Prediction of Generosity of Income groups in a Zip Code

Matthew Elliott Monica Hegde Jules Testard
**University of California,
San Diego**

ABSTRACT

In this paper, we describe the methodologies used by us to predict the generosity of different income groups living within a zip code. These results are based on data collected from IRS and US Census for every zip code in America. Generosity is the percentage of one’s income a person donated to charity. This is declared on the return forms when taxes are filed. The income groups of people is determined based on the income bracket that they belong to when filing taxes. We describe the models we used for the prediction task and discuss the implications of the accuracy of the models and the interpretation of the results.

CCS Concepts

• **Computing methodologies → Machine learning → Machine learning**

• **Applied computing →Law, social and behavioral sciences →Economics**

Keywords

Data Mining; Linear Regression; Latent Factor Models; Census data; Tax returns data; Philanthropy

1. INTRODUCTION

The National Center for Charitable Statistics (NCCS) has concluded that individual contribution for 2012 in USA amounted to $228.93 billion,accounting for 72% of all charitable donations. [1] Since 2007, the center has also been working on the **Fundraising Effectiveness Project** (FEP),which aims to help grow philanthropy’s share of the GDP. Our efforts in this project are to consider not only the IRS data as used by NCCS but also to consider various other factors that affect the percentage of charitable donations. Such information could be useful to NGOs and other organizations that rely on fundraising in order to operate.

It is important to determine the percent of income that is donated to charity so that charitable organizations can utilize that to narrow down the areas where efforts can be concentrated to achieve maximum turnovers, especially in emergent situations such as natural disaster relief. Since charitable donation is claimed on tax return forms, we can build a model which can determine the areas of the country and the classes of people who are most likely to make large donations, based on a small set of tax returns forms. Such data is useful to nearly all charities that rely on advertising themselves for revenue. We will discuss two models that we used to make the prediction, namely, Linear Regression and Latent Factor Models.

2. LITERATURE

The NCCS has done considerable statistical analyses in this area with its results published on the website http://nccs.urban.org. The report published with charitable contributions of each state was of particular interest to us. The main goal of NCCS is to maintain data on the non-profit, government, commercial and civil sectors, and to help grow philanthropy. As such, it has only used tax returns and non-profits data in its analyses.

The newspaper, The Chronicle of Philanthropy, dedicated a portion of its website called "How America Gives" to post findings that use tax data to analyze generosity. The article showed the level of contributions for every zip code in America. It also highlighted how contributions within zipcodes change based on wealth and other factors like religious affiliation [2].

Matthew Elliot originally collected the data used in this paper while he was studying with Professor Emmanuel Saez at UC Berkeley. The data set is a combination of data from IRS tax data and US Census data. Its original purpose was to analyze intergenerational wealth mobility using tax data from every zip code in America [3].However, the data set lends itself well to a vast amount of research topics, because it contains 100’s of features for nearly every zip code in America. In our work, we reuse this dataset to predict generosity in America, using features like Contributions divided by income, Gini index, ethnic diversity index, and income by wealth, and standard deductions.

3. DATA SET

The data set consists of 119,000 data points. The data points come from each of the 7 income groups that exist within the 17,000 zip codes within America. This data was generated by cross-referencing tax return information from the IRS with US Census data for each zip code.

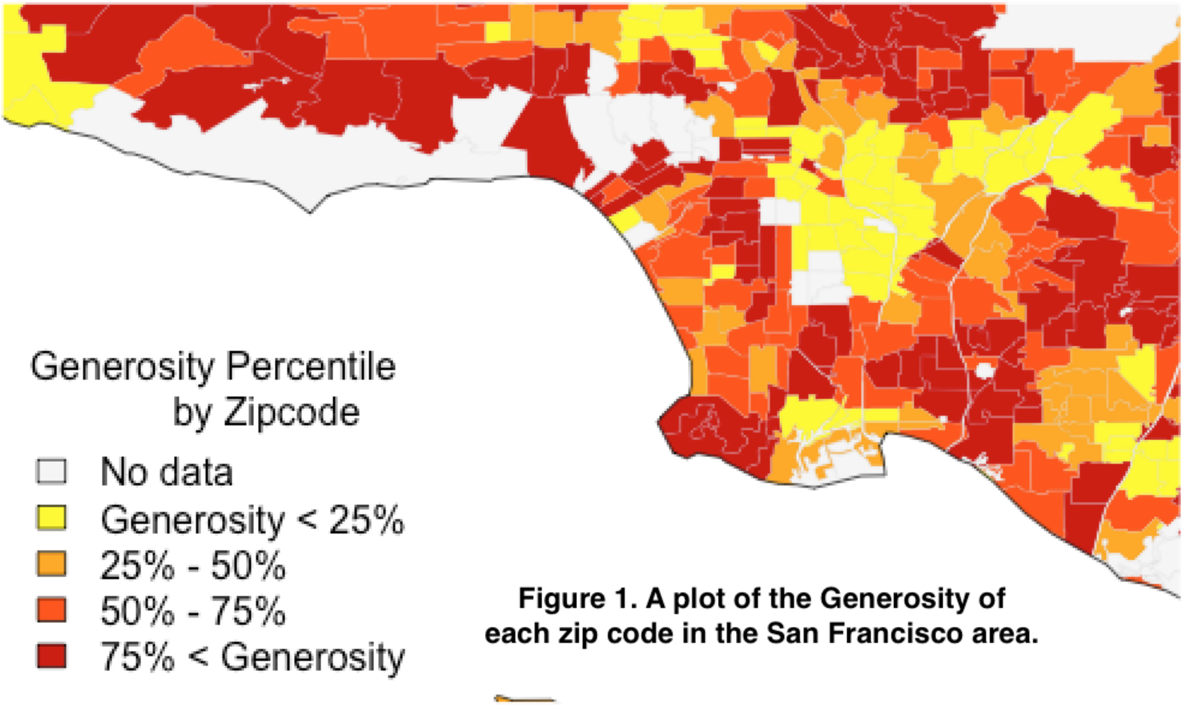


Figure 1. A plot of the Generosity of each zip code in the San Francisco area.

The data used pertains to the year 2007. Data was compiled for other years as well, but for this assignment we considered only 1 cross section in time. Each data point has 6 attributes specific to the income class and 49 attributes common across all income classes for a given zipcode. Features in the data set such as Gini Index, Fragmentation Index, and Generosity were computed from features that were available in the original data. The goal is to predict the generosity value for every zip code and income group combination. It is important to consider each income group in determining the generosity so that the best groups can be targeted. The data set was divided into 50% training data and 50% test data to run the prediction models and calculate their accuracy.

Figure 1 shows a plot of the generosity of the Los Angeles metropolitan area by zip code across all wealth groups. Figure 2 (see appendix) shows a plot of the wealth distribution by zip code in the same area. We were surprised by the level of diversity in generosity within a single region. Generosity varies considerably within a single city. Wealthier communities such as Hollywood, Palos Verdes, Santa Monica and the Malibu Hills tend to be very generous, while poorer communities around south and central LA less so.

4. PREDICTIVE TASK

Our goal is to predict the generosity (Cont.AGI) for a given income group in a given zip code. The baseline model is a model that predicts the average generosity value for the entire country for each data point.

We decided to use the 50% test data to assess the accuracy of the results. The measure of the accuracy of the three models (the two models that we will subsequently discuss and the baseline) will be in terms of Mean Square Error (MSE). The attributes described were carefully extracted from IRS and Census data programmatically, while ignoring irrelevant details such as dependents, different investments, number of pets, number of employees for a household, etc. Hence, all the attributes listed are relevant to the prediction task at hand.

5. MODELS

We considered two models in this prediction task and compared the results. The first model uses Linear Regression, while the second model uses latent-factors. Both techniques were used as seen in class.

5.1 Linear Regression

We decided to run Linear Regression on the model since the task was to predict the percent of charitable donations of the income and there were real-valued, categorical as well as binary features.

Some of the features we considered to be most important were average income, number of standard deductions, and average savings. These features are a strong indicator of wealth and exist for every income group of each zip code. Other features shared by all incomes groups for a given zip code such as average age, property value, and family size was also used.

Though our linear model has over 45 features, most of these features are specific only to the zip code, and not to the income groups within the zip code. These features have limited ability to predict minutia for each income group within a zipcode, and thus we are not concerned about over fitting. For this reason, no lambda term was used in our model.

We ran the least squares linear regression on the training set using all the attributes listed in Table 1 except zip code and State. We used the formula below to generate the best values for θ.

$$\arg\min(\theta) = \frac{1}{N}\|y - X\theta\|_2^2$$

Where, $N \rightarrow$ *number of data points*

$y \rightarrow$ *matrix for Cont.AGI values*

$X \rightarrow$ *matrix of features*

$\theta \rightarrow s$

The strength of this model is that it helps us interpret the results and discuss the effect of the parameters on the predicted value. However, we wanted to use another model to see if it performed better at prediction than linear regression.

5.2 Latent Factor Model

We noticed that neighborhoods with high average income donated more, thus we extrapolated that within a zip code, and generosity will fluctuate considerably with income. Therefore, a latent-factor model predicting generosity for (income, zip code) pairs would likely be very accurate.

We used the latent factor model by calculating bias terms for each income group and each zip code. We chose a value of 0 as initialization for the β_i and β_z. We chose the initial of α to be the mean generosity across the nation.

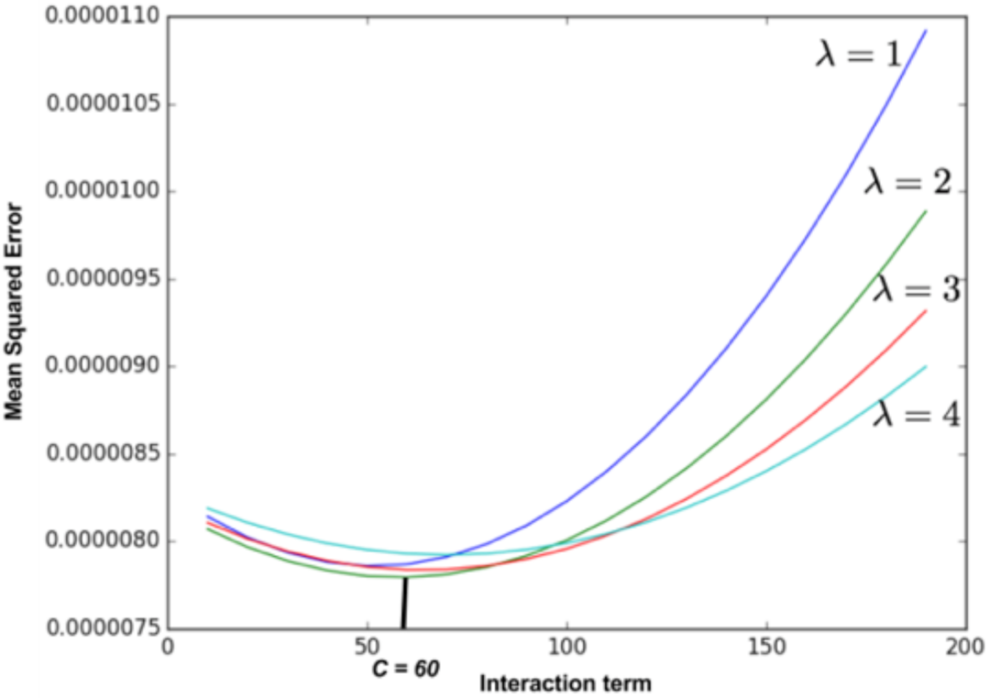
$$f(i,z) = \alpha + \beta_i + \beta_z$$

Where, $z \rightarrow$ *zip code*

$i \rightarrow$ *income group*

We tried a variety of values for regularization of our

Figure 2. MSE vs. the constant in the interaction term for different values of λ



latent-factor model, from $\lambda = 1$ to $\lambda = 4$. We found that $\lambda = 2$ gave us the best results. We did not attempt to compute a more accurate λ value by increasing granularity.

We then added an interaction term to improve the performance of the model, in turn finding the interaction between income groups and zip codes.We denote the interaction term using the Greek letter γ. We attempted interaction terms values from γ=10 to γ = 200. We report our results on figure 2 (see bottom of page). We found the minimal value of the MSE to occur at λ = 2and γ = 60. Notice how the MSE is higher when $\lambda = 1$ as well as when $\lambda = 3$, meaning we expect the λ value found to be very near optimal.

6.RESULTS AND CONCLUSION

As anticipated, the latent factor model performed much better than the linear regression model. The results for each model were as follows.

6.1 Baseline model

For the baseline, we calculated the mean Generosity (Cont.AGI) value and the MSE on the test set using the following mean.

$$\text{Mean} = 0.015503652746$$

$$\text{MSE} = 0.000143$$

6.2 Linear Regression

$$\text{MSE} = 5.412299\text{e-}05$$

Linear regression performs much better than the baseline model using the best attributes as detailed. The θ values have been added in the Appendix. The features that were most significant were the ones that came from the IRS tax data set for every income group within a zipcode. The features that related to wealth were the most significant, while features relating to cultural aspects such as language spoken at home were less significant.

When considering features we did not consider the polynomial values of any features. After initially checking polynomial values for the most significant features, no change was made in the regression. Thus we discarded the polynomial terms. We did, however, take the log of a few of the features such as average gross income of a wealth group, and Gini index. This had a significant affect on the regression.

We did notice that we could obtain more precise results predicting the log of generosity instead of generosity itself. This is because generosity is heavily left skewed instead of being normally distributed. While such a model yields more accurate predictions, interpreting the results of such a model becomes more difficult because the relationship between the feature and generosity is not linear.

6.3 Latent Factor Model

The latent factor model produced the results below, after applying gradient descent to find the best possible values for the terms.

$$\text{MSE} = 6.80456785995\text{e-}05$$

$$\lambda = 2$$

After adding the interaction term as a constant multiplier to the product of the β's we obtained the following mean squared error:

$$\text{MSE} = 7.79470499255\text{e-}06$$

From the results, it was clear that latent-factor model performs much better than the linear regression.We

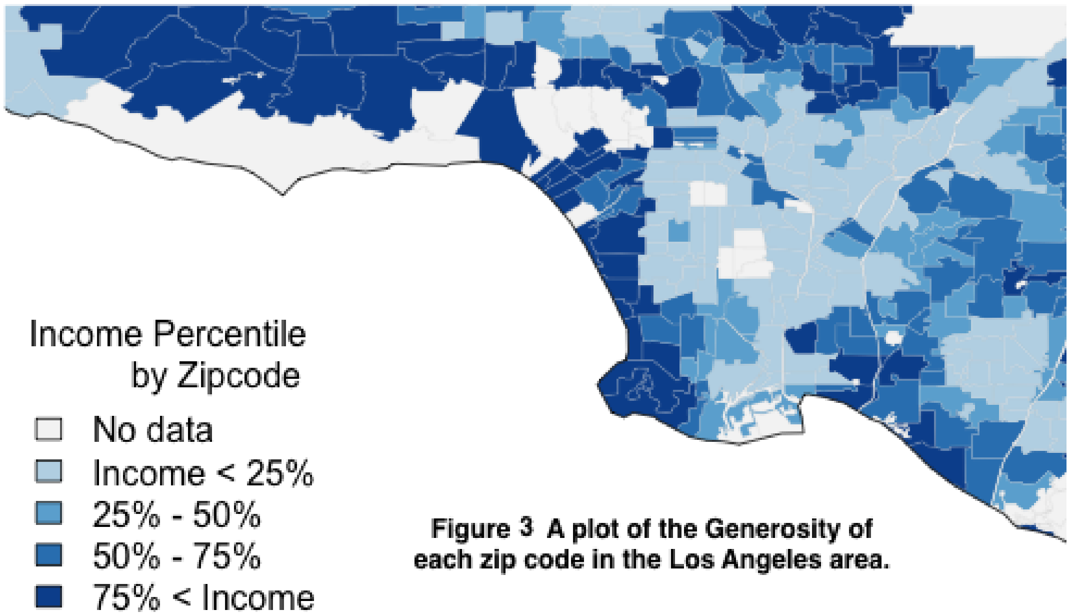


Figure 3 A plot of the Generosity of each zip code in the Los Angeles area.

believe the linear regression falls short because we are incapable of accounting for all the volatility of generosity in terms of features that can be represented in the model. On the other hand, the latent-factor is able to account for all the volatility the linear regression is not able to predict by considering every zip code as its own features. This takes into account minute difference in regional generosity.

Linear regression, on the other hand, provides more insight into which features have a greater effect on generosity. For instance, we find that areas with high levels of education, families and older populations tend to be considerably more generous; while young urban areas tend to be considerably less. Our findings also showed that areas with more ethnic diversity and socioeconomic diversity also tend to be less generous. However, when considering an individual regression for each wealth group, ethnic and socioeconomic diversity had a positive effect.

REFERENCES

[1] http://nccs.urban.org/nccs/statistics/Charitable-Giving-in-America-Some-Facts-and-Figures.cfm

[2] ”America’s Generosity Divide.” The Chronicle of Philanthropy. 19 Aug 2012.http://philanthropy.com/article/America-s-Generosity- Divide/133775/

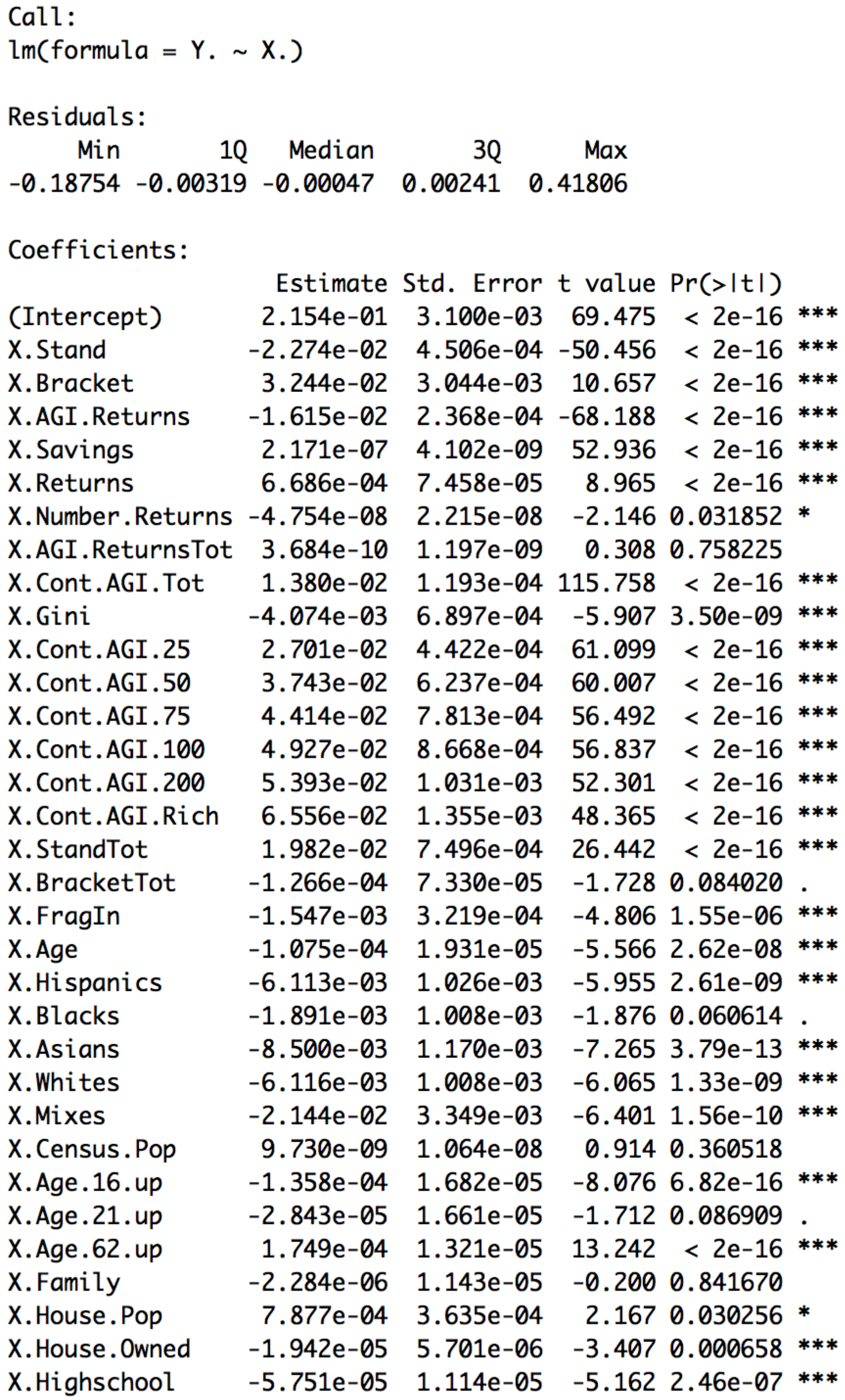
[3] http://eml.berkeley.edu/~saez/chetty-friedman-kline-saezQJE14mobility.pdf

APPENDIX

Table1. Description of features used in linear regression

Attribute	Meaning
Zip.Code	Zip code
State	State
Number.Returns	Number of tax returns filed
AGI.ReturnsTot	Average Gross Income (AGI) by total returns
Cont.AGI.Tot	Total donations by AGI
Gini	Gini Index (denotes income diversity)
Cont.AGI	**Total contributions by AGI for income group**
Cont.AGI.25	Group with AGI < $25,000
Cont.AGI.50	Group with $25,000 <= AGI < $50,000
Cont.AGI.75	Group with $50,000 <= AGI < $75,000
Cont.AGI.100	Group with $75,000 <= AGI < $100,000
Cont.AGI.200	Group with $100,000 <= AGI < $200,000
Cont.AGI.Rich	Group with $200,000 <= AGI
Stand	Standard deductions by number of returns for income group
StandTot	Standard deductions by number of returns
Bracket	Approximated tax bracket for income group
BracketTot	Approximated tax bracket
AGI.Returns	AGI by returns for the income group
FragIn	Fragmentation Index (denotes ethnic diversity)
Age	Average age of the group
Hispanics	Percent of non-white Hispanics
Blacks	Percent of African-Americans
Asians	Percent of Asians
Whites	Percent of Whites
Mixes	Percent of mixed race population
Census.Pop	Recorded Census population
Age.16.Up	Percent above 16 years of age
Age.21.Up	Percent above 21 years of age
Age.62.Up	Percent above 62 years of age
Family	Percent of family households
House.Pop	Average household population
House.Owned	Percent of households that own their house
Highschool	Percent with highschool education
Bachelors	Percent with Bachelors education
Graduate	Percent with Graduate education
No.English	Percent who do not speak English
House.Rooms	Average number of rooms in a house
No.Cars	Percent who do not own cars
No.Heat	Percent of homes with no heat
Rent	Average rent paid
House.Price	Average price of a house
No.Mortgage	Percent households with no mortgage
Grapi.15	Percent whose rent as a percentage of income is less than 15%
Grapi.35	Percent whose rent as a percentage of income is less than 35%
Smocapi.M.20	Percent mortgage owners whose home costs as a % of income is less than 20%
Smocapi.M.30	Percent mortgage owners whose home costs as a % of income is less than 35%
Smocapi.NM.10	Percent non-mortgage owners whose home costs as a % of income is less than 10%
Smocapi.NM.35	Percent non-mortgage owners whose home costs as a % of income is less than 35%
House.2000	Percent of houses built after 2000
House.1990	Percent of houses built after 1990
House.1980	Percent of houses built after 1980
House.1970	Percent of houses built after 1970
House.1960	Percent of houses built after 1960
House.1950	Percent of houses built after 1950
House.1940	Percent of houses built after 1940
House.Old	Percent of houses built before 1940
Savings	Average amount in Savings Account for the income group (in $)

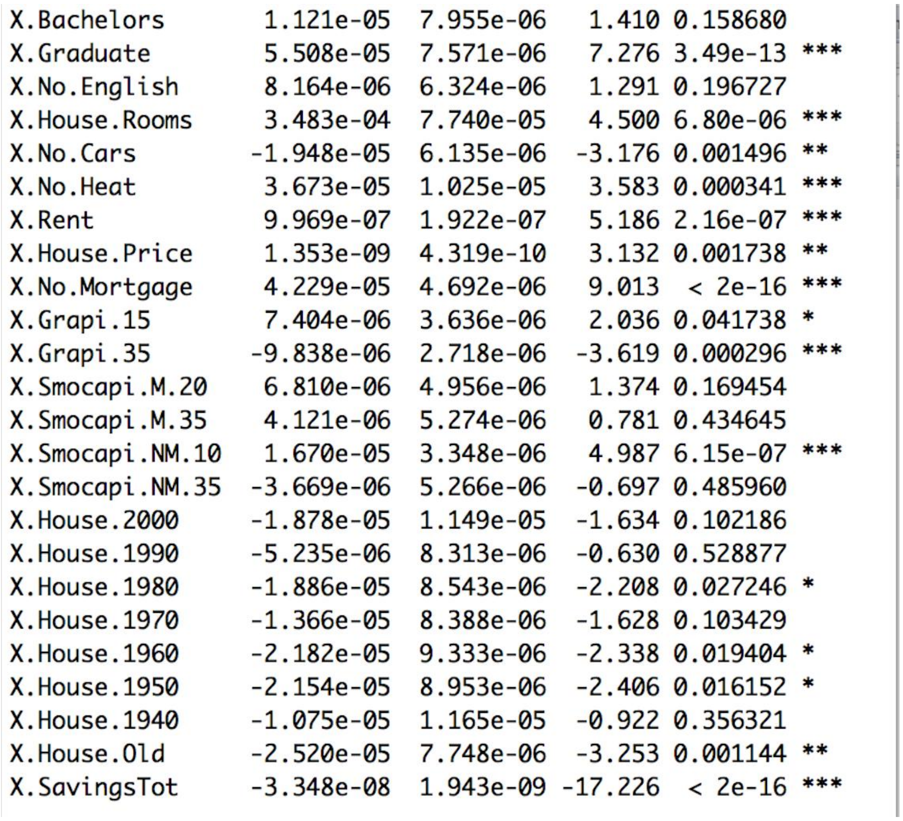
Figure 4 : Coefficients of the linear regression



```
Call:
lm(formula = Y. ~ X.)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.18754 -0.00319 -0.00047  0.00241  0.41806 

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)    
(Intercept)       2.154e-01  3.100e-03  69.475  < 2e-16 ***
X.Stand          -2.274e-02  4.506e-04 -50.456  < 2e-16 ***
X.Bracket         3.244e-02  3.044e-03  10.657  < 2e-16 ***
X.AGI.Returns    -1.615e-02  2.368e-04 -68.188  < 2e-16 ***
X.Savings         2.171e-07  4.102e-09  52.936  < 2e-16 ***
X.Returns         6.686e-04  7.458e-05   8.965  < 2e-16 ***
X.Number.Returns -4.754e-08  2.215e-08  -2.146 0.031852 *  
X.AGI.ReturnsTot  3.684e-10  1.197e-09   0.308 0.758225    
X.Cont.AGI.Tot    1.380e-02  1.193e-04 115.758  < 2e-16 ***
X.Gini           -4.074e-03  6.897e-04  -5.907 3.50e-09 ***
X.Cont.AGI.25     2.701e-02  4.422e-04  61.099  < 2e-16 ***
X.Cont.AGI.50     3.743e-02  6.237e-04  60.007  < 2e-16 ***
X.Cont.AGI.75     4.414e-02  7.813e-04  56.492  < 2e-16 ***
X.Cont.AGI.100    4.927e-02  8.668e-04  56.837  < 2e-16 ***
X.Cont.AGI.200    5.393e-02  1.031e-03  52.301  < 2e-16 ***
X.Cont.AGI.Rich   6.556e-02  1.355e-03  48.365  < 2e-16 ***
X.StandTot        1.982e-02  7.496e-04  26.442  < 2e-16 ***
X.BracketTot     -1.266e-04  7.330e-05  -1.728 0.084020 .  
X.FragIn         -1.547e-03  3.219e-04  -4.806 1.55e-06 ***
X.Age            -1.075e-04  1.931e-05  -5.566 2.62e-08 ***
X.Hispanics      -6.113e-03  1.026e-03  -5.955 2.61e-09 ***
X.Blacks         -1.891e-03  1.008e-03  -1.876 0.060614 .  
X.Asians         -8.500e-03  1.170e-03  -7.265 3.79e-13 ***
X.Whites         -6.116e-03  1.008e-03  -6.065 1.33e-09 ***
X.Mixes          -2.144e-02  3.349e-03  -6.401 1.56e-10 ***
X.Census.Pop      9.730e-09  1.064e-08   0.914 0.360518    
X.Age.16.up      -1.358e-04  1.682e-05  -8.076 6.82e-16 ***
X.Age.21.up      -2.843e-05  1.661e-05  -1.712 0.086909 .  
X.Age.62.up       1.749e-04  1.321e-05  13.242  < 2e-16 ***
X.Family         -2.284e-06  1.143e-05  -0.200 0.841670    
X.House.Pop       7.877e-04  3.635e-04   2.167 0.030256 *  
X.House.Owned    -1.942e-05  5.701e-06  -3.407 0.000658 ***
X.Highschool     -5.751e-05  1.114e-05  -5.162 2.46e-07 ***
```



```
X.Bachelors       1.121e-05  7.955e-06   1.410 0.158680    
X.Graduate        5.508e-05  7.571e-06   7.276 3.49e-13 ***
X.No.English      8.164e-06  6.324e-06   1.291 0.196727    
X.House.Rooms     3.483e-04  7.740e-05   4.500 6.80e-06 ***
X.No.Cars        -1.948e-05  6.135e-06  -3.176 0.001496 ** 
X.No.Heat         3.673e-05  1.025e-05   3.583 0.000341 ***
X.Rent            9.969e-07  1.922e-07   5.186 2.16e-07 ***
X.House.Price     1.353e-09  4.319e-10   3.132 0.001738 ** 
X.No.Mortgage     4.229e-05  4.692e-06   9.013  < 2e-16 ***
X.Grapi.15        7.404e-06  3.636e-06   2.036 0.041738 *  
X.Grapi.35       -9.838e-06  2.718e-06  -3.619 0.000296 ***
X.Smocapi.M.20    6.810e-06  4.956e-06   1.374 0.169454    
X.Smocapi.M.35    4.121e-06  5.274e-06   0.781 0.434645    
X.Smocapi.NM.10   1.670e-05  3.348e-06   4.987 6.15e-07 ***
X.Smocapi.NM.35  -3.669e-06  5.266e-06  -0.697 0.485960    
X.House.2000     -1.878e-05  1.149e-05  -1.634 0.102186    
X.House.1990     -5.235e-06  8.313e-06  -0.630 0.528877    
X.House.1980     -1.886e-05  8.543e-06  -2.208 0.027246 *  
X.House.1970     -1.366e-05  8.388e-06  -1.628 0.103429    
X.House.1960     -2.182e-05  9.333e-06  -2.338 0.019404 *  
X.House.1950     -2.154e-05  8.953e-06  -2.406 0.016152 *  
X.House.1940     -1.075e-05  1.165e-05  -0.922 0.356321    
X.House.Old      -2.520e-05  7.748e-06  -3.253 0.001144 ** 
X.SavingsTot     -3.348e-08  1.943e-09 -17.226  < 2e-16 ***
```