predictor is not very effective for predicting income among those receiving welfare. It also shows that the 10 man sample we were working with earlier was not necessarily accurate when it comes to education level. This may be due to the fact that the 10 indiividuals looked at first were closely grouped in terms of education level and as such differcnes in income between the 10 of them was due to family size and age.

We also look at the F test for individual effects and find that we reject the null and embrace the alternative that there are significant effects going on between the fixed effect model and the pooled model. As such we will work with the within model going forward and compare it to the random test.

```
#Comparing Fixed Within model to the Random Model
phtest(Kwelfare.within, Kwelfare.random)
```

```
##
##  Hausman Test
##
## data:  nincome ~ education_level + yb + region + family_member
## chisq = 1288, df = 4, p-value < 2.2e-16
## alternative hypothesis: one model is inconsistent
```

As evidence by the low p-value here, the Fixed Effects Within model is preferable to the Random Effects model, as the Random effects model is considered inconsistent relative to the Fixed Effects Within model.

3.5 Conclusion for Korean Welfare Model

Overall the Fixed Effects model proved to be the most effective when it came to parsing out the variables that most heavily affect the income of individuals with welfare. It was surprising to see that individuals

with larger families made more money, and that region didn't necessarily matter. We're not the most knowledgeable on Korena geogrpahy but it wouldn't be surprising to see that region doesn't matter if most of the individuals tracked are from cities rather than rural places.

Age also affected income, but not as much, especially since the distribution was almost uniform across the ages from 40-70. Since you have to be within a certain income range in order to receive welfare, it's most likely that those of elderly age have no inherent advantage, or a slight one in the marketplace.

Education was a foregone conclusion as those with degrees can command greater pay due to certain characteristics that having a higher education may imply. As a result, even in the case that such individuals are placed on welfare, they can receive greater pay than their peers on average.

Econ 104L: Group

Project #3 Part 2: Retirement Age Prediction

Constructing Logit and Probit models for Retirment Age

Ye Wang, Omer Abdelrahim, Shane Barry, Yale Yang

Contents

1 Part 1 **1**

1.1 What We Want to Answer With This Model 1

2 Part 2 **1**

2.1 Descriptive Analysis of the Variables 1

3 Part 3 **16**

3.1 Fitting Models 16

3.2 Probit Model 21

3.3 Probit Model Confusion Matrix 23

4 Part 4 **24**

4.1 Probit Prediction Models 24

1 Part 1

1.1 What We Want to Answer With This Model

We study the factors influencing American's retirement. This includes things such as age, race and income in order to find the largest and most accurate predictor of retirment in America.

Of course we can expect age and income to be the foremost predictors, but race and education also should give some invaluable information as to whether an indiviudal will be retired at a certain point in time.

2 Part 2

2.1 Descriptive Analysis of the Variables

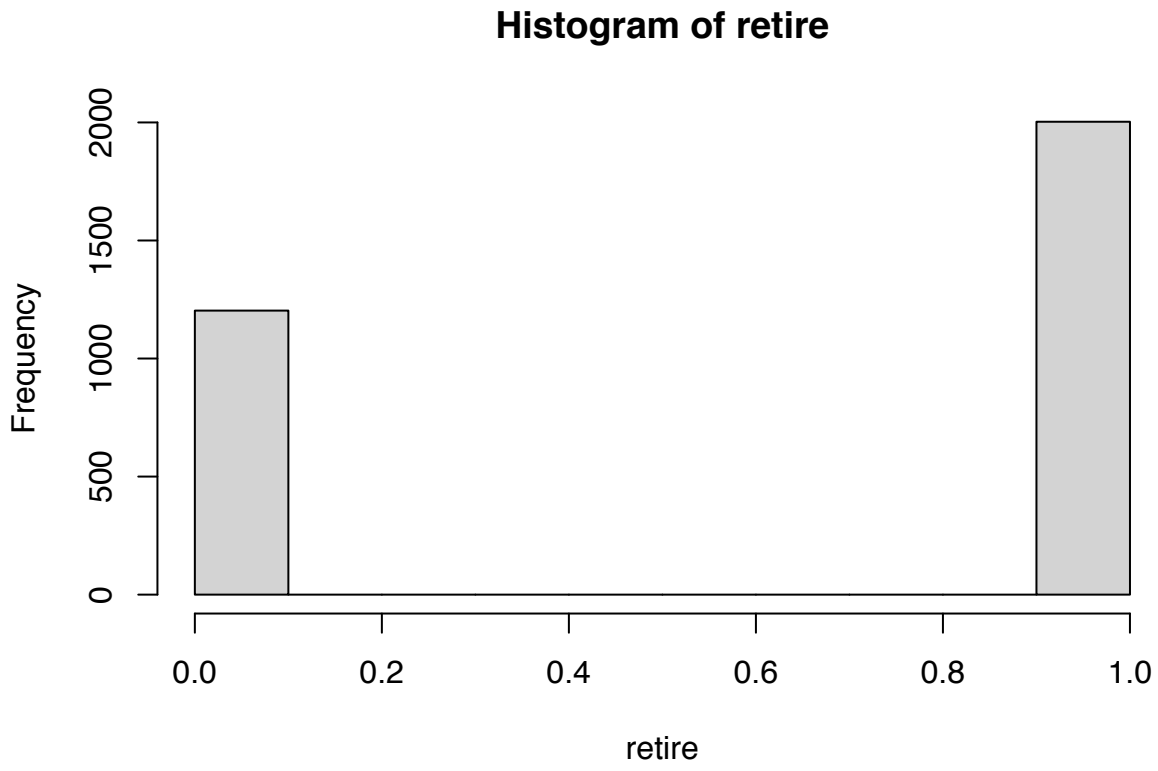
Using data set from the Health and Retirement Study (HRS), wave 5 (2002) collected by the National Institute of Aging.

Totally 8 variables in the data:

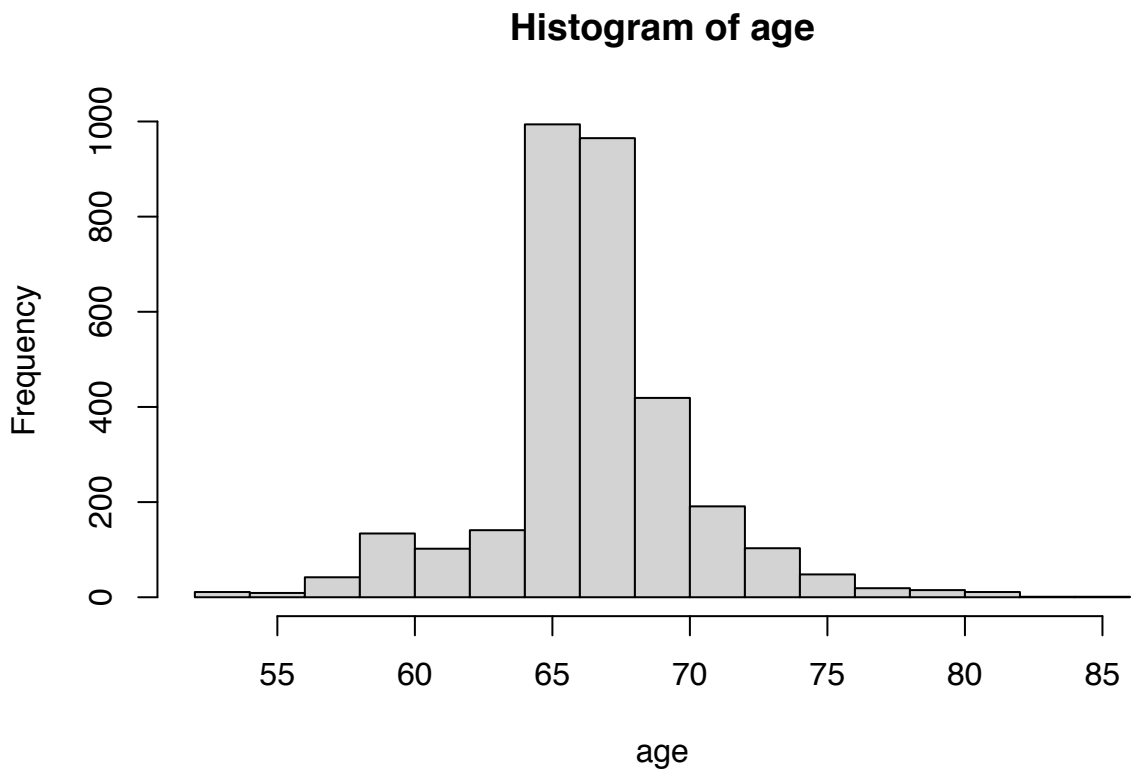
Dependent variable: whether or not a person has become retired (0 or 1).

Independent variables: ins = whether or not people have insurance, age, hstatusg = good health status, hhincome = household income, educyear = education years, married, hisp = Hispanic.

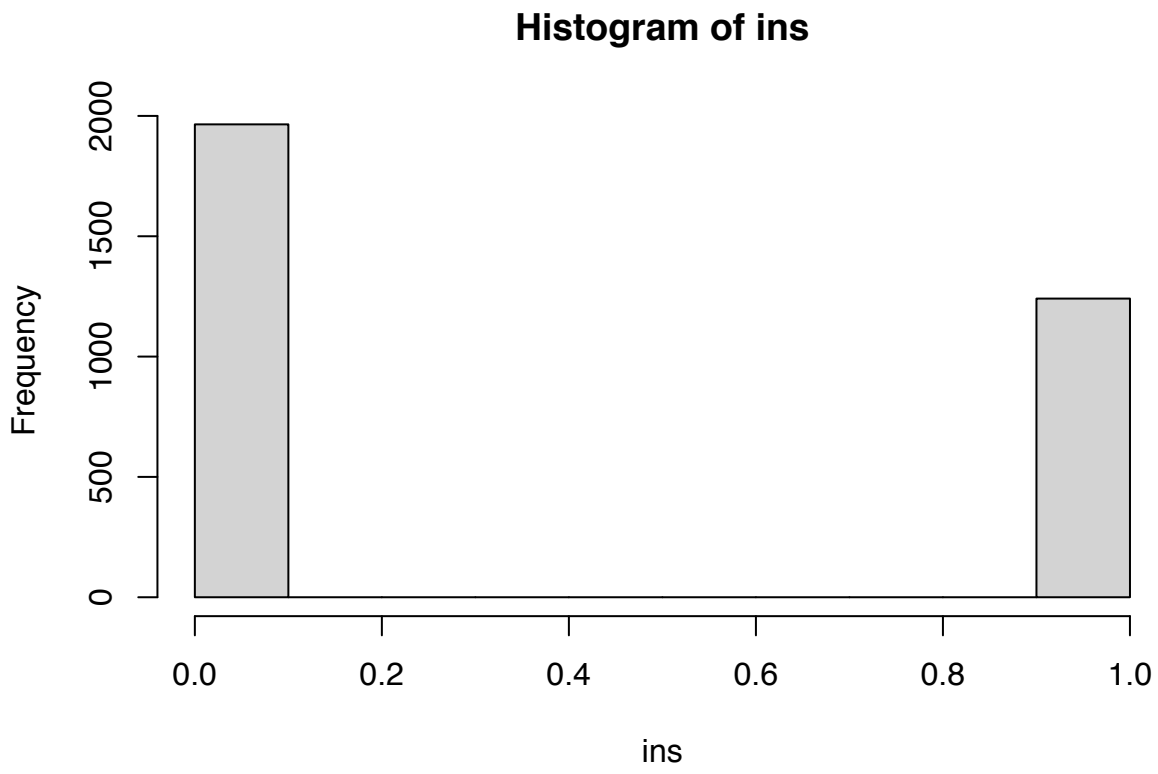
```
mydata<- read.csv("/Users/omerabdelrahim/Downloads/probit_insurance.csv")

attach(mydata)
hist(retire)
```



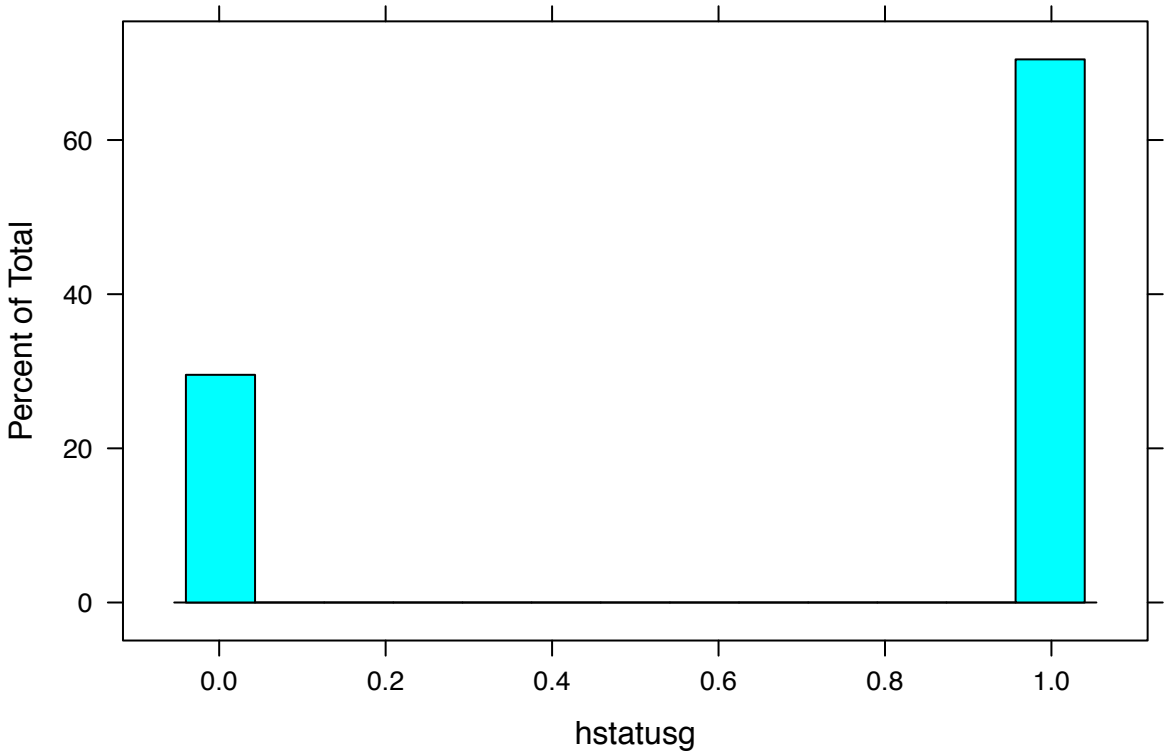

```
hist(age)
```


Histogram of age
Frequency
0
200
400
600
800
1000
55
60
65
70
75
80
85
age

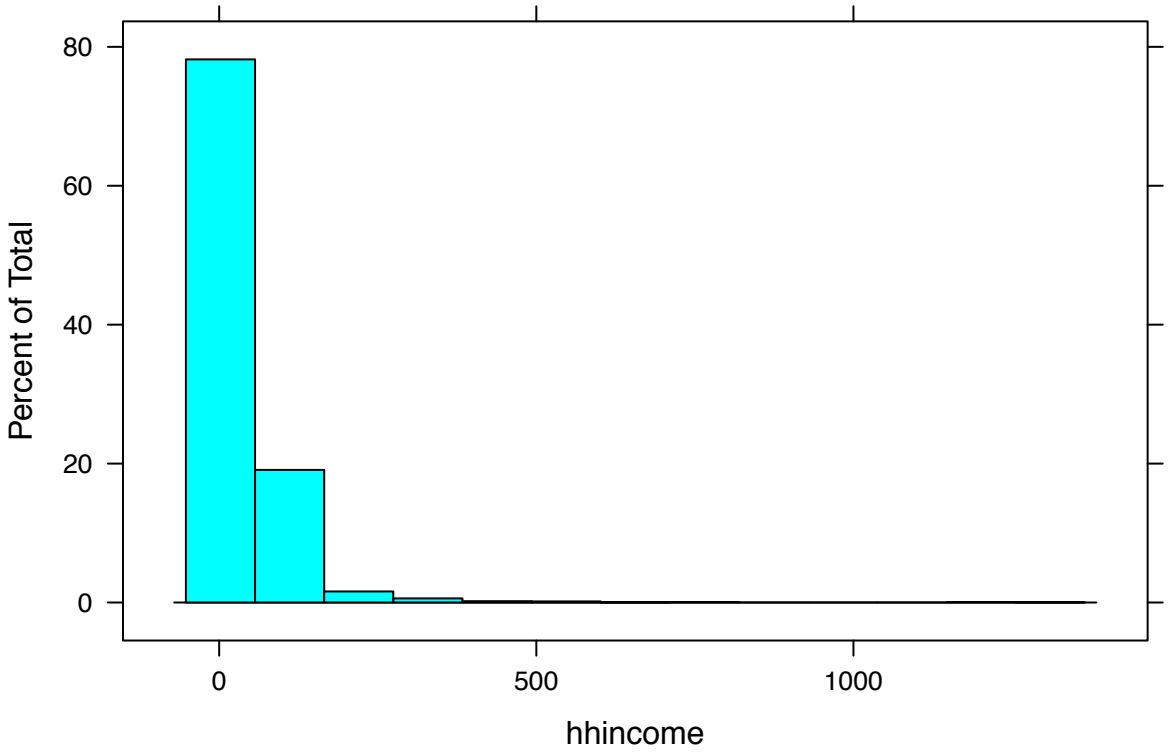
```
hist(ins)
```



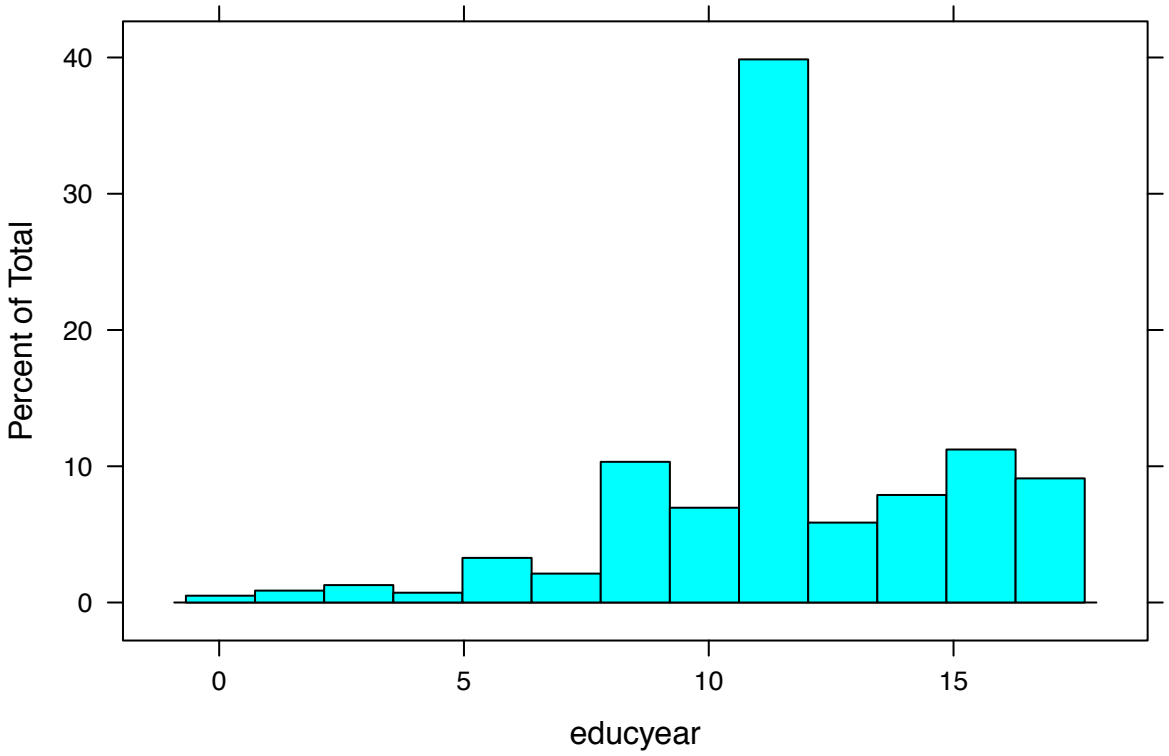

```
histogram(hstatusg)
```



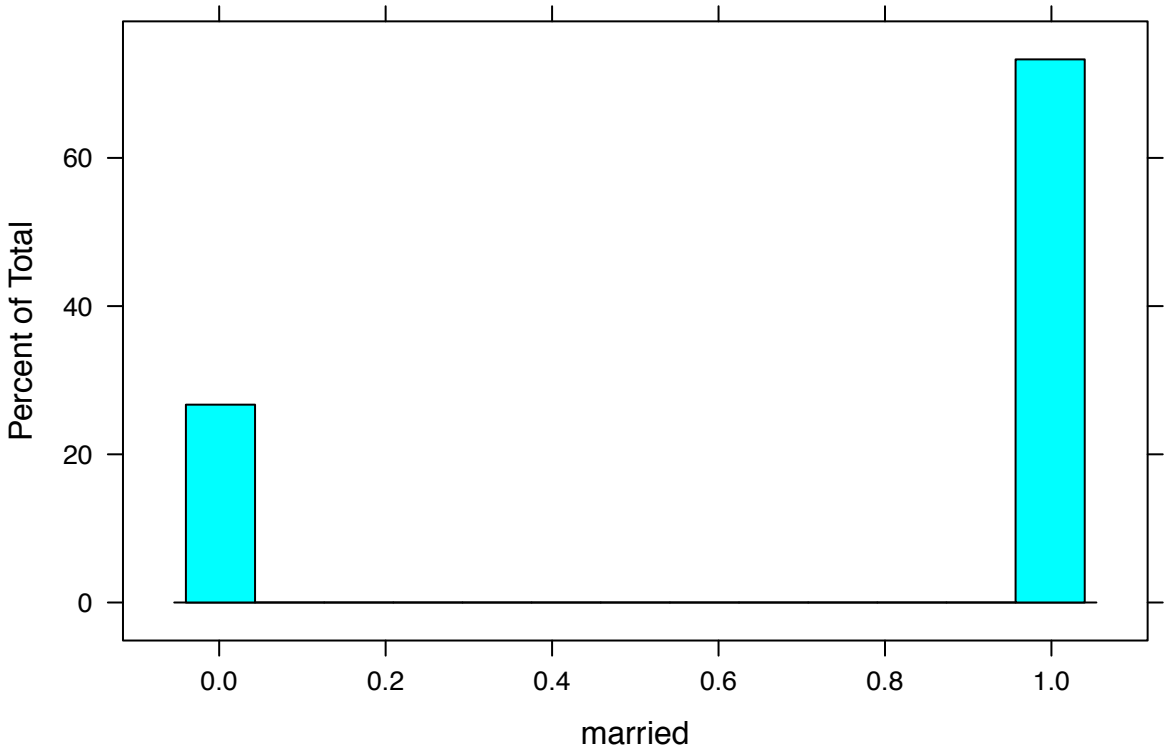

```
histogram(hhincome)
```



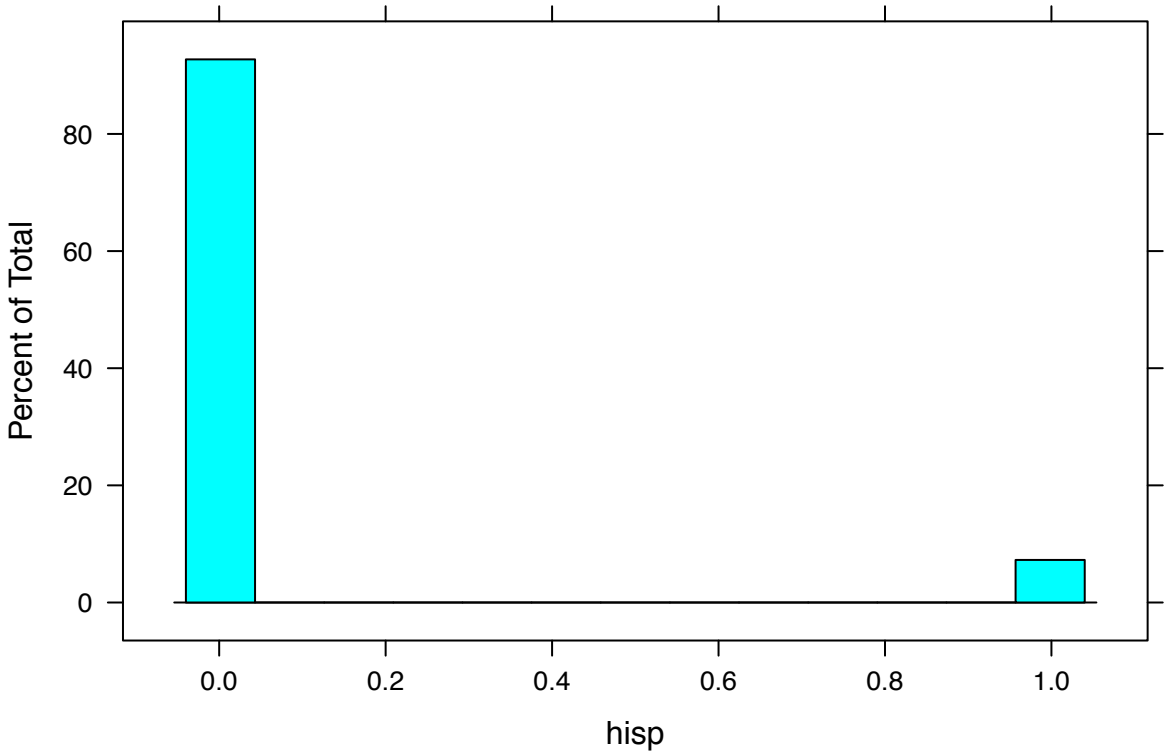

```
histogram(educyear)
```




```
histogram(married)
```



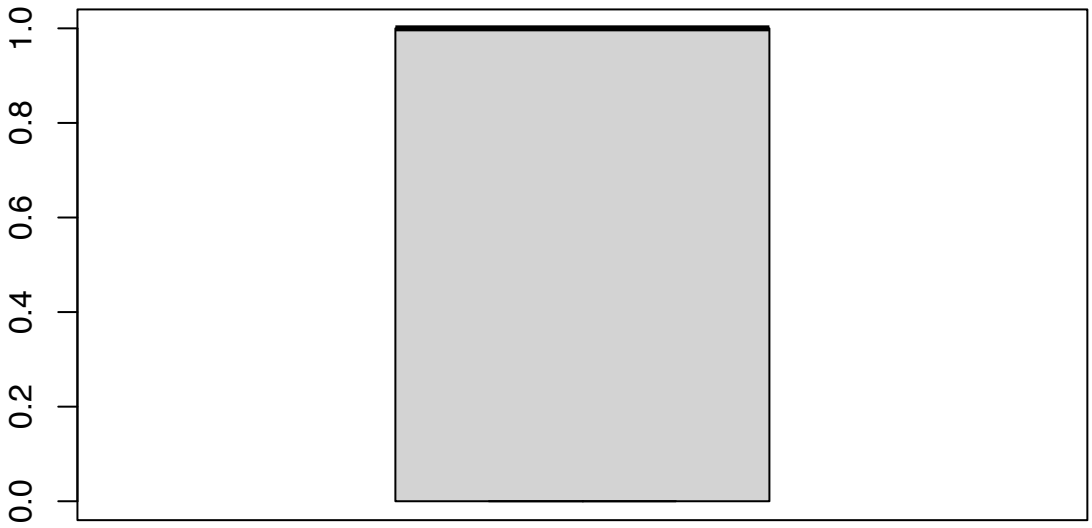

```
histogram(hisp)
```



From the histogram graph, we can overall concluded that in our sample individual groups, around 2/3 people are retired and 1/3 people are not retired. The majority of people (around 78%) have relatively low household income that are lower than 100K. Around 38% percent of people have insurance while the rest 62% of people do not have insurance.

The histogram of education year presents a beta distribution with relatively long left tail. The majority of people are married, having good health status. Around 10% of sample individuals are Hispanic.

```
boxplot(retire)
```

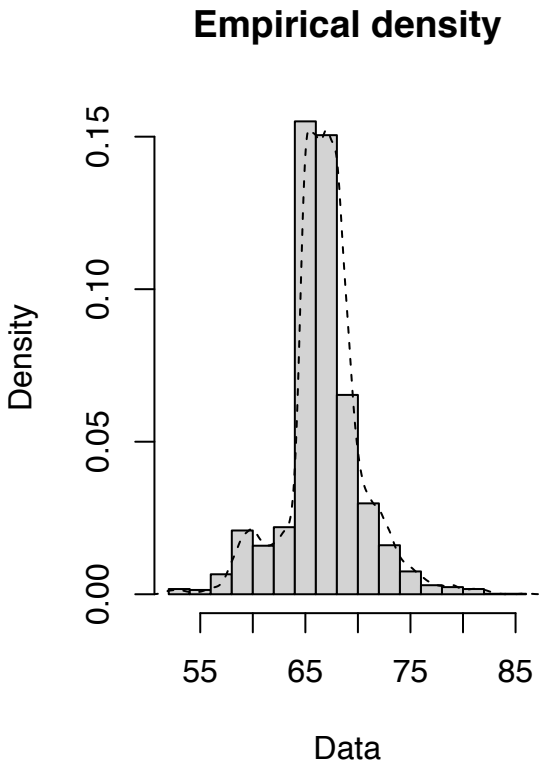


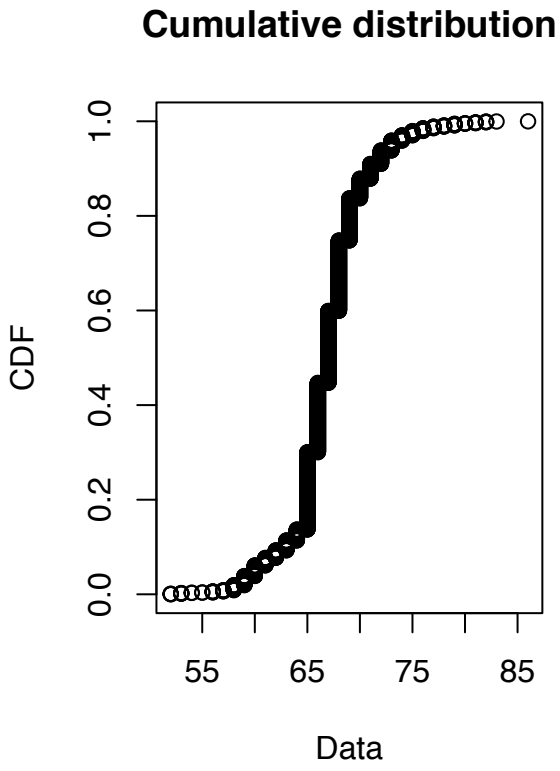

```
fivenum(retire)
```

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

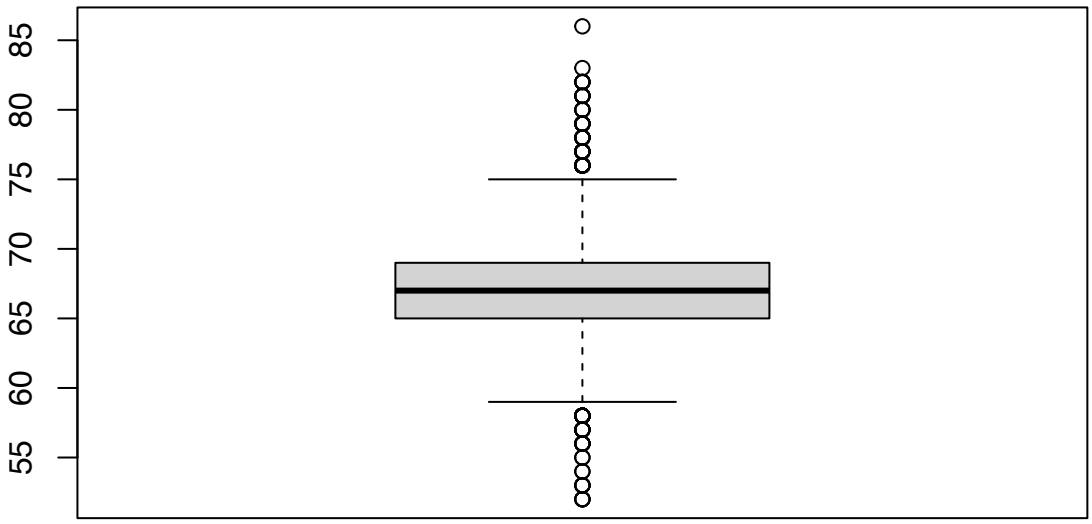
The majority of those in the dataset are retired.

```
plotdist(age,  histo = TRUE, demp = TRUE)
```


Empirical density
Density
0.00
0.05
0.10
0.15
55
65
75
85
Data


Cumulative distribution
CDF
0.0
0.2
0.4
0.6
0.8
1.0
55
65
75
85
Data

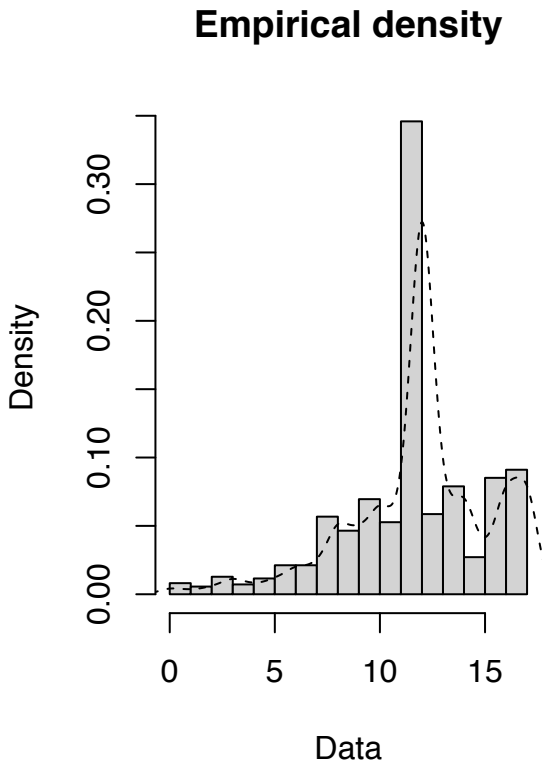
```
boxplot(age)
```

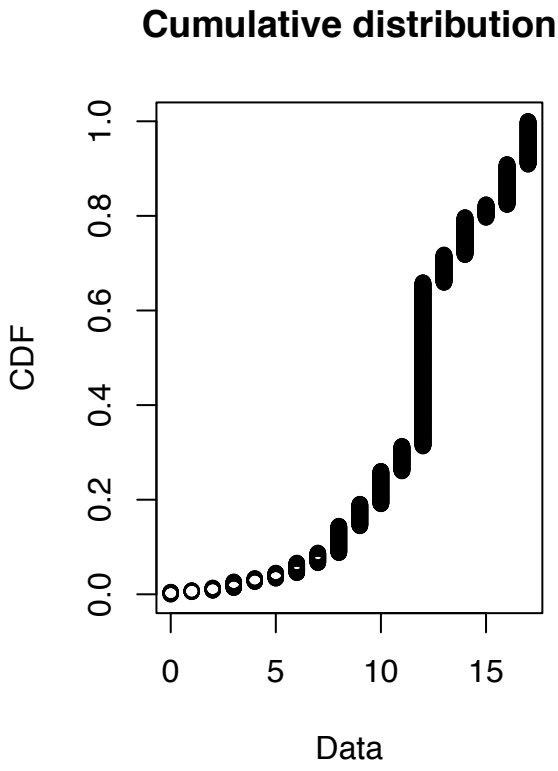



```
fivenum(age)
```

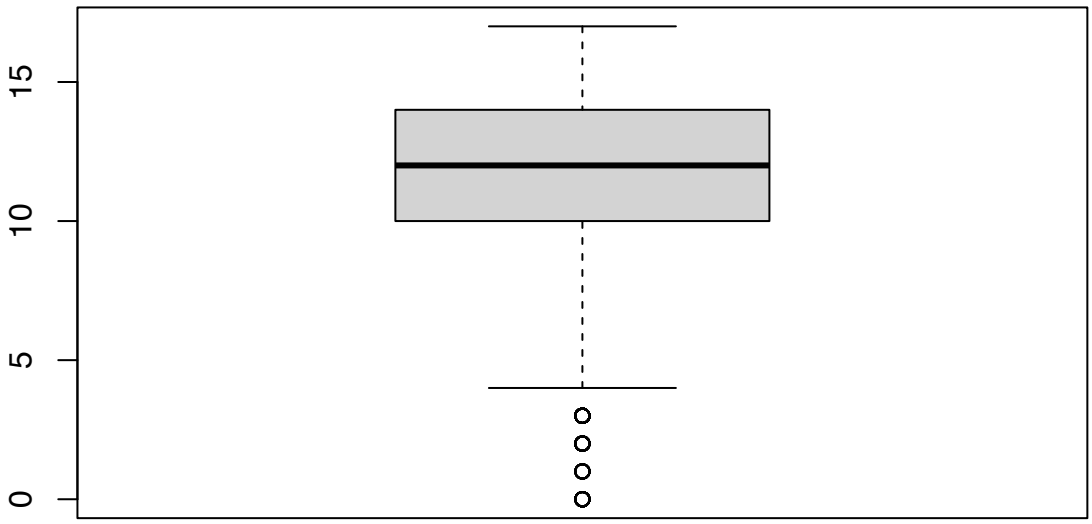
```
## [1] 52 65 67 69 86
```

```
plotdist(educyear,  histo = TRUE, demp = TRUE)
```





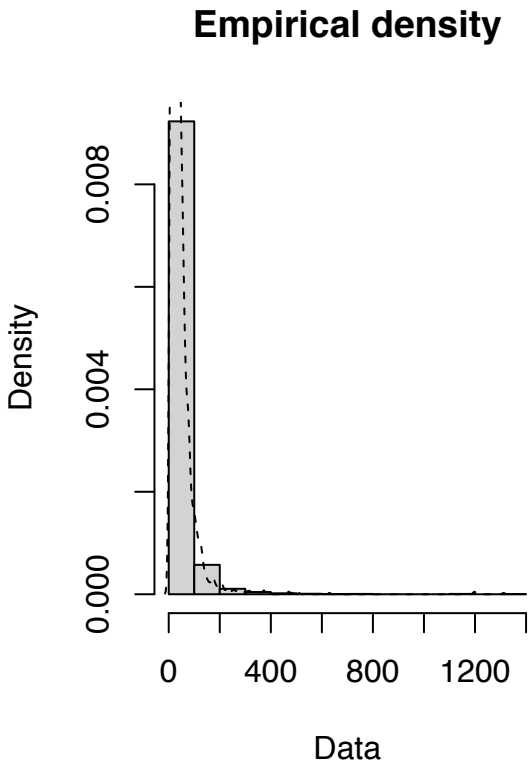

```
boxplot(educyear)
```

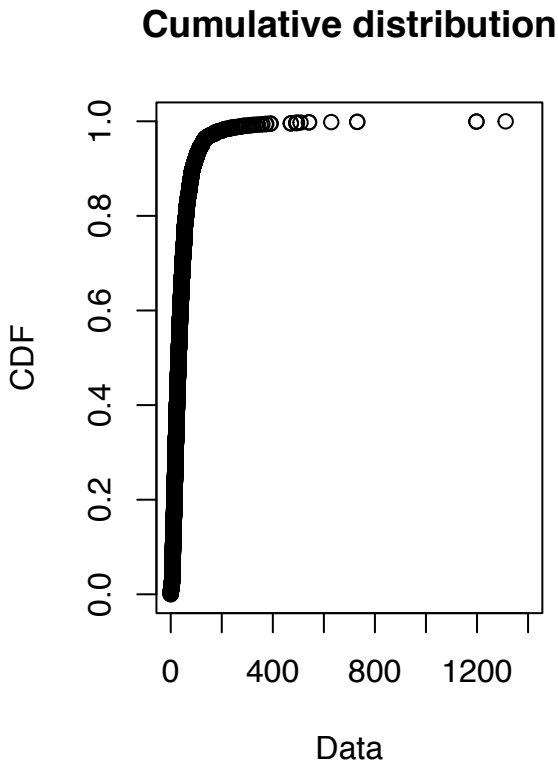



```
fivenum(educyear)
```

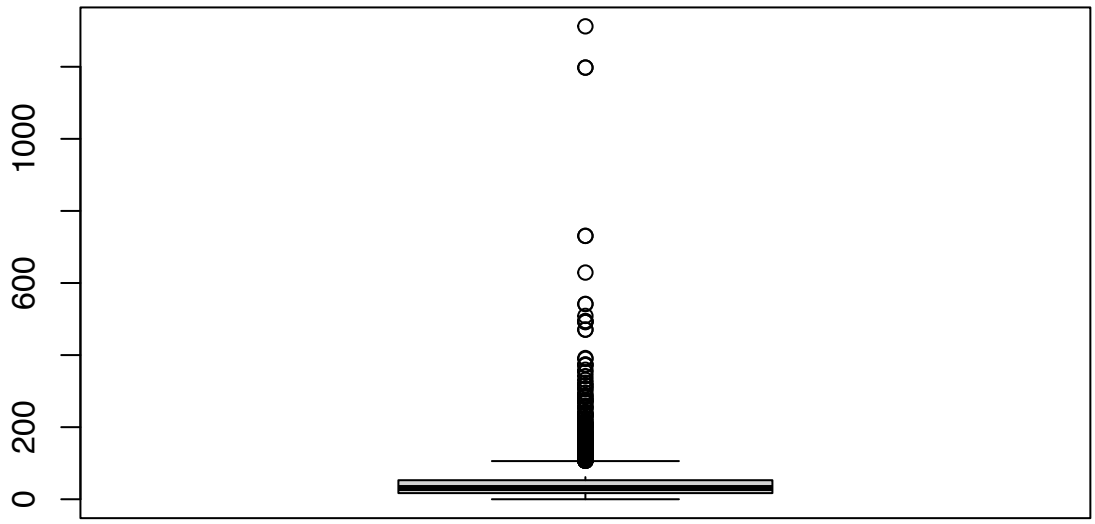
```
## [1]  0 10 12 14 17
```

```
plotdist(hhincome, histo = TRUE, demp = TRUE)
```






```
boxplot(hhincome)
```




```
fivenum(hhincome)
```

```
## [1]    0.000   17.000   31.104   52.800 1312.124
```

This is can also been seen from the individual's age boxplot graph that people approximately normally distributed around age 67.

The mean of education year is 12 years, and most of people's education year are around 10 to 14 years. It is also shown from the boxplot that education year has a little longer left tail.

3 Part 3

3.1 Fitting Models

```
mydata<- read.csv("/Users/omerabdelrahim/Downloads/probit_insurance.csv")

attach(mydata)

# Define variables
Y <- cbind(retire)
X <- cbind(ins, age, hstatusg, hhincome, educyear, married, hisp)

# Descriptive statistics
summary(Y)
```

```
##      retire
##  Min.   :0.0000
##  1st Qu.:0.0000
##  Median :1.0000
##  Mean   :0.6248
##  3rd Qu.:1.0000
##  Max.   :1.0000
```

```
summary(X)
```

```
##       ins              age           hstatusg         hhincome
##  Min.   :0.0000   Min.   :52.00   Min.   :0.0000   Min.   :   0.00
##  1st Qu.:0.0000   1st Qu.:65.00   1st Qu.:0.0000   1st Qu.:  17.00
##  Median :0.0000   Median :67.00   Median :1.0000   Median :  31.10
##  Mean   :0.3871   Mean   :66.91   Mean   :0.7046   Mean   :  45.26
##  3rd Qu.:1.0000   3rd Qu.:69.00   3rd Qu.:1.0000   3rd Qu.:  52.80
##  Max.   :1.0000   Max.   :86.00   Max.   :1.0000   Max.   :1312.12
##     educyear       married           hisp
##  Min.   : 0.0   Min.   :0.000   Min.   :0.00000
##  1st Qu.:10.0   1st Qu.:0.000   1st Qu.:0.00000
##  Median :12.0   Median :1.000   Median :0.00000
##  Mean   :11.9   Mean   :0.733   Mean   :0.07268
##  3rd Qu.:14.0   3rd Qu.:1.000   3rd Qu.:0.00000
##  Max.   :17.0   Max.   :1.000   Max.   :1.00000
```

```
table(Y)
```

```
## Y
##    0    1
## 1203 2003
```

```
table(Y)/sum(table(Y))
```

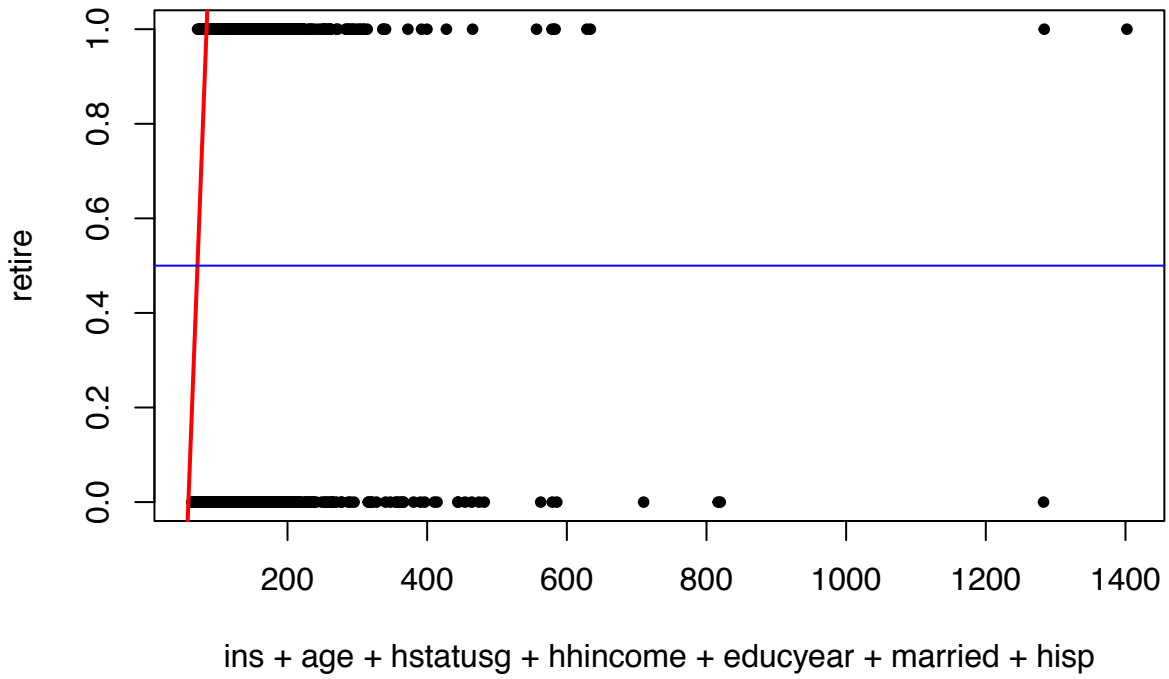
```
## Y
##         0         1
## 0.3752339 0.6247661
```

```
# Linear Probability Model
olsreg = lm(Y ~ X)
summary(olsreg)
```

```
##
## Call:
## lm(formula = Y ~ X)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.1384 -0.4908  0.2216  0.3695  1.0654
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept) -2.2305895  0.1506494 -14.806  < 2e-16 ***
## Xins         0.0384200  0.0171355   2.242   0.0250 *
## Xage         0.0386281  0.0022447  17.208  < 2e-16 ***
## Xhstatusg    0.0347885  0.0188889   1.842   0.0656 .
## Xhhincome   -0.0007153  0.0001330  -5.377 8.10e-08 ***
## Xeducyear    0.0172361  0.0027927   6.172 7.60e-10 ***
## Xmarried     0.0846282  0.0188365   4.493 7.28e-06 ***
## Xhisp       -0.0485429  0.0327036  -1.484   0.1378
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.453 on 3198 degrees of freedom
## Multiple R-squared:  0.1268, Adjusted R-squared:  0.1249
## F-statistic: 66.36 on 7 and 3198 DF,  p-value: < 2.2e-16
```

```
plot(ins+age+hstatusg+hhincome+educyear+married+hisp, retire,pch=20)
abline(olsreg,col ="red", lwd=2)
abline(h=0.5,col="blue")
```




```
confint(olsreg)
```

```
##                     2.5 %        97.5 %
## (Intercept) -2.5259687566 -1.9352102782
## Xins         0.0048221995  0.0720177025
## Xage         0.0342268139  0.0430293974
```

```
## Xhstatusg   -0.0022470086  0.0718239618
## Xhhincome   -0.0009761078 -0.0004544884
## Xeducyear    0.0117604478  0.0227117830
## Xmarried     0.0476953489  0.1215611270
## Xhisp       -0.1126649658  0.0155791776
```

```
polsreg<- predict(olsreg)
summary(polsreg)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -0.1447  0.5366  0.6303  0.6248  0.7234  1.4022
```

```
olsreg.predict <- ifelse(fitted(olsreg) > 0.5, 1, 0)
table(olsreg.predict, Y)
```

```
##                Y
## olsreg.predict    0    1
##              0  415  195
##              1  788 1808
```

```
mean(olsreg.predict == Y)
```

```
## [1] 0.6933874
```

##Logit Model

```
# Logit model coefficients
logit<- glm(Y ~ X, family=binomial (link = "logit"))
summary(logit)
```

```
##
## Call:
## glm(formula = Y ~ X, family = binomial(link = "logit"))
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5567  -1.1443   0.6764   0.9320   2.5179
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.439e+01  8.970e-01 -16.039  < 2e-16 ***
## Xins         1.858e-01  8.428e-02   2.204   0.0275 *
## Xage         2.048e-01  1.335e-02  15.343  < 2e-16 ***
## Xhstatusg    1.122e-01  9.162e-02   1.224   0.2208
## Xhhincome   -3.785e-03  7.597e-04  -4.983 6.26e-07 ***
## Xeducyear    8.454e-02  1.401e-02   6.034 1.60e-09 ***
## Xmarried     4.034e-01  8.958e-02   4.503 6.69e-06 ***
## Xhisp       -2.174e-01  1.586e-01  -1.370   0.1705
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4242.7  on 3205  degrees of freedom
## Residual deviance: 3799.1  on 3198  degrees of freedom
## AIC: 3815.1
##
## Number of Fisher Scoring iterations: 4
```

```
# Logit model odds ratios
exp(logit$coefficients)
```

```
##  (Intercept)         Xins         Xage    Xhstatusg    Xhhincome    Xeducyear
## 5.646557e-07 1.204180e+00 1.227269e+00 1.118709e+00 9.962216e-01 1.088213e+00
##     Xmarried        Xhisp
## 1.496925e+00 8.046246e-01
```

```
confint(logit)
```

```
##                    2.5 %       97.5 %
## (Intercept) -16.17269079 -12.655377218
## Xins          0.02081553   0.351273732
## Xage          0.17902587   0.231365442
## Xhstatusg    -0.06802414   0.291221047
## Xhhincome    -0.00532560  -0.002356619
## Xeducyear     0.05716973   0.112109845
## Xmarried      0.22770699   0.578954375
## Xhisp        -0.52802342   0.094293021
```

```
table(true = Y, pred = round(fitted(logit)))
```

```
##     pred
## true    0    1
##    0  430  773
##    1  223 1780
```

```
# Logit model average marginal effects
LogitM <- mean(dlogis(predict(logit, type = "link")))
LogitM * coef(logit)
```

```
##   (Intercept)          Xins          Xage     Xhstatusg     Xhhincome
## -2.9297183868  0.0378352458  0.0417028383  0.0228428615 -0.0007708696
##     Xeducyear      Xmarried         Xhisp
##  0.0172147843  0.0821493644 -0.0442662259
```

```
plogit<- predict(logit, type="response")
summary(plogit)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.03359 0.53528 0.64601 0.62477 0.74584 0.99074
```

```
logit.pred <- ifelse(fitted(logit) > 0.5, 1, 0)
table(logit.pred, Y)
```

```
##           Y
## logit.pred    0    1
##          0  430  223
##          1  773 1780
```

```
mean(logit.pred == Y)
```

```
## [1] 0.6893325
```

3.2 Probit Model

```
# Probit model coefficients
probit<- glm(Y ~ X, family=binomial (link="probit"))
summary(probit)
```

```
##
## Call:
## glm(formula = Y ~ X, family = binomial(link = "probit"))
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6794  -1.1460   0.6844   0.9387   2.4453
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -8.4767710  0.5109627 -16.590  < 2e-16 ***
## Xins         0.1135930  0.0506818   2.241    0.025 *
## Xage         0.1205241  0.0076105  15.837  < 2e-16 ***
## Xhstatusg    0.0757453  0.0554046   1.367    0.172
## Xhhincome   -0.0020650  0.0004356  -4.740 2.13e-06 ***
## Xeducyear    0.0498821  0.0083675   5.961 2.50e-09 ***
## Xmarried     0.2385886  0.0544462   4.382 1.18e-05 ***
## Xhisp       -0.1454455  0.0959780  -1.515    0.130
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4242.7  on 3205  degrees of freedom
## Residual deviance: 3804.3  on 3198  degrees of freedom
## AIC: 3820.3
##
## Number of Fisher Scoring iterations: 4
```

```
confint(probit)
```

```
##                    2.5 %       97.5 %
## (Intercept) -9.477941532 -7.491744588
## Xins         0.014377543  0.212938548
## Xage         0.105857476  0.135433535
## Xhstatusg   -0.033386506  0.184544146
## Xhhincome   -0.002816422 -0.001321293
## Xeducyear    0.033514046  0.066295779
## Xmarried     0.131813563  0.345280207
## Xhisp       -0.333099820  0.042528311
```

```
# Probit model average marginal effects
ProbitM <- mean(dnorm(predict(probit, type = "link")))
ProbitM * coef(probit)
```

```
##   (Intercept)          Xins          Xage     Xhstatusg     Xhhincome
## -2.8691726600  0.0384483592  0.0407943759  0.0256378583 -0.0006989496
##     Xeducyear      Xmarried         Xhisp
##  0.0168838325  0.0807562126 -0.0492296254
```

```
# Percent correctly predicted values
table(true = Y, pred = round(fitted(probit)))
```

```
##     pred
## true    0    1
##    0  425  778
##    1  215 1788
```

```
#Predict probit model
pprobit<- predict(probit, type="response")
summary(pprobit)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0229  0.5350  0.6420  0.6235  0.7402  0.9971
```

```
probit.pred <- ifelse(fitted(probit) > 0.5, 1, 0)
table(probit.pred, Y)
```

```
##            Y
## probit.pred    0    1
##           0  425  215
##           1  778 1788
```

```
mean(probit.pred == Y)
```

```
## [1] 0.6902682
```

(1) Binary outcome model coefficients interpretation:

insured individuals (in comparison to non-insured individuals), older individuals, individuals with good health status, higher household income, higher education, married are more likely to retired; people with less household income and Hispanic people are less likely being retired.

(2) Marginal effects interpretation for both probit and logit:

insured individuals are 3.8% more likely to become retired (in comparison with those that are not insured). One year older will bring around 4.1% of more possibility of retirement. Married people can have For each individual one more year in education are 8% more likely to have retirement. Hispanics are 4% to 5% less likely to become retired than non-Hispanics.

(3) Unlike the coefficients are different, the marginal effects are almost identical in the three models. Also, the sign of the coefficients and marginal effects are same for both the logit and probit models.

The average of predicted probabilities for having insurance is about 62.3% which is similar to the actual frequency for becoming retired.

The logit and probit models correctly predict around 69% of the values and the rest are misclassified.

Because the probit model has the highest accuracy around 69.03%, therefore we choose the probit model as the most preferable model.

3.3 Probit Model Confusion Matrix

```
library(caret)

inTraining <- createDataPartition(mydata$retire, p = .75, list = FALSE)

training <- mydata[ inTraining,]
testing <- mydata[-inTraining,]
train_control <- trainControl(method = "cv",
number = 5)
logit_model <- train(as.factor(retire) ~ .,
data = training,
method = "glm",
family = "binomial",
trControl = train_control)
# Predict (probabilities) using the testing data
pred_ins = predict(logit_model, newdata = testing)
# Evaluate performance
confusionMatrix(data=pred_ins, reference=as.factor(testing$retire))
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0 143  50
##          1 176 432
##
##                Accuracy : 0.7179
##                  95% CI : (0.6853, 0.7488)
##     No Information Rate : 0.6017
##     P-Value [Acc > NIR] : 4.485e-12
##
##                   Kappa : 0.3692
##
```

```
##  Mcnemar's Test P-Value : < 2.2e-16
##
##             Sensitivity : 0.4483
##             Specificity : 0.8963
##          Pos Pred Value : 0.7409
##          Neg Pred Value : 0.7105
##              Prevalence : 0.3983
##          Detection Rate : 0.1785
##    Detection Prevalence : 0.2409
##       Balanced Accuracy : 0.6723
##
##        'Positive' Class : 0
##
```

From the confusion metrix we could see that the provit model has a relatively high accuracy, which is around 70% percent. Balanced accuracy is reliable as well.

4 Part 4

4.1 Probit Prediction Models

```
#initial probit prediction model
mydata<- read.csv("/Users/omerabdelrahim/Downloads/probit_insurance.csv")

attach(mydata)
probit<- glm(retire ~ ins+age+hstatusg+hhincome+educyear+married+hisp, family=binomial (link="probit"))

predict(probit, data.frame(age = 67, ins = 1, hhincome = 31.104, educyear = 12, hstatusg = 1, married =
```

```
##         1
## 0.7124748
```

Using all average data from the dataset, we could see that the predicted possibility of retirement is around 71.24%.

```
# Case 1: Average individual 50 years old
mydata<- read.csv("/Users/omerabdelrahim/Downloads/probit_insurance.csv")

attach(mydata)
probit<- glm(retire ~ ins+age+hstatusg+hhincome+educyear+married+hisp, family=binomial (link="probit"))

predict(probit, data.frame(age = 50, ins = 1, hhincome = 31.104, educyear = 12, hstatusg = 1, married =
```

```
##          1
## 0.06833837
```

In this case, we could see that an 50 years old, insured, married, non-Hispanic individual with average household income, good health condition, has about 6.8% possibility of getting retired. This possibility is very small, because age 50 is still very far from the legal retirement age (65 to 67) to receive full retirement benefits. Therefore most people will not get retired in their age of 50.

```
# Case 2: Average individual 70 years old
mydata<- read.csv("/Users/omerabdelrahim/Downloads/probit_insurance.csv")

attach(mydata)
probit<- glm(retire ~ ins+age+hstatusg+hhincome+educyear+married+hisp, family=binomial (link="probit"))

predict(probit, data.frame(age = 70, ins = 1, hhincome = 31.104, educyear = 12, hstatusg = 1, married =
```

```
##         1
## 0.8217883
```

In this case, we could see that an 70 years old, insured, married, non-Hispanic individual with average household income, education year, and good health condition, has about 6.8% possibility of getting retired. Compared with case 1, the only changed variable is the age. The possibility of retirement has increased hugely from 6.8% to 82%. This is because this individual has already over 65 years old/legal retured age and has a big possibility of getting retired.

```
#case 3: 5 years' more of education than average 12 years
mydata<- read.csv("/Users/omerabdelrahim/Downloads/probit_insurance.csv")

attach(mydata)
probit<- glm(retire ~ ins+age+hstatusg+hhincome+educyear+married+hisp, family=binomial (link="probit"))

predict(probit, data.frame(age = 67, ins = 1, hhincome = 31.104, educyear = 17, hstatusg = 1, married =
```

```
##         1
## 0.7910413
```

In this case, we could see that insured, married, non-Hispanic individual with average sample age 67, household income, good health condition, and extra 5 years of education has about 79% possibility of getting retired. Compared with the initial prediction case, The possibility of retirement has increased about 8%. This shows a possible relationship that higher education/more education years leads to bigger possibility of retirement when reaching age 67, compared with others with less education years.

```
#case 4: Unmarried individual
mydata<- read.csv("/Users/omerabdelrahim/Downloads/probit_insurance.csv")

attach(mydata)
probit<- glm(retire ~ ins+age+hstatusg+hhincome+educyear+married+hisp, family=binomial (link="probit"))

predict(probit, data.frame(age = 67, ins = 1, hhincome = 31.104, educyear = 17, hstatusg = 1, married =
```

```
##         1
## 0.7161531
```

With other conditions stay same, an unmarried individual has a slightly more possibility to get retired than a married one, about 0.37%. This can potentially because that married individual may need more income for household than single individual, so they will delay their retirement.