Shihan Qian

ECON 178

August 31, 2022

Prediction Model of the Total Wealth

In this study, I seek to find a model to predict the total wealth based on different feature variables about household information. The data I use comes from the 1991 Survey of Income and Program Participation (SIPP). It’s representative data of the whole United States Household. The sample size I use for training the data is 7933 interviewed households. The sample contains household data in which the reference persons are aged 25-64 years old. The observation units correspond to the household reference persons. The observations in our samples are persons ranging from 25 to 65 years old. There is a condition that no person is self-employed and at least one person in the household is employed. The units of the observations in the data are referred to as the household interview persons.

To have a brief idea of the study, I first introduce the dataset's dependent variable and feature variables. The dependent variable is total wealth (tw). Total wealth means the total market value of all the assets owned by a household. I give the calculation definition as follows: “Total wealth equals net financial assets, including Individual Retirement Account (IRA) and 401(k) assets, plus housing equity plus the value of business, property, and motor vehicles”. Feature variables are individual retirement account (ira), e401, non-401k financial assets (nifa), income (inc), home mortgage (hmort) , home value (hval), home value minus home mortgage (hequity), educ: education (educ), male (1 if male, 0 otherwise), twoearn (1 if two earners in the household, 0 otherwise), education dummies: no highschool, high-school, some college, college, age, family size (fsize), and marital status (marr, 1 if married, 0 otherwise). There are a few terms that need to be clarified: individual retirement account (ira), e401. e401 plans are tax-deferred retirement savings accounts and it’s a company-sponsored account. They are offered by employers and only employees in companies offering such plans can participate while ira is an individual-created savings account.

The first thing before conducting any analysis, I need to check for perfect collinearity and high multicollinearity problem. Therefore, I verify the multicollinearity of the variables using the correlation check and Variance Information Factor for each independent variable (Fig 1) in the model since high multicollinearity will inflate the standard error and further weaken the significance of the variable. Generally, Variance Information Factor above 5 is an indication of high multicollinearity. In order to reduce high multicollinearity, I will first try to simply remove them from the model. hequity and hmort and hval are perfect collinearity because hequity = hval-hmort, so I decide to drop hequity. Also, I find that nohs+hs+smcol+col = 1; it’s perfect collinearity, therefore, I decide to drop the col dummy. In addition, by the VIF command, I get the VIF value for each feature variable and find out that the VIF for educ is 6 .936390, VIF for nohs is 8.221861, VIF for hs is 6.542482. These are high values above 5, so I

drop hs, nohs,educ in the dataset and name the new dataset df\_1.

I run four basic linear regressions on the data based on the family situation, which includes the variables age, marital status, family size, twoearn, and house value; the retirement information, which includes ira , e401, and age; the variables that have a correlation score with total wealth about 0.5 or above, which includes ira, nifa, hval, inc; and all the variables. From each quantile-quantile plot, I can see a lot of outliers in these regressions given the huge deviation of dots (Fig 2) since a perfect fit model, the QQ plot will be a roughly straight line. Thus, several regression analysis methods need to be applied to improve the accuracy of the prediction of the regression. I use 10 folds cross-validation to compare the mean squared error of four models on the plain dataset: forward stepwise regression, backward stepwise regression, lasso regression, and ridge regression. Lasso and ridge regression perform feature selection and help improve the interpretability of the model and prevent overfitting, while stepwise regression is more helpful in feature selection. By comparing the results of cross-validation, the forwards stepwise give the loIst mean, which is 1732037235. I also perform the 10 folds cross-validation with the dataset that doesn't drop the high VIF variable. The result of the lowest mean is 1733283481, which is a little bit higher than the model without high VIF. Therefore, we would prefer the first dataset that dropped the high VIF variables.

Nevertheless, there are still some improvements that could be added to the model. When doing the data exploration, I notice that in the graphs, the relationships between total wealth and some variables have some curvature shape or flat relationship. Therefore, it will be useful to do some data transformation before applying the analysis method. We can see that home mortgage, home value, and age have a flat relationship shown in the scatter plot (Fig 3). Therefore, I decide to apply splines to these variables. A spline is a piecewise cubic polynomial that helps smoothly fit a non-linear model. Because the distributions of these three variables versus total wealth are pretty flat, therefore I select the degree of freedom ranging from 5 to 20 and use for loop to record the mean square error of each degree of freedom. Inside the for loop, I use 9/10 of the dataset as the training set and 1/10 of the dataset as the testing set for comparing the mean square error. First, I run the spline regression on the home mortgage and record the mse in a vector. The minimum mean squared error is 2037519176 with a degree of freedom of 20. The mse is quite higher compared to the previous method selection because, in this part, I don’t use cross-validation and only select 1/10 for testing, which may increase the mean squared error. Then, I continue using the for loop to select the degree of freedom ranging from 5 to 20 and run the regression on the home value. The spline regression of home value is added to the dataset with 20 degrees of freedom b-spline home mortgage variable. I compare the mean squared error and select the degrees of freedom with the minimum mean squared error. I get the minimum mean squared error is 2035821365 with a degree of freedom of 17. It’s obvious that I can see that the new regression with splines of home mortgage has a lower mean square error. After selecting the optimal degree of freedom for a home mortgage, we continue the selection of the degree of freedom of age. Following the previous routine, I add the spline transformation of the age to the model. I get the minimum mean squared error is 2027919542 with the degrees of freedom of 19. Therefore, I finish the optimal selection of the three variables. Combining all these spline transformations of these variables, I use the forward stepwise selection to do the feature selection of the transformed data. The R squared of the model is pretty, about 0.8547.

Furthermore, I try to apply the second transformation technique, the interaction of variables. I decide to use the interaction between Individual Retirement Account(ira) and age because ira is a retirement feature and it will vary very differently at different ages since elder people are more likely to pay attention to the retirement feature. Thus, I try to include the interaction of age and ira. After applying the interaction of ira and age to the plain dataset, the r squared is 0.853567. It’s less than the model with a spline transformation dataset. Therefore, I plan to add the interaction to the previous spline transformation model and get the R squared of 0.8549. The next interaction I want to include is the interaction between male and marital status because existing scholarly papers and observations have proved that “marriage is positively correlated with wealth” (Vespa&Painter, 2011). In addition, the married men will have even higher average income given the empirical results (Hersch&Stratton, 2000). Since there are different effects for each subgroup of the variable, thus, it’s suitable to include the interaction between male and married status. The R squared of the model included this interaction improve to 0.855.

Then, I decide to test for the third transformation technique, the polynomial transformation. In the first testing model, I only include variables that have a correlation with total wealth around 0.5 or above, which is ira, nifa, inc, hval, and do the polynomial transformation of these variables to the degrees of 3. Then I use a forward stepwise and backward stepwise and get the R squared result of both 0.7998. Since the R square is smaller than our previous model, so I need to discard the polynomial transformation of the whole variables. Therefore, I decide to investigate one particular variable. From the graph of family size (Fig 4), we can see that there is actually a curve trend at a family size of two, so linear regression may not be the best fit. Therefore, we decide to do some polynomial transformations on family size. I use 3-fold cross-validation because I already have a lot of transformed variables included in our model and it already takes a few minutes to run the loops, so for simplicity and efficiency purposes, I have to limit the number of loops and the polynomial degree. After running the loops of comparing the polynomial degrees of 2 and 5, I get the same mean squared error, therefore, we check the model. From the summary of the model, we can see that it doesn't include any polynomial transformation variables of fsize by using forward step-wise selection, thus we know that it may not be a wise choice to include the polynomial transformation of variable fsize.

The fourth technique I use is the generalized additive model. I check for the graph of nifa and inc (Fig 5&6) and find that they all have data concentrated in one part, therefore, I decide to use the ns() to generate Natural Cubic Splines of the three variables and include all the other plain variables in the regression. I also use for loops to select minimum in data mse to choose the best degree of freedom for each variable. The degree of freedom for age is 10, for nifa is 5, and for inc is 5. The R squared is 0.855. Then I try to add the interaction term of male and martial status to the model and get the R squared of 0.856. Though we see the improvements in the R squared, it doesn’t indicate a actual increase in the accuracy of our prediction model. Thus, we need to validate these models I build.

Lastly, I use 5 folds cross-validation to select the best performance model from the previous 4 techniques I mentioned: the first transformation model only includes using splines of three variable: age, hval, and hmort; the second transformation model include splines and interactions: age\*ira, male\*marr, the polynomial model that includes polynomial degrees of 2 of the variable family size; the generalized additive model. After running the cross-validation, I get the minimum mean squared error of 1646350358 by the generalized additive model. Therefore, I decide to use this model for our final prediction. Because I add two interaction terms, I need to first do some data manipulation of the test data. Finally, I use the command write.table(my\_predictions, file = 'my\_predictions.txt') to save the prediction in to txt file for the later prediction accuray check.

In conclusion, the summary of my final model is shown as following:

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -6.029e+02 4.476e+03 -0.135 0.892872

ira 4.476e-01 2.901e-01 1.543 0.122854

ns(age, 10)1 1.806e+03 3.563e+03 0.507 0.612124

ns(age, 10)2 -3.212e+03 4.348e+03 -0.739 0.460055

ns(age, 10)3 -2.457e+03 4.010e+03 -0.613 0.540130

ns(age, 10)4 1.574e+03 4.204e+03 0.374 0.708162

ns(age, 10)5 -5.795e+03 3.908e+03 -1.483 0.138185

ns(age, 10)6 5.994e+03 4.053e+03 1.479 0.139226

ns(age, 10)7 6.988e+03 4.245e+03 1.646 0.099736 .

ns(age, 10)8 -9.022e+02 3.819e+03 -0.236 0.813277

ns(age, 10)9 2.147e+03 6.643e+03 0.323 0.746605

ns(age, 10)10 1.239e+04 3.586e+03 3.457 0.000550 \*\*\*

e401 8.747e+03 1.025e+03 8.535 < 2e-16 \*\*\*

ns(nifa, 5)1 1.912e+03 1.895e+03 1.009 0.312980

ns(nifa, 5)2 8.733e+03 1.761e+03 4.960 7.21e-07 \*\*\*

ns(nifa, 5)3 5.753e+05 1.102e+04 52.212 < 2e-16 \*\*\*

ns(nifa, 5)4 9.738e+05 1.052e+04 92.536 < 2e-16 \*\*\*

ns(nifa, 5)5 1.375e+06 2.083e+04 66.006 < 2e-16 \*\*\*

ns(inc, 5)1 4.167e+02 3.654e+03 0.114 0.909205

ns(inc, 5)2 3.645e+02 4.321e+03 0.084 0.932775

ns(inc, 5)3 1.771e+04 5.647e+03 3.136 0.001721 \*\*

ns(inc, 5)4 6.773e+04 1.180e+04 5.739 9.88e-09 \*\*\*

ns(inc, 5)5 1.201e+05 1.773e+04 6.771 1.37e-11 \*\*\*

hmort -1.021e+00 1.854e-02 -55.059 < 2e-16 \*\*\*

hval 1.070e+00 1.067e-02 100.227 < 2e-16 \*\*\*

twoearn -5.623e+03 1.160e+03 -4.847 1.27e-06 \*\*\*

smcol 1.451e+03 1.090e+03 1.331 0.183296

fsize 2.139e+02 3.346e+02 0.639 0.522654

interaction\_1 8.514e+03 2.090e+03 4.073 4.69e-05 \*\*\*

interaction 2.122e-02 5.712e-03 3.714 0.000205 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 41150 on 7903 degrees of freedom

Multiple R-squared: 0.856, Adjusted R-squared: 0.8555

F-statistic: 1620 on 29 and 7903 DF, p-value: < 2.2e-16

My model include 10 degrees of freedom of the natural spline of age, and 5 degrees of freedom of the natural spline of nifa and income. In addition, I add the two interactions, one between male and martial status and the other between age and ira. The interaction term in the model is kind of important because ​​I notice that the effect of martial will be different when intereact with male condition and the same as the age and ira. Also, we can see the usefulness of using spline by the plots of transformed variables(Fig 7&8&9) compared to the simple linear regression plot (Fig 3). That the transformed spline lines are more fitted to the scatter plots.

There is some possible caveats and improvements I can do for this project. If the time and computer CPU is enough, we can do a grid search for different combination of variables and model choices,which will give us better results. Also, for the simplicity and convenience purposes, some parameter selection, I only calculate the in sample mean square error, rather than the out of sample mean square error. It may cause some deviation in the final prediction model. In addition, for the final model because I don’t add a feature selection step to the genralized additive model, so if I have a chance to redo the project, I will continue adding on the spline to the model and use more restricted feature selection method to choose from these variables.

From the process of doing the project, I discover that the forward and backward stepwise method is very useful in feature selection, but it may sometimes cause overfiting compared to lasso and ridge regression since lasso and ridge has constraints and regularization to prevent overfitting, so we need to be cautious with choosing the method. In addition, I notice that it’s very important to check the data distribution first and then do the analysis, for example, I see that there is a flat relationship between hmort and total wealth, then I know it could be useful to use spline on the variable. Importantly, cross validation is a tool that helps us validate and select our model and it’s not limited only on comparing the model, but also selecting the parameters inside one model.

Appendix

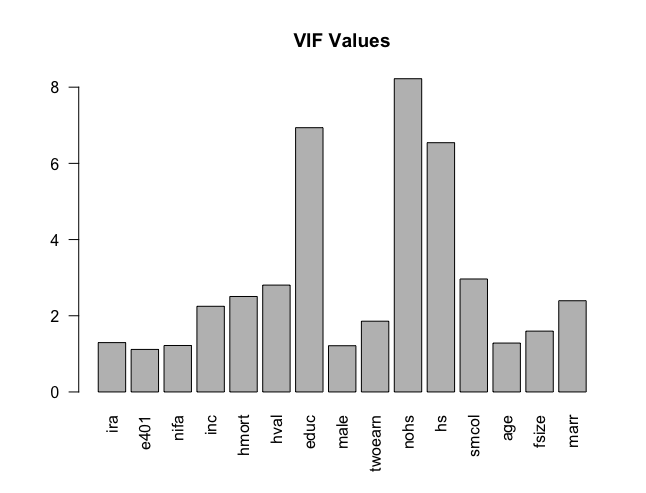
Fig 1

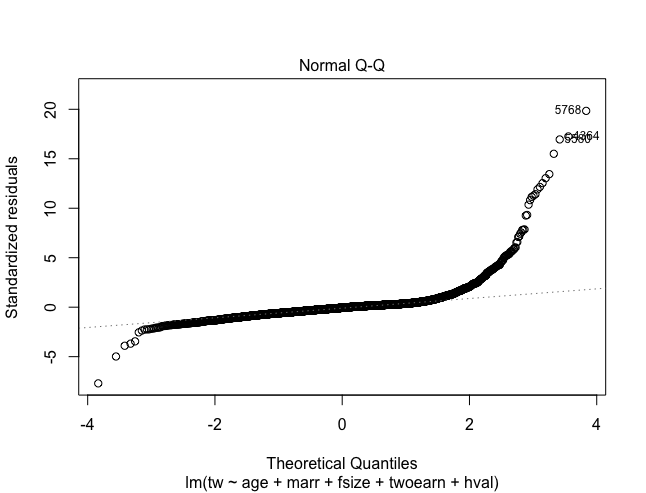
Fig 2

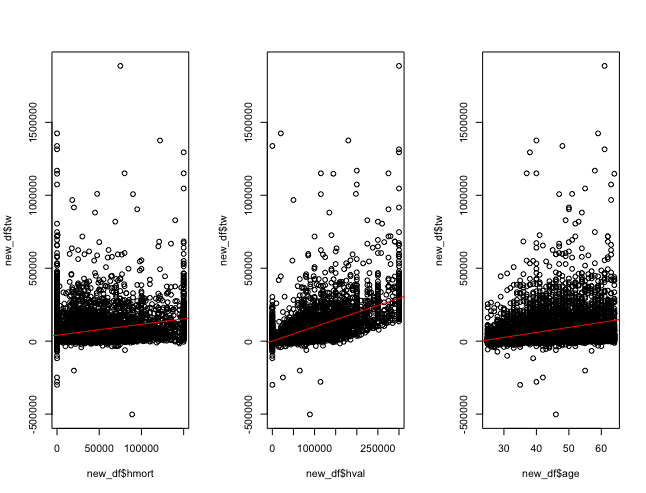
Fig 3

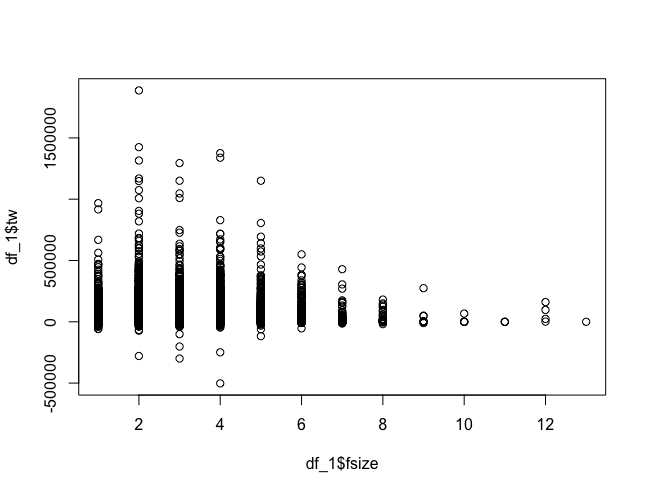
Fig 4

Fig 5

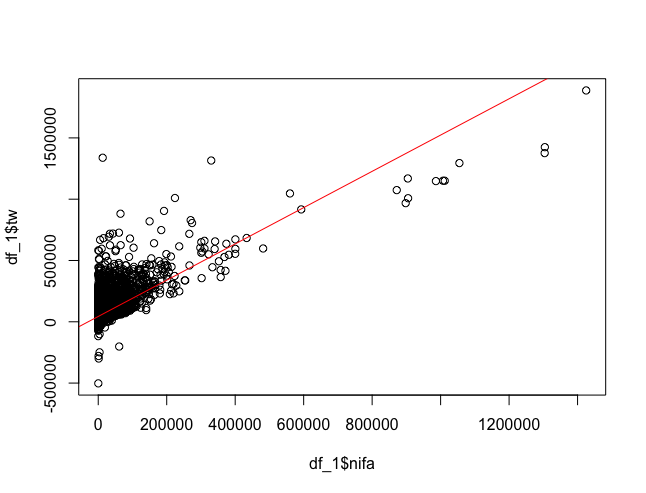


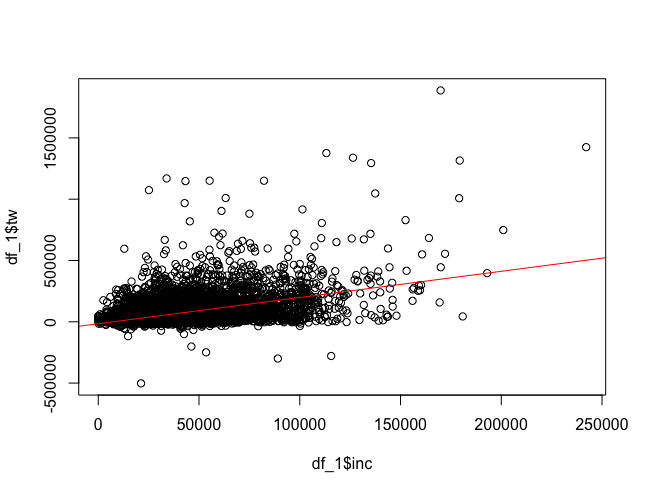
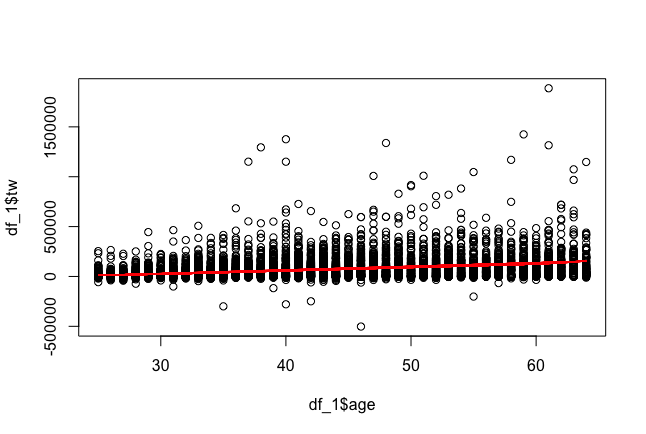
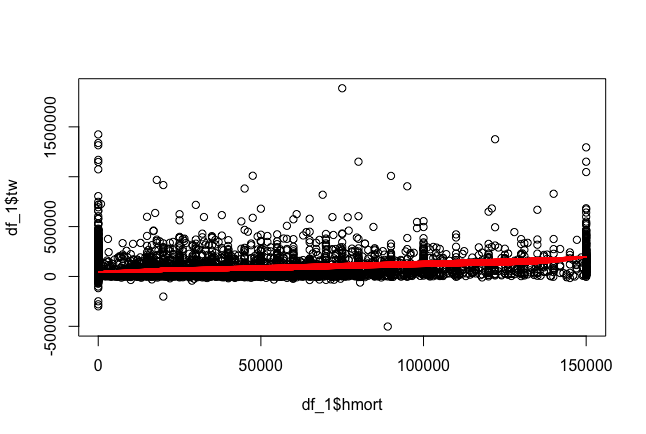
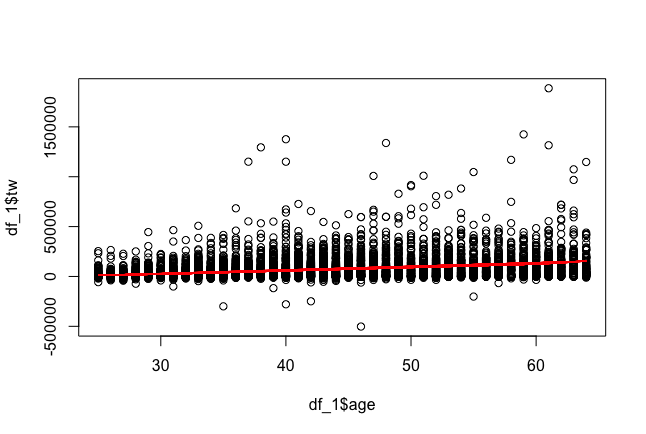
Fig 6

Fig 7&8&9







Reference

Jonathan Vespa, Matthew A. Painter; Cohabitation History, Marriage, and Wealth Accumulation. *Demography* 1 August 2011; 48 (3): 983–1004. doi: <https://doi.org/10.1007/s13524-011-0043-2>

Hersch, J., & Stratton, L. S. (2000). Household Specialization and the Male Marriage Wage Premium. *ILR Review*, *54*(1), 78–94. <https://doi.org/10.1177/001979390005400105>