Final Project

Examining the Influence of Personal Financial and Social Factors on Total Wealth

Introduction

Background

The value of all the valuable assets that a person, group of people, business, or nation owns is measured as wealth. The total market worth of all held tangible and intangible assets is calculated, and all debts are then subtracted. Wealth is essentially the accumulation of limited resources.

When certain individuals, groups, or nations are able to amass a large quantity of expensive items or resources, they are said to be wealthy. Income and wealth can be compared since income is a flow whereas wealth is a stock that can be seen either in absolute or relative terms.

The Survey of Income and Program Participation (SIPP) began gathering household information on individual total wealth as well as other financial and social factors in 1992. The goal of this study is to use these financial and social variables to build a model that successfully predicts total wealth in order to determine the most important variables and their impact on total wealth.

Data

7,933 observations and 18 variables make up the data. Data from households of people in the sample range in age from 25 to 64. The person who serves as the household reference person makes up the observation units, and none of the individuals are self-employed.

Total wealth (in US dollars) is the variable to be predicted. Total wealth is equal to net financial assets, including Individual Retirement Account (IRA) and 401(k) assets, plus home equity and the value of businesses, real estate, and cars.

Among the predictor factors are;

Retirement-related variables (features)

Individual Retirement Account (IRA), expressed in US dollars. • e401: if you qualify for a 401(k), 1; otherwise, 0 financial aspects (variables): Non-401(k) Financial Assets (in US dollars): • income (in US dollars). Home ownership-related variables (features): Mortgage on a residence, expressed in US dollars. • HVAL (in US dollars): home value • Hequity is the worth of a home less the mortgage. Additional covariates (features) include: (in years). • Male: If male, 1; if not, 0. • twoearn: 1 if the household has two earners, 0 otherwise.

Objective

In order to develop a model that predicts each person's total wealth, the goal of this study is to examine the link between the variables provided in the data tr.txt dataset and the 'tw' or total wealth.

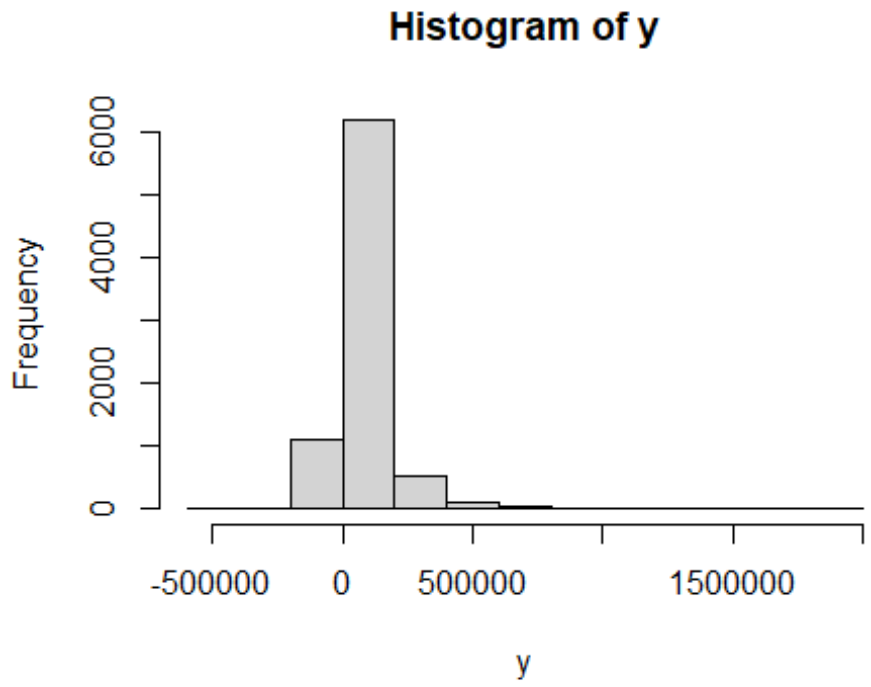
Ordinary Least Square Regression, OLS Cross Validation, and K-fold Cross Validation are the techniques used to calculate the MSPE of each model and assess correctness.

Cross validation in Stepwise, Lasso, and Ridge . K-fold cross validation is used to identify the best correct model.

Statistical analyses

Investigating outcome variable

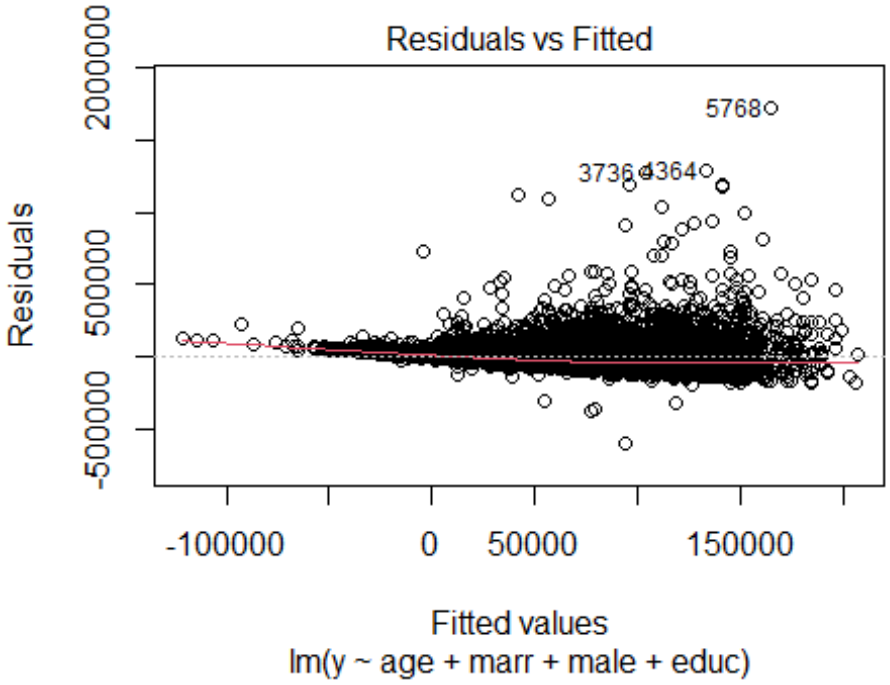
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -502302    3246   25225   63629   82173 1887115
```

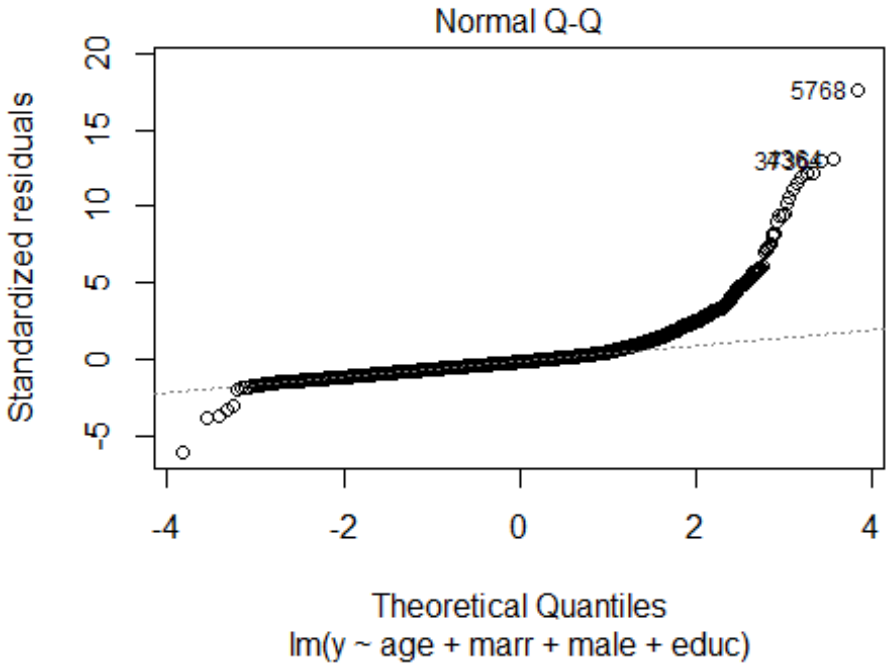


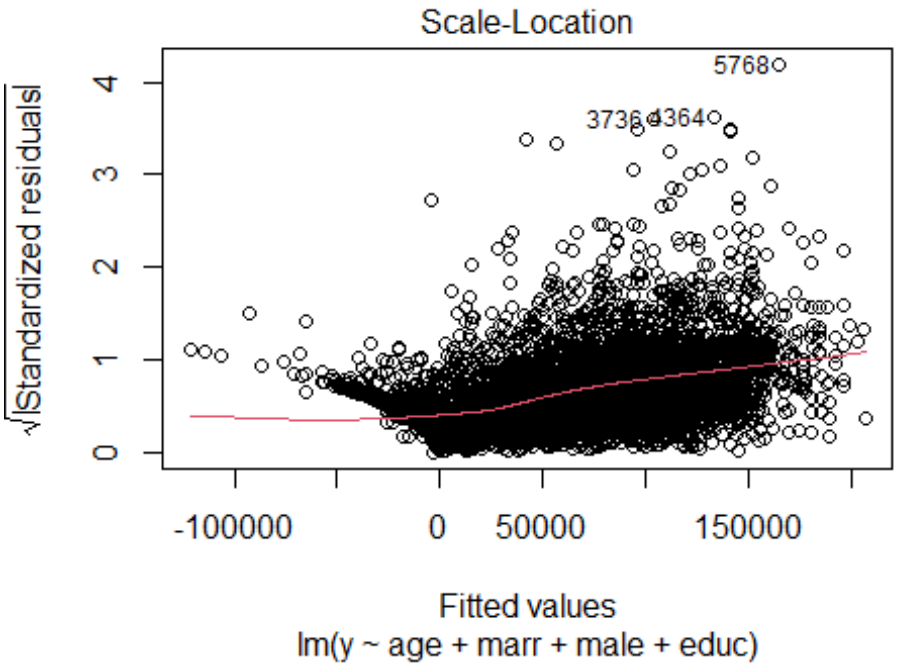
The mean of the response variable is 63629 with the maximum being 1887115.

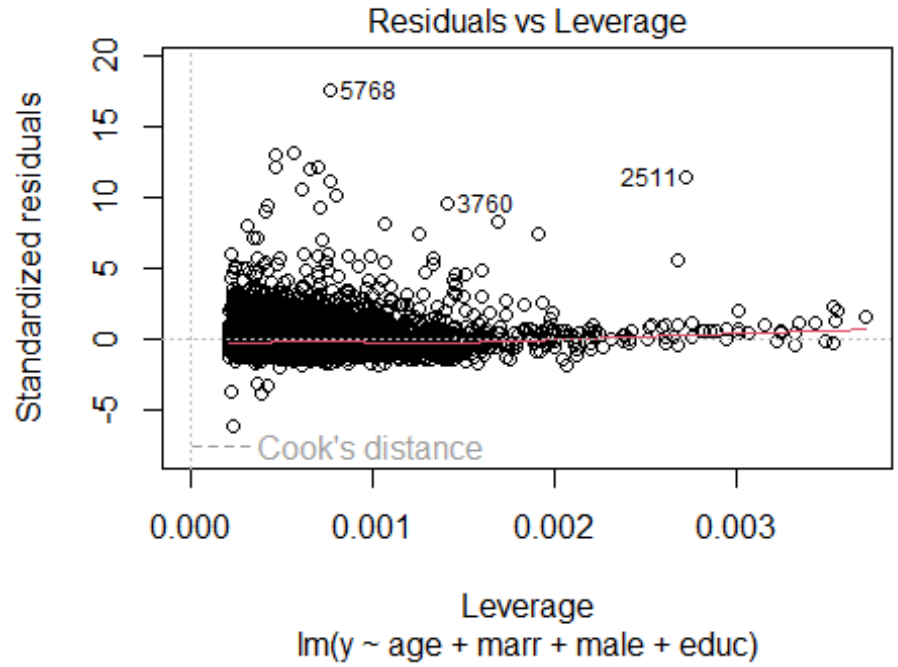
Ordinary Least Square Regression

First regression only includes social features such as age, marital status, education, and gender.


Residuals vs Fitted
Residuals
2000000
500000
-500000
5768
4364
-100000
0
50000
150000
Fitted values
lm(y ~ age + marr + male + educ)


Normal Q-Q
Standardized residuals
20
15
10
5
0
-5
5768
4364
-4
-2
0
2
4
Theoretical Quantiles
lm(y ~ age + marr + male + educ)





For residuals versus fitted and residuals versus leverages the horizontal line is almost at zero. The scale location line is around 1 as required. This shows that there are no large

outliers that would cause bias in our model since the line is horizontal and almost centered around zero. However, the graph shows the presence of some outliers.

In the normal Q-Q plot, the slope is close to 1 with a bit of minor deviation. This shows that the data is normal hence meeting the required assumptions.

```
##
```

```
## Call:
## lm(formula = y ~ age + marr + male + educ, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -596214  -48762  -16455   20804 1721883
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -221873.0     7689.8 -28.853  < 2e-16 ***
## age            3688.0      108.3  34.058  < 2e-16 ***
## marr          42142.4     2417.0  17.436  < 2e-16 ***
## male          14084.7     2955.5   4.765 1.92e-06 ***
## educ           7999.7      400.5  19.976  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 98150 on 7928 degrees of freedom
## Multiple R-squared:  0.1781, Adjusted R-squared:  0.1777
## F-statistic: 429.6 on 4 and 7928 DF,  p-value: < 2.2e-16
```

From this model, all variables are statistically significant with a p-value less than 0.05 at 95% confidence interval. All variables; age, marriage, gender and education are positively associated with the outcome variable total wealth.

This model fit is not good with its predictor variables explaining 17.77% and statistically significant with a p-value less than 2.2e-16.

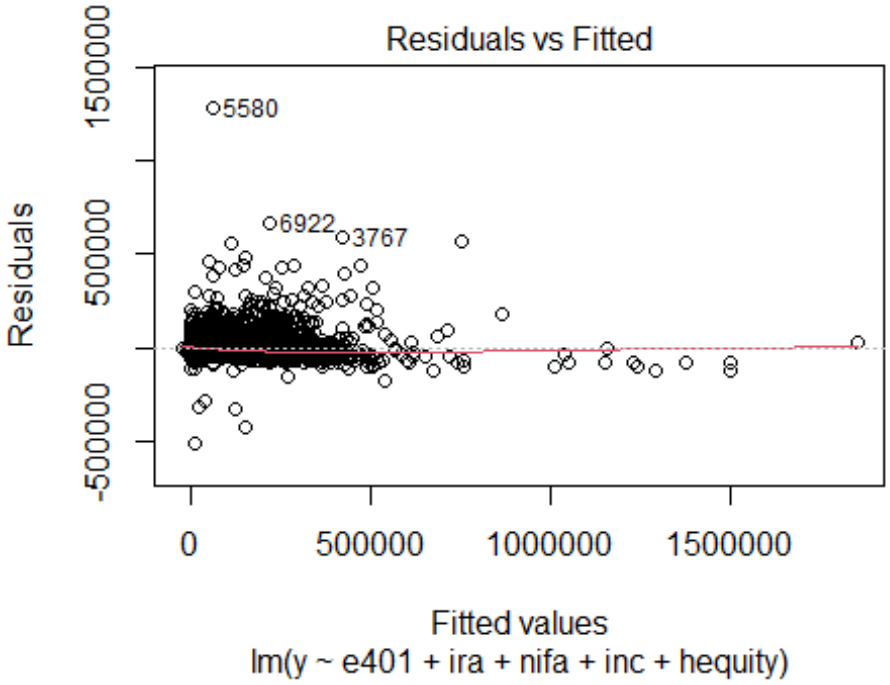
Second regression only includes financial, retirement, and home equity variables

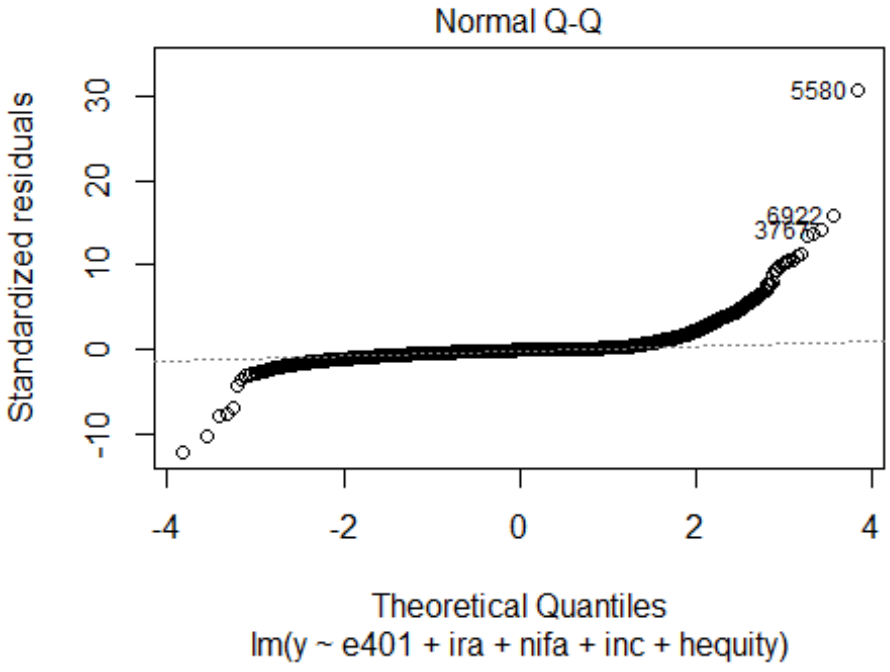
```
##
## Call:
## lm(formula = y ~ e401 + ira + nifa + inc + hequity, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -509184  -12523   -3144    3715 1279531
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -6.585e+03  8.750e+02  -7.526 5.82e-14 ***
```

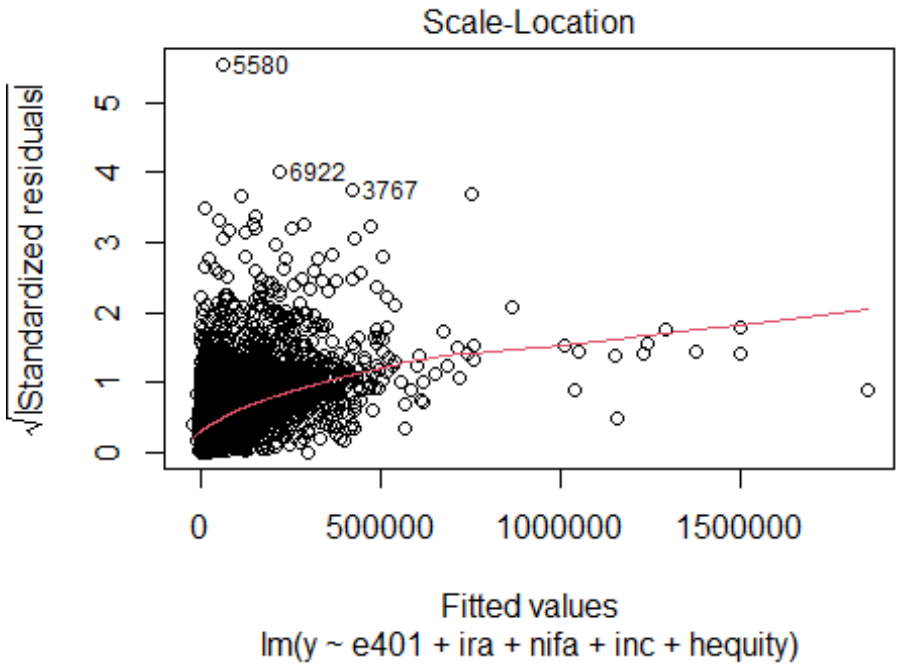
```
## e401         7.676e+03  1.017e+03   7.544 5.05e-14 ***
## ira          1.673e+00  5.442e-02  30.737  < 2e-16 ***
## nifa         1.067e+00  9.889e-03 107.886  < 2e-16 ***
## inc          2.626e-01  2.264e-02  11.600  < 2e-16 ***
## hequity      1.104e+00  1.004e-02 109.933  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 41710 on 7927 degrees of freedom
## Multiple R-squared:  0.8516, Adjusted R-squared:  0.8515
## F-statistic:  9096 on 5 and 7927 DF,  p-value: < 2.2e-16
```

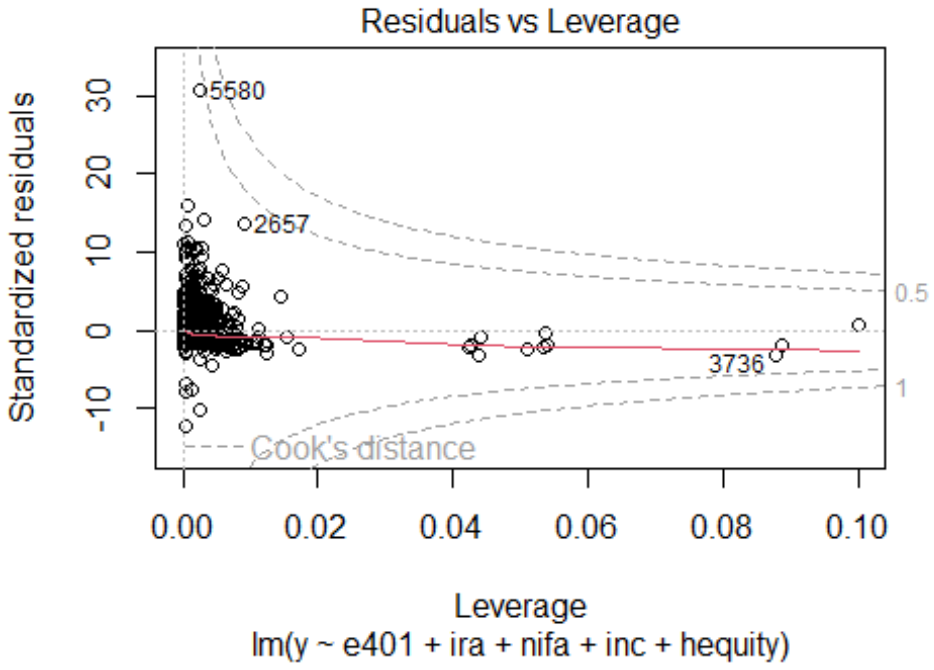
Similar to the first regression, in this model, all variables are statistically significant with a p-value less than 0.05 at 95% confidence interval. All variables; age, marriage, gender and education are positively associated with the outcome variable total wealth.

This model fit is good with its predictor variables explaining 85.15% and statistically significant with a p-value less than 2.2e-16.


Residuals vs Fitted
Residuals
Fitted values
lm(y ~ e401 + ira + nifa + inc + hequity)
5580
6922
3767
1500000
500000
-500000
0
500000
1000000
1500000


Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(y ~ e401 + ira + nifa + inc + hequity)
5580
6922
3767
30
20
10
0
-10
-4
-2
0
2
4





The Third regression is the combination of the variables in first and second regression.

```
##
## Call:
## lm(formula = y ~ e401 + ira + nifa + inc + hequity + age + male +
##     educ + marr, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -509352  -12908   -3306    4442 1275856
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.790e+04  3.493e+03  -5.125 3.05e-07 ***
## e401         7.694e+03  1.016e+03   7.574 4.04e-14 ***
## ira          1.622e+00  5.508e-02  29.445  < 2e-16 ***
## nifa         1.061e+00  9.908e-03 107.060  < 2e-16 ***
## inc          2.980e-01  2.576e-02  11.570  < 2e-16 ***
## hequity      1.088e+00  1.049e-02 103.739  < 2e-16 ***
## age          3.064e+02  5.020e+01   6.102 1.09e-09 ***
## male         3.524e+03  1.257e+03   2.802  0.00509 **
## educ        -8.683e+01  1.888e+02  -0.460  0.64557
## marr        -2.248e+03  1.146e+03  -1.962  0.04980 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 41580 on 7923 degrees of freedom
## Multiple R-squared:  0.8526, Adjusted R-squared:  0.8524
## F-statistic:  5092 on 9 and 7923 DF,  p-value: < 2.2e-16
```

Combining all variables in the first and second regression fit, education and marriage variables becomes statistically insignificant with the other variables being significant.

This model fit is good with its predictor variables explaining 85.24% and statistically significant with a p-value less than 2.2e-16.

Compute the in sample MSPE of each model

```
## [1] 9626754407 1738563515 1726694810
```

Computing the in-sample MSPE for each model, third regression becomes the most optimized or accurate with the lowest in sample mspe.

These results agree with the goodness of the fit. It's the fit where its explanatory variables explain the highest percentage of total wealth.

5-fold cross validation to compare the three ols models

```
## [1] 12354218453  1788255904  1783801037
```

Using this comparison method to compare the mspe's of the three models, third regression is the most optimized or accurate model with MSPE of 1783801037.

Third regression is the most effective model under both approaches, having the highest R-squared value and the lowest MSPE, allowing us to conclude that combining financial and social individual data will yield the most accurate total wealth prediction.

Therefore the third model becomes reliable to predict the total wealth even if the mspe of the second regression is very close to the mspe of the third one.

Stepwise,Lasso, Ridge

5-fold cross validation to compare stepwise, lasso, ridge

```
## [1] 1733071536 1732050431 2014425515 2002979166
```

On the basis of 5-fold cross validation on the entire dataset, it is found that stepwise backward is the best accurate model because it has the lowest MSPE.

The MSPE of the stepwise backward model and the third regression from the preceding section, however, are not significantly different from one another. Both of them employ nine different variables, which combine financial and social factors. However, it appears that home equity, education, and marital status are less statistically significant than home mortgage, home value, and whether the household has two earners.

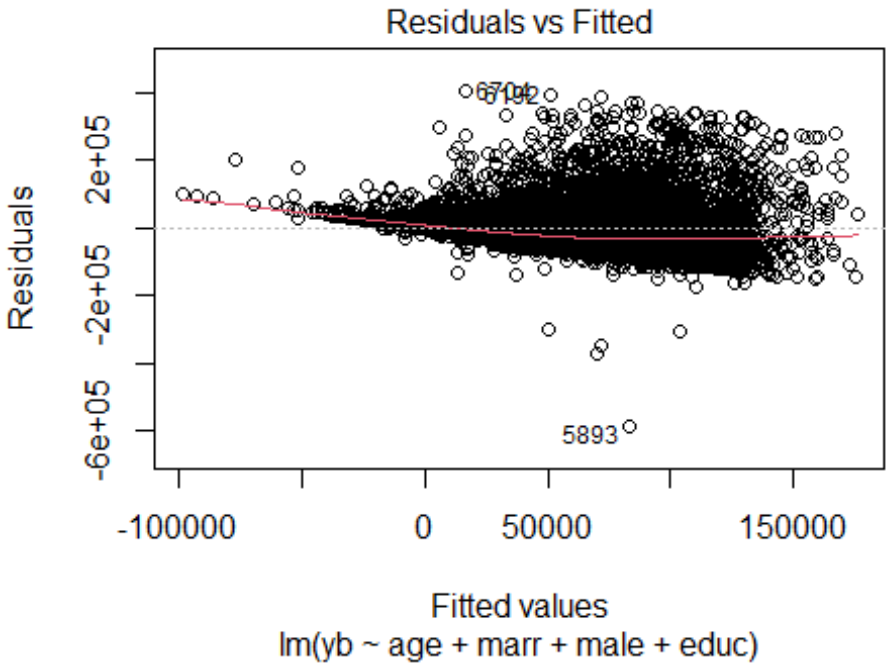
```
##
## Call:
## lm(formula = tw ~ ira + e401 + nifa + inc + hmort + hval + male +
##     twoearn + age, data = data)
##
## Coefficients:
## (Intercept)          ira         e401         nifa          inc
hmort
##  -1.889e+04    1.617e+00    7.791e+03    1.055e+00    3.122e-01
-1.046e+00
##        hval         male      twoearn          age
##   1.082e+00    2.972e+03   -6.588e+03    2.999e+02
```

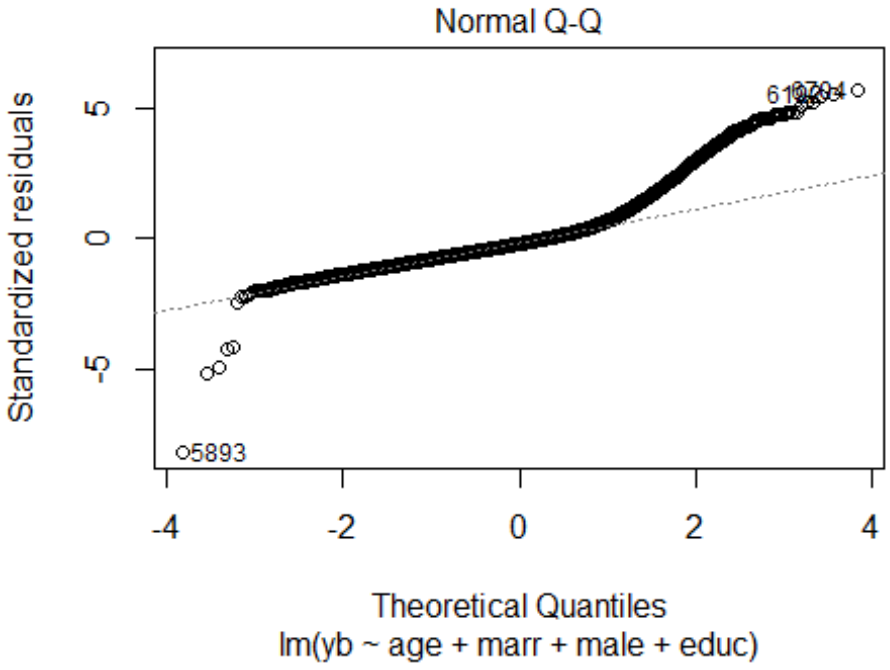
```
## [1] 1718301666
```

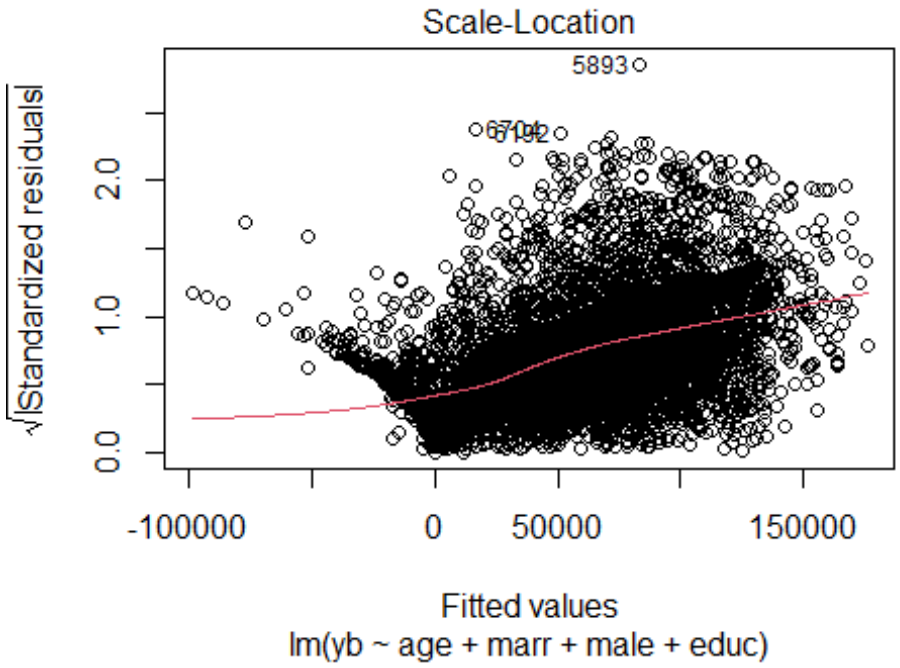
The model, which had the lowest MSPE of 1726409016, may benefit from backward stepwise regression, according to our suggestion. With a smaller MSPE and a higher r-squared value of 0.8531, stepwise regression is more understandable than third regression, according to a comparison of the coefficients of the two regression models. Therefore, we decide to predict total wealth using backward stepwise regression.

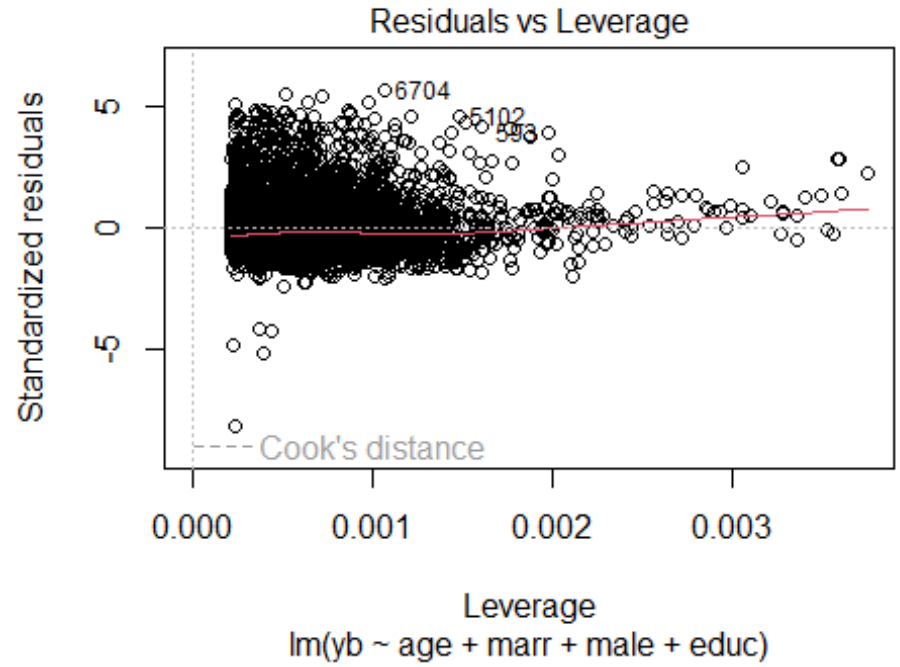
Removing outliers from the data set

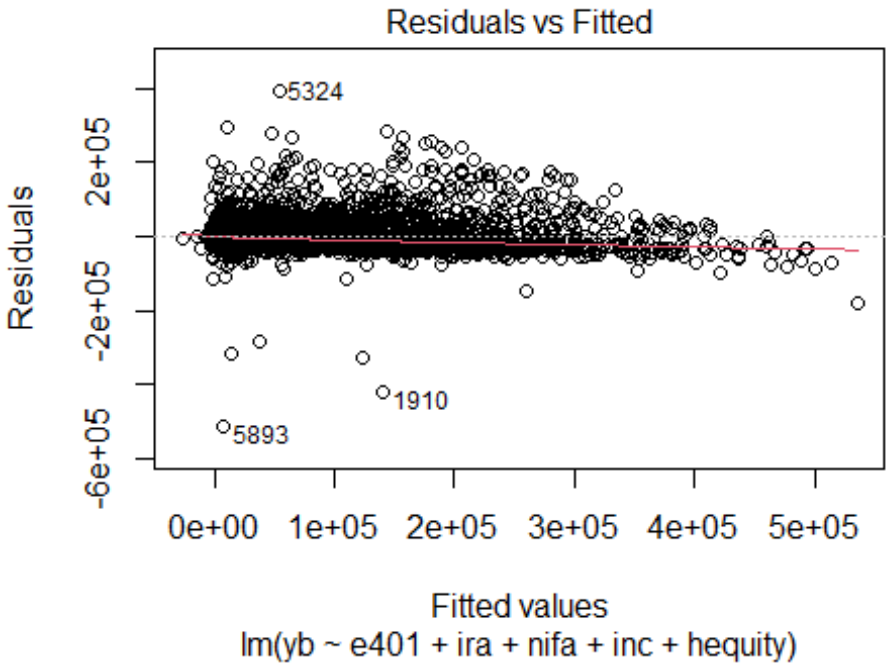
```
##      99%
## 461425.1
```

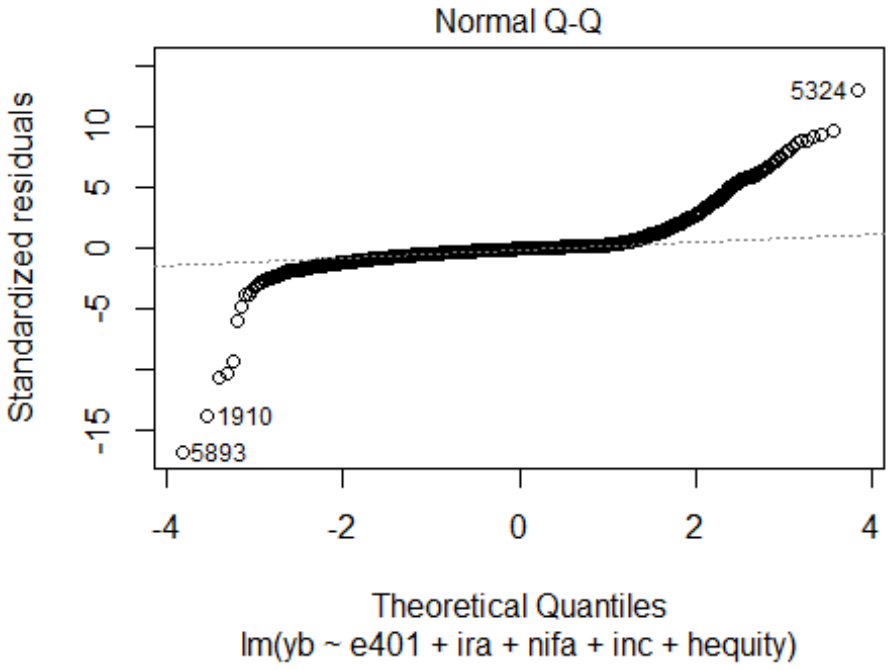

Residuals vs Fitted
Residuals
2e+05
-2e+05
-6e+05
5893
-100000
0
50000
150000
Fitted values
lm(yb ~ age + marr + male + educ)

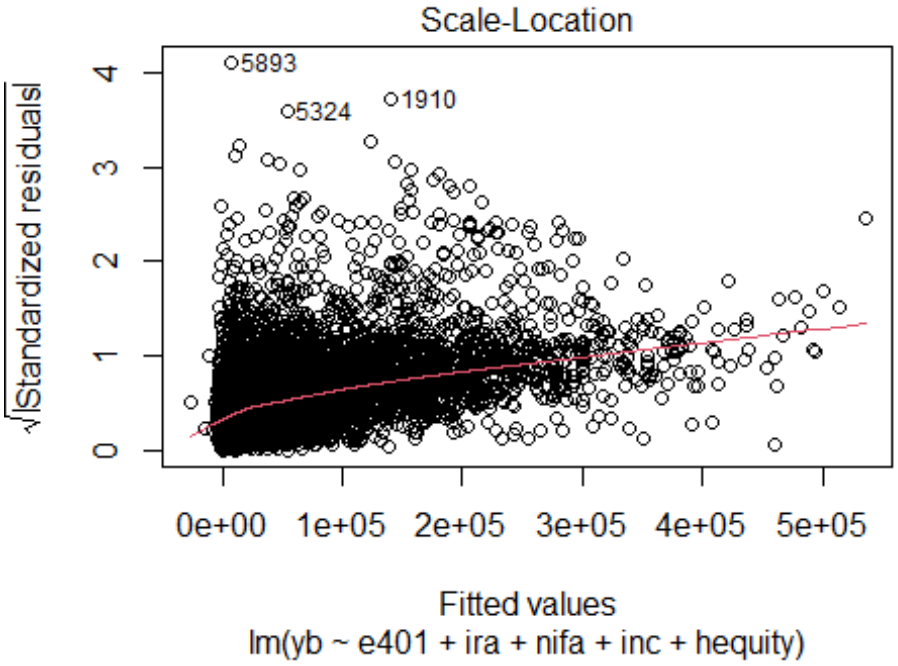

Normal Q-Q
Standardized residuals
5
0
-5
5893
-4
-2
0
2
4
Theoretical Quantiles
lm(yb ~ age + marr + male + educ)

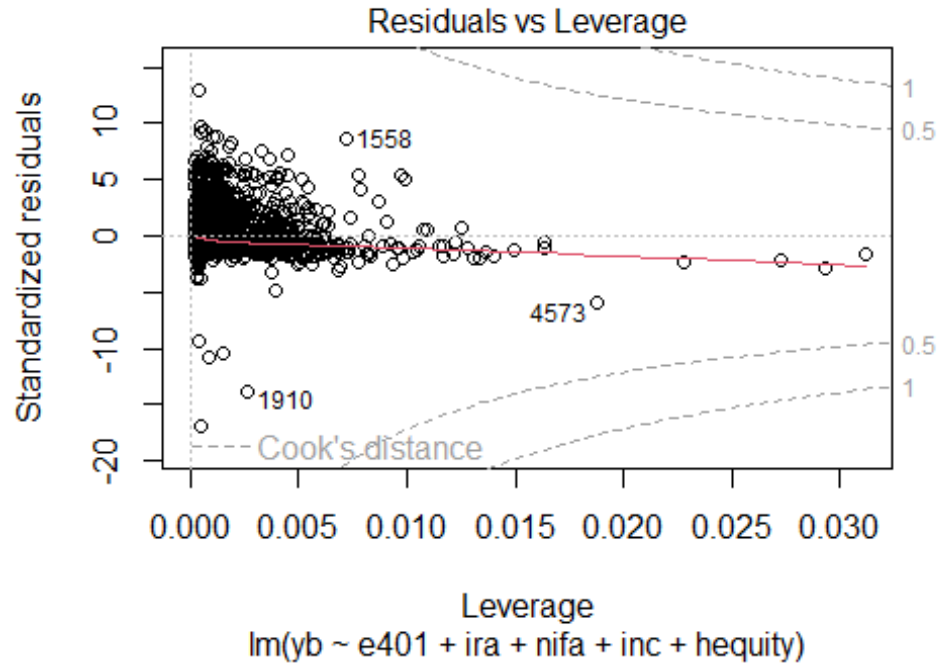

Scale-Location
√|Standardized residuals|
5893
6704
6102
0.0
1.0
2.0
-100000
0
50000
150000
Fitted values
lm(yb ~ age + marr + male + educ)

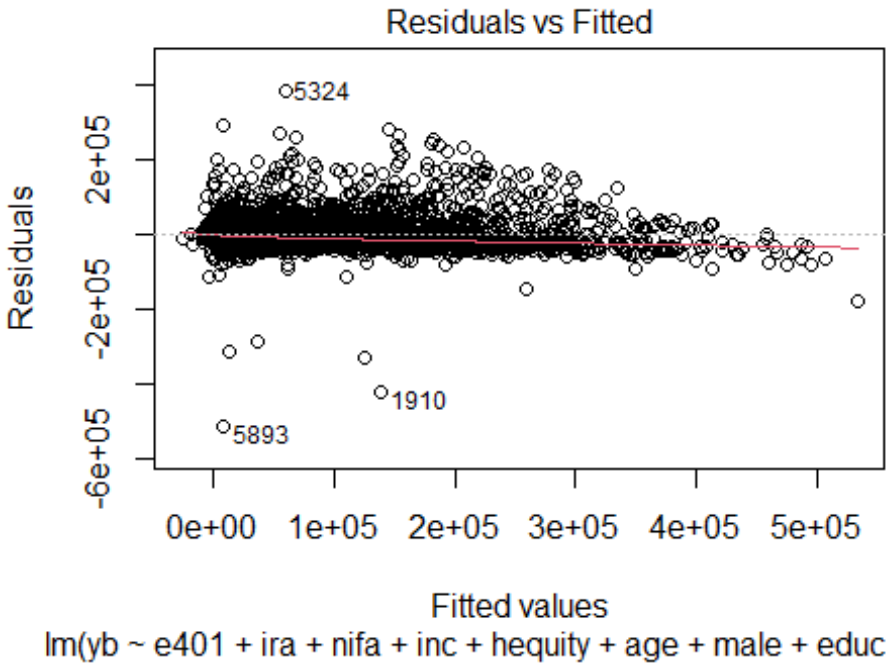

Residuals vs Leverage
Standardized residuals
6704
5102
593
Cook's distance
5
0
-5
0.000
0.001
0.002
0.003
Leverage
lm(yb ~ age + marr + male + educ)

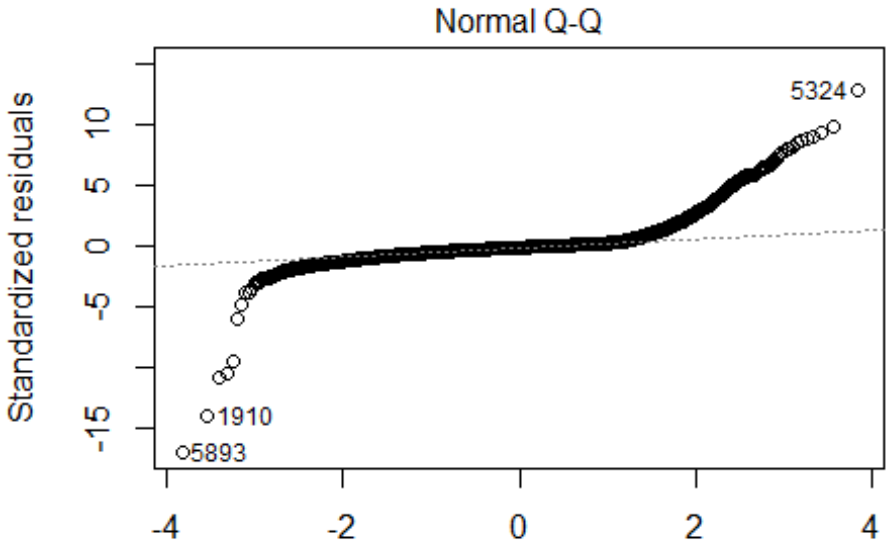

Residuals vs Fitted
Residuals
Fitted values
lm(yb ~ e401 + ira + nifa + inc + hequity)
5324
1910
5893
-6e+05
-2e+05
2e+05
0e+00
1e+05
2e+05
3e+05
4e+05
5e+05

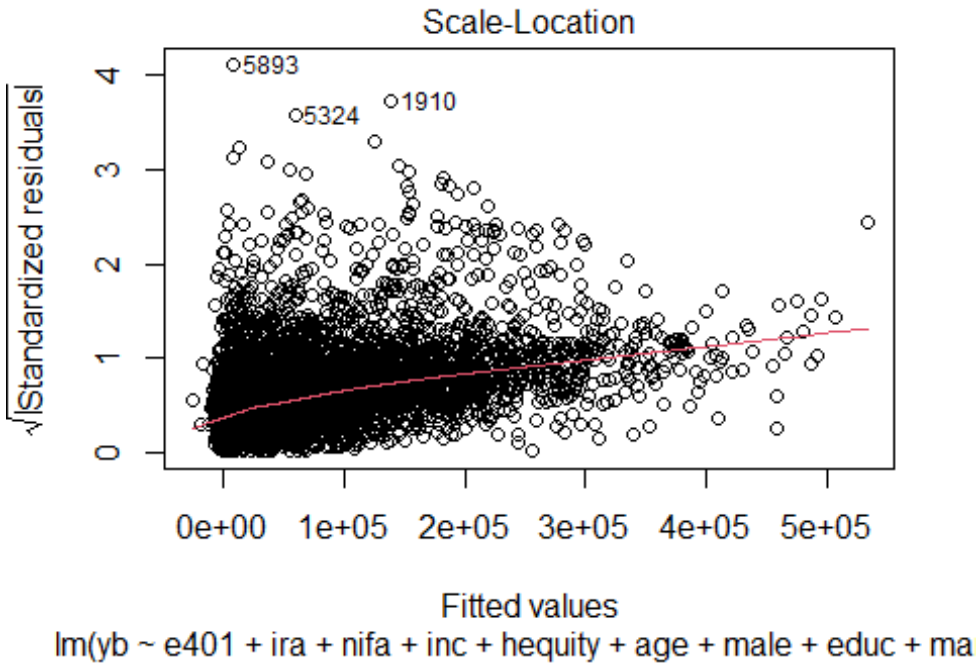

Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(yb ~ e401 + ira + nifa + inc + hequity)
5324
1910
5893
-15
-5
0
5
10
-4
-2
0
2
4

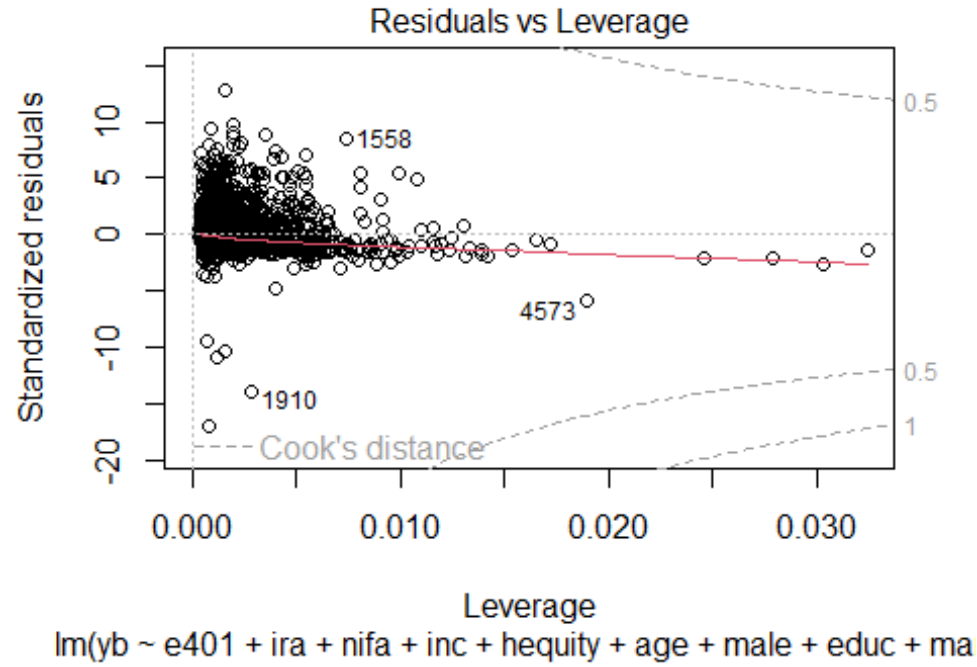

Scale-Location
√|Standardized residuals|
5893
5324
1910
0
1
2
3
4
0e+00
1e+05
2e+05
3e+05
4e+05
5e+05
Fitted values
lm(yb ~ e401 + ira + nifa + inc + hequity)


Residuals vs Leverage
Standardized residuals
1558
4573
1910
Cook's distance
1
0.5
0.5
1
-20
-10
0
5
10
0.000
0.005
0.010
0.015
0.020
0.025
0.030
Leverage
lm(yb ~ e401 + ira + nifa + inc + hequity)


Residuals vs Fitted
Residuals
Fitted values
lm(yb ~ e401 + ira + nifa + inc + hequity + age + male + educ + ma
5324
1910
5893
-6e+05
-2e+05
2e+05
0e+00
1e+05
2e+05
3e+05
4e+05
5e+05


Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(yb ~ e401 + ira + nifa + inc + hequity + age + male + educ + ma
5324
1910
5893
-15
-5
0
5
10
-4
-2
0
2
4






```
##
## Call:
## lm(formula = yb ~ age + marr + male + educ, data = data_trimmed)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -585430  -42009  -13933   20463  405007
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -182417.20    5686.66 -32.078  < 2e-16 ***
## age            3073.13      79.95  38.436  < 2e-16 ***
## marr          35230.84    1776.30  19.834  < 2e-16 ***
## male           9457.97    2175.23   4.348 1.39e-05 ***
## educ           6842.31     295.86  23.127  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 71840 on 7848 degrees of freedom
## Multiple R-squared:   0.22,  Adjusted R-squared:  0.2196
## F-statistic: 553.5 on 4 and 7848 DF,  p-value: < 2.2e-16
```

Removing the outliers, the adjusted R squared for this model (first regression model) increases from 0.1777 to 0.2196 .

```
##
## Call:
## lm(formula = yb ~ e401 + ira + nifa + inc + hequity, data = data_trimmed)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -510462  -10630   -2895    2573  391150
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3.997e+03  6.427e+02   -6.22 5.23e-10 ***
## e401         7.833e+03  7.428e+02   10.54  < 2e-16 ***
## ira          1.497e+00  4.142e-02   36.14  < 2e-16 ***
## nifa         1.100e+00  1.529e-02   71.94  < 2e-16 ***
## inc          1.932e-01  1.697e-02   11.38  < 2e-16 ***
## hequity      1.059e+00  7.551e-03  140.21  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 30270 on 7847 degrees of freedom
## Multiple R-squared:  0.8615, Adjusted R-squared:  0.8614
## F-statistic:  9762 on 5 and 7847 DF,  p-value: < 2.2e-16
```

Removing the outliers, the adjusted R squared for this model (second regression model) increases from 0.8515 to 0.8614 .

```
##
## Call:
## lm(formula = yb ~ e401 + ira + nifa + inc + hequity + age + male +
##     educ + marr, data = data_trimmed)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -511322  -10964   -3202    3489  385359
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.714e+04  2.562e+03  -6.687 2.43e-11 ***
## e401         7.806e+03  7.408e+02  10.538  < 2e-16 ***
## ira          1.452e+00  4.179e-02  34.737  < 2e-16 ***
## nifa         1.085e+00  1.535e-02  70.676  < 2e-16 ***
## inc          2.123e-01  1.929e-02  11.008  < 2e-16 ***
## hequity      1.044e+00  7.872e-03 132.598  < 2e-16 ***
## age          2.827e+02  3.674e+01   7.694 1.60e-14 ***
## male         2.759e+03  9.168e+02   3.009  0.00263 **
## educ         1.356e+02  1.385e+02   0.979  0.32761
## marr        -1.143e+03  8.353e+02  -1.368  0.17137
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 30140 on 7843 degrees of freedom
## Multiple R-squared:  0.8628, Adjusted R-squared:  0.8626
## F-statistic:  5479 on 9 and 7843 DF,  p-value: < 2.2e-16
```

Removing the outliers, the adjusted R squared for this model (third regression model) increases from 0.8524 to 0.8626 .

Compute the in sample MSPE of each model

```
## [1] 5157050986  915768454  907374865
```

Removing the outliers, third regression remains the best model however with a reduced mspe. For all OLS regressions, mspe is lower with the trimmed data set.

5-fold cross validation to compare the three ols models

```
## [1] 4708113332  806392236  800268595
```

Comparing the mspe's of the three models, third one is the most optimized or accurate model.

Cross validation to compare stepwise, lasso, ridge

```
## [1] 903997404 903997404 956902950 986555952
```

```
##
## Call:
```

```
## lm(formula = tw ~ ira + e401 + nifa + inc + hmort + hval + male +
##     twoearn + age, data = data_trimmed)
##
## Coefficients:
## (Intercept)          ira         e401         nifa          inc
hmort
##  -1.554e+04    1.456e+00    7.781e+03    1.079e+00    2.099e-01
-9.811e-01
##        hval         male      twoearn          age
##   1.036e+00    2.520e+03   -4.266e+03    2.808e+02
```

```
## [1] 901116648
```

Conclusion

Trimmed data to remove outliers is better for prediction. If we only used MSPE as the criterion to determine which regression is better, we would choose stepwise regression because it has a slightly lower MSPE than multiple linear regression after comparing the MSPE on the two regressions. Stepwise regression is a superior model for the data because there is such a minor variation in the r-square values.

In this study, we looked at how the 'tw' or total wealth variable related to the variables provided in the data tr.txt dataset. By applying multiple linear regressions to the data, we can observe that financial variables like income, home equity, 401(k) and IRA assets, non-401(k) financial assets, and married status are statistically more important in predicting total wealth than age, gender, marital status, and education. Cross validation was then utilized to establish that combining both financial and social individual data will provide the best accurate prediction of total wealth. The most statistically significant in predicting overall wealth are two-earner households.

Then, we used stepwise regression to create a better-fitting model, which revealed that the presence of two earners in a home is inversely connected with overall wealth. This finding defies logic because single-earner households typically have enough income to support the entire family. We were able to create far more accurate and understandable models after excluding the top 1% of total wealth from the data set, which revealed to us that many of the people in the top 1% of total wealth do not have a high level of education. The most statistically significant factors in the forecast of total wealth, according to our findings, are individual retirement accounts, 401(k) and non-401(k) financial assets, income, home mortgage and home value, gender, age, and two-earner households.