Predicting Total Wealth: A Predictive Analysis Using the 1991 SIPP Data

Xueshan (Kevin) Peng

Introduction

Loading and Inspecting the Data

Let's take a look at the first 6 rows of the data.

```
data <- read.table('data_tr.txt', head = T)[,-1]
head(data)
```

```
##       tw  ira e401  nifa   inc hmort   hval hequity educ male twoearn nohs hs
## 1  53550    0    0   100 28146 60150  69000    8850   12    0       0    0  1
## 2 124635    0    0 61010 32634 20000  78000   58000   16    0       0    0  0
## 3 192949 1800    0  7549 52206 15900 200000  184100   11    1       1    1  0
## 4   -513    0    0  2487 45252     0      0       0   15    0       1    0  0
## 5 212087    0    0 10625 33126 90000 300000  210000   12    0       0    0  1
## 6  24400    0    0  9000 76860 99600 120000   20400   15    0       1    0  0
##   smcol col age fsize marr
## 1     0   0  31     5    1
## 2     0   1  52     5    0
## 3     0   0  50     3    1
## 4     1   0  28     4    1
## 5     0   0  42     3    0
## 6     1   0  49     6    1
```

We can see that the data is in good shape, where categorical variables are already transformed into dummy variables. We can also see that there exists multi-collinearity between education levels.

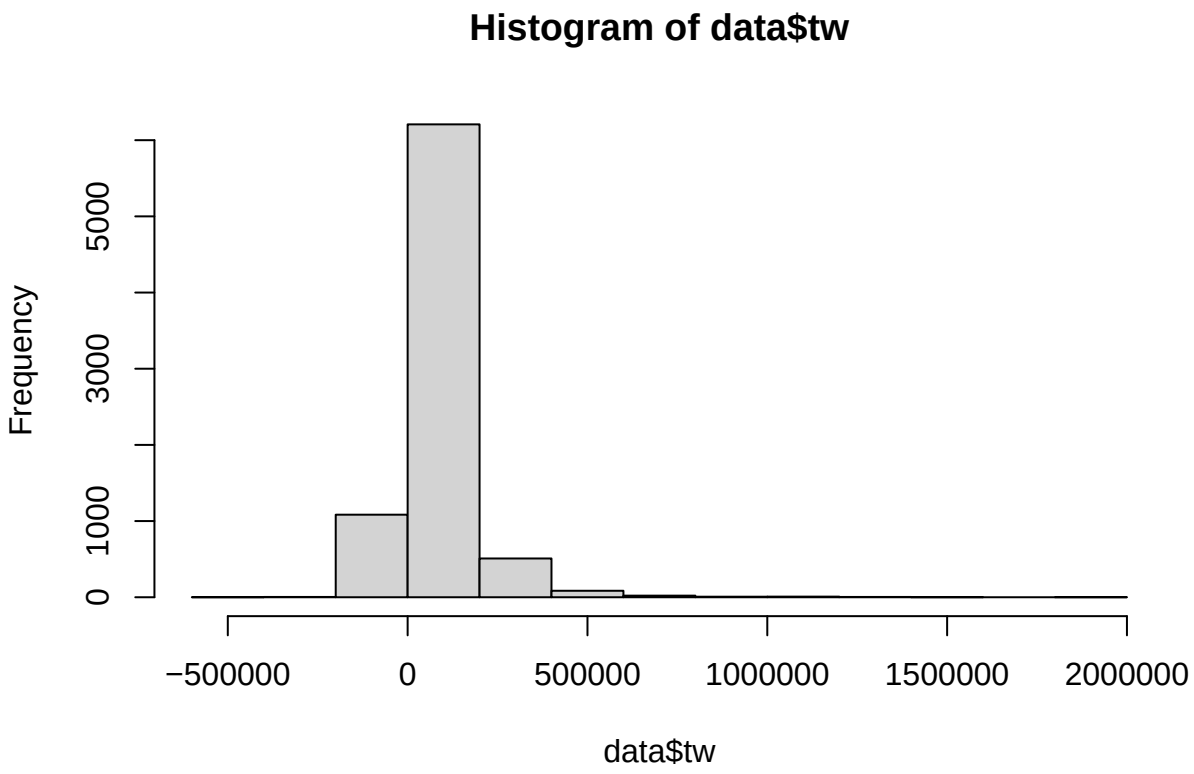
```
summary(data)
```

```
##        tw                ira              e401             nifa
##  Min.   :-502302   Min.   :     0   Min.   :0.0000   Min.   :      0
##  1st Qu.:   3246   1st Qu.:     0   1st Qu.:0.0000   1st Qu.:    200
##  Median :  25225   Median :     0   Median :0.0000   Median :   1687
##  Mean   :  63629   Mean   :  3471   Mean   :0.3714   Mean   :  13611
##  3rd Qu.:  82173   3rd Qu.:     0   3rd Qu.:1.0000   3rd Qu.:   8875
##  Max.   :1887115   Max.   :100000   Max.   :1.0000   Max.   :1425115
##       inc              hmort             hval           hequity
##  Min.   :    -9   Min.   :     0   Min.   :     0   Min.   :-40000
##  1st Qu.: 19413   1st Qu.:     0   1st Qu.:     0   1st Qu.:     0
##  Median : 31575   Median :  8000   Median : 50000   Median : 10000
```

```
##  Mean   : 37177   Mean   : 30207   Mean   : 63965   Mean   : 33757
##  3rd Qu.: 48615   3rd Qu.: 52000   3rd Qu.: 95000   3rd Qu.: 48000
##  Max.   :242124   Max.   :150000   Max.   :300000   Max.   :300000
##       educ            male            twoearn            nohs
##  Min.   : 1.0   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:12.0   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000
##  Median :12.0   Median :0.0000   Median :0.0000   Median :0.0000
##  Mean   :13.2   Mean   :0.2018   Mean   :0.3808   Mean   :0.1277
##  3rd Qu.:15.0   3rd Qu.:0.0000   3rd Qu.:1.0000   3rd Qu.:0.0000
##  Max.   :18.0   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
##        hs              smcol             col              age
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :25.00
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:32.00
##  Median :0.0000   Median :0.0000   Median :0.0000   Median :40.00
##  Mean   :0.3819   Mean   :0.2422   Mean   :0.2482   Mean   :41.08
##  3rd Qu.:1.0000   3rd Qu.:0.0000   3rd Qu.:0.0000   3rd Qu.:48.00
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :64.00
##      fsize            marr
##  Min.   : 1.00   Min.   :0.0000
##  1st Qu.: 2.00   1st Qu.:0.0000
##  Median : 3.00   Median :1.0000
##  Mean   : 2.87   Mean   :0.6075
##  3rd Qu.: 4.00   3rd Qu.:1.0000
##  Max.   :13.00   Max.   :1.0000
```

The variables **ira**, **nohs**, **smcol**, **col**, and **male** exhibited a value of 0 at the 3rd quantile. They are therefore susceptible to outliers, with a significant number of data points taking on the value of 0. It should also be noted that most observations are created by female participants.

```
hist(data$tw)
```



It is obvious that there are outliers with enormous health, and we should take caution in our subsequent analysis.

Testing and Removing Multi-collinearity

Let's test whether removing different educational level predictors affect my model's performance, gauged by (MSPE). For simplicity sake, I did not use k-fold cross validation.

```
k <- 10
set.seed(123)
rand <- sample(nrow(data), floor(nrow(data)/k))
train <- setdiff(c(1:nrow(data)), rand)
y_rand <- data$tw[rand]

regnohs <- lm(tw ~ 1 + hs + smcol + col, data = data[train,])
reghs <- lm(tw ~ 1 + nohs + smcol + col, data = data[train,])
regsmcol <- lm(tw ~ 1 + nohs + hs + col, data = data[train,])
regcol <- lm(tw ~ 1 + nohs + hs + smcol, data = data[train,])

prnohs <- predict(regnohs, newdata = data[rand,])
prhs <- predict(reghs, newdata = data[rand,])
prsmcol <- predict(regsmcol, newdata = data[rand,])
prcol <- predict(regcol, newdata = data[rand,])

MSEnohs <- mean((y_rand-prnohs)^2)
```

```r
MSEhs <- mean((y_rand-prhs)^2)
MSEsmcol <- mean((y_rand-prsmcol)^2)
MSEcol <- mean((y_rand-prcol)^2)

c(MSEnohs, MSEhs, MSEsmcol, MSEcol)
```

```
## [1] 11322526335 11322526335 11322526335 11322526335
```

No difference in performance is found between removing different terms for multi-collinearity. For interpretability, we choose to remove **hs** and **hequity**.