Final Project

1. Introduction

This project aims to predict the total wealth of American households by using data from the 1991 Survey of Income and Program Participation. In order to achieve this scope, different models, such as Lasso and Ridge regressions, are employed for prediction. In-sample and out-of-sample evaluations are used to select the best performing model. For the first instance, Mean Square Error is compared from the results of the different models. For the second one, a 10-fold cross-validation is performed, and the Mean Squared Prediction Error is used to compare the models: the regression that obtains the lowest MSPE is picked as best model since it is the best one in generalizing on unseen data.

2. Variable Selection

The data for this project is collected by the 1991 Survey of Income and Program Particitaption, consisiting of 7933 observations. To predict the total wealth, households with at least an employed person between 25 and 64 years old are taken into account. The dataset presents 18 variables, among which the total wealth is going to be the dependent variable and some of them will be employed for the prediction.

2.1 Dependent Variable

The dependent variable predicted in this project is the total wealth of American households.

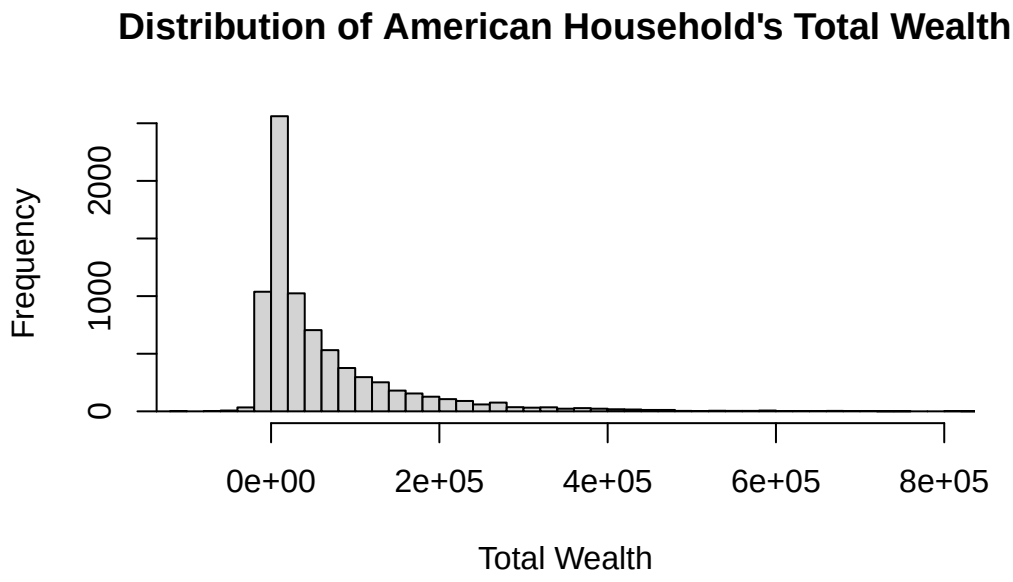


Figure 1. Distribution of total Wealth of American Households.

Figure 1 shows how the dependent variable distributes: a right skewed distribution, with more people possessing a lower total wealth. This right skeweness is possible to see also trhough a diagnostic of the data. The mean of this distribution accounts for $63,628.48 while the median for $25,225. This shows how some outliers with a higher total wealth influence the average of the variable as the median is significantly lower thant the first parameter.
In the visualization, this concept is represented by a peak in the lower range, meaning that the majority of the observations concentrates in there. As wealth increases, the observations decrease, showing that less households own high wealth. It is also possible to notice the outliers with higher wealth, which stand in the higher section.
In short, total wealth is right skewed, with the majority of people having a total wealth concentrated in the lower range, some cases in the intermediate, and some outliers in the highest range.

2.2 Independent Variables

The dataset presents 17 independent variables that can be employed as predictors; they divide in four main groups: retirement related variables, financial related variables, home ownership variables, and demographic variables. Among them, a selection will be done to choose the right predictors and avoid problems like multicollinearity and rsik of overfitting. For what concerns the variables related to retairement, there are the ira and the e401 variables. The first one inclueds the amount of dollars in the IRA retairment plan, which is delieberalely chosen by the individual. The second one, the e401, regards the retairment plan for the indivduals, but it has more requirements. Indeed, not all companies sponsor this plan, making the variable not a wealth based choice of the individual, but a decision taken from the company, which is not linked to the willingness

of the person. In other words, e401 is selected following sandards outside the indivudal's wealth. This means that it provides data just for restricted observations, and the presence or absence of an account is not related to the wealth but to a decision of the company. A further analysis of this variable shows the presence of zeros for almost all the observations. For these reasons, the variable will not be included in the models. Regarding the financial variables, there are nifa and income. Nifa inlcues all the financial asstes that are not 401k, and it is expressed in dollars, as well as income which is reported in the same currency. Both of these variables are included in the regression models. Another kind of variable present in the dataset is the financial variables. The home mortgage is the amount of dollars that the household holds in mortgage, the home value is the value in dollar of the house for each household while hequity is the subtraction of the two. Since it is of primary importance to include these values in the models, they will employ these values as predictors. However, including the three variables will lead to the problem of multicollinearity since they are related to each other, resulting in issues for the regressions. For this purpose, the following models will include just the hequity variable, which best represents the net worth of the previous two. The demographic variables relate to education, age, family size, marriage, earinings, and gender. For what concerns education, the varibale that it wil be used includes the years of education for each individuals. There are also dummies variables related to education, such as the attendance or not attendance of high-school and college. However, all of these variables are related to each others, and their inclusion might lead to multicollinearity. For this reason, only the years of education will be included as parameter. The age of the individuals will be included as predictor as well. Moreover, the earining varibales is another dummy variable which expresses if the households includes the presence of two earnings or not. This variable is included as predictor, but there were concerns that it would be related to family size and marriage. For this purpose, some tests to check multicollinearity among these variables are run. Among these, an Ordinary Least Square is run: if there is multicollinearity among these predictors, they will be automatically dropped in the regression.

Table 1. Ordinary Least Square
Without Gender Variable

(Intercept)	-14884.282 ***
	(3653.308)
ira	1.606 ***
	(0.055)
nifa	1.052 ***
	(0.010)
inc	0.387 ***
	(0.026)
hequity	1.086 ***
	(0.011)
twoearn	-7105.684 ***

	(1307.318)
age	270.553 ***
	(50.426)
educ	-56.059
	(190.949)
marr	894.846
	(1457.662)
fsize	-229.178
	(376.613)
N	7933
R2	0.852

*** p < 0.001; ** p < 0.01; * p < 0.05.

Column names: names, Table 1. Ordinary Least Square Without Gender Variable

Table 1. Ordinary Least Square employing all the variables discussed (without gender predictor).

As it is possible to see in Table 1, all the predictors discussed are kept for the regression, showing that there is no presence of multicollinearity. The last independent variable to discuss belongs to demographic variable: gender. In order to understand if gender is a good predictor, another OLS is run: it has all the predictors as the one in Table 1, but it will add the gender variable (male).

Table 2. Ordinary Least Square With Gender Variable

(Intercept)	-17166.424 ***
	(3772.278)
ira	1.608 ***
	(0.055)
nifa	1.052 ***
	(0.010)
inc	0.382 ***
	(0.026)
hequity	1.086 ***
	(0.011)
twoearn	-6977.384 ***
	(1307.996)

age	286.448 ***
	(50.838)
educ	-32.636
	(191.137)
marr	1529.449
	(1480.683)
fsize	-68.845
	(382.298)
male	3093.385 *
	(1279.880)
N	7933
R2	0.852

*** p < 0.001; ** p < 0.01; * p < 0.05.

Column names: names, Table 2. Ordinary Least Square With Gender Variable

Table 2. Ordinary Least Square including the male variable.

As it is shown in Table 2, the variable male, which accounts for the gender of the individuals, is significant (p-value = 0.0157) meaning that it has a linear relation with total wealth. Moreover, the Residual Standard Error slightly decreases, signifying a little improvement of the overall model. The same can be seen from the R-squared. Indeed, even if the difference with Table 1 is minor, the R-squared increases, meaning that the second modle is explaining more variance of the data compared to the previous one. Because of these three reasons, the variable male is included as a prdictor for the following regressions. In summary, the independent variables selected for the following models include retirement, financial, home ownership, and demographic variables, but they take into account different factors that can benefit the models, such as multicollinearity and significance.

2.3 Relation Between Dependent and Independent Variables

In order to understand the models that best fit the data in order to get the best predictions, understanding the relationship between dependent and independent variables is necessary. As a first thing, a correlation matrix will be run to understand the relation of the selected variables, then, some scatterplot will help in visualizing the kind of relationships among them.

2.3.1 Correlation Matrix

Figure 2 represents a correlation matrix that visualizes the relations between the dependent variable and all the independent variables selected for the models. Even if this is the representation of mainly linear relations, and it is important to take into account non linear relations as well, this figure helps in the visual detection of multicollinearity. Some varibles, as family size and marriage, are more correlated than others (a more intense color shows a stronger relation); however, the previous tests reassure a lack of multicollinearity. Something important to notice is also the more intense relation of ira and hequity with the total wealth output. This might mean that a more linear relation exists between each independent variable and the prediction. However, a more detailed analysis is needed. In order to do so, some scatterplots will help to start this diagnostic.

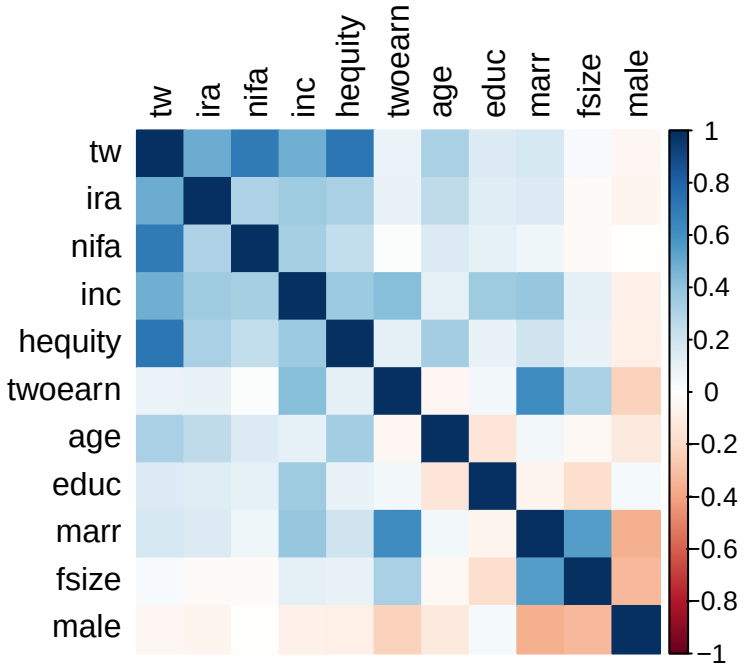


Figure 2. Correlation Matrix of the independent variable and all the dependent variables selected.

2.3.2 Visualization with Scatter Plots

In order to visualize the independent variables, scatterplots are an efficient tools. A series of these representations will follow, focusng first on the financial variables and then on the demographic ones excluding dummy variables.

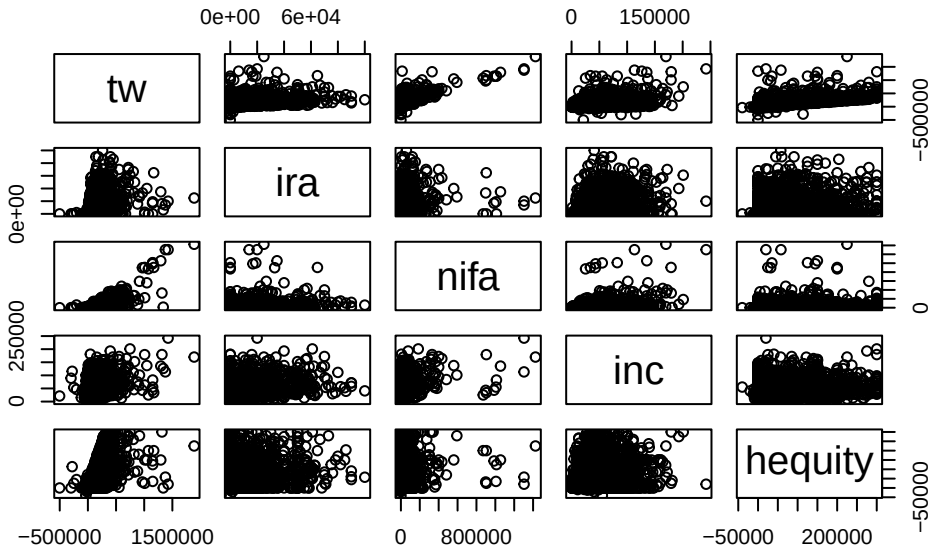


Figure 3. Representation of financial variables in relation to the dependent variable total wealth.

Figure 3 shows the relation of financial varibales to the total wealth. As it is possible to see, the relation of the predictors to the total wealth is positive, but it is not a strong linear relationship for neither of them. Even if nifa and income present the most linear relationships, the scatterplots show that running a linear regression will leave out some relationships due to non linearity.

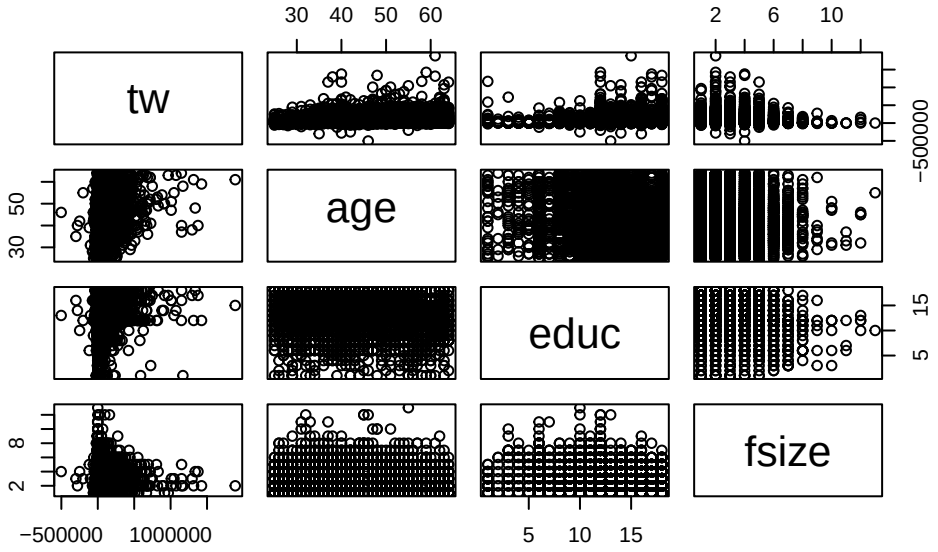


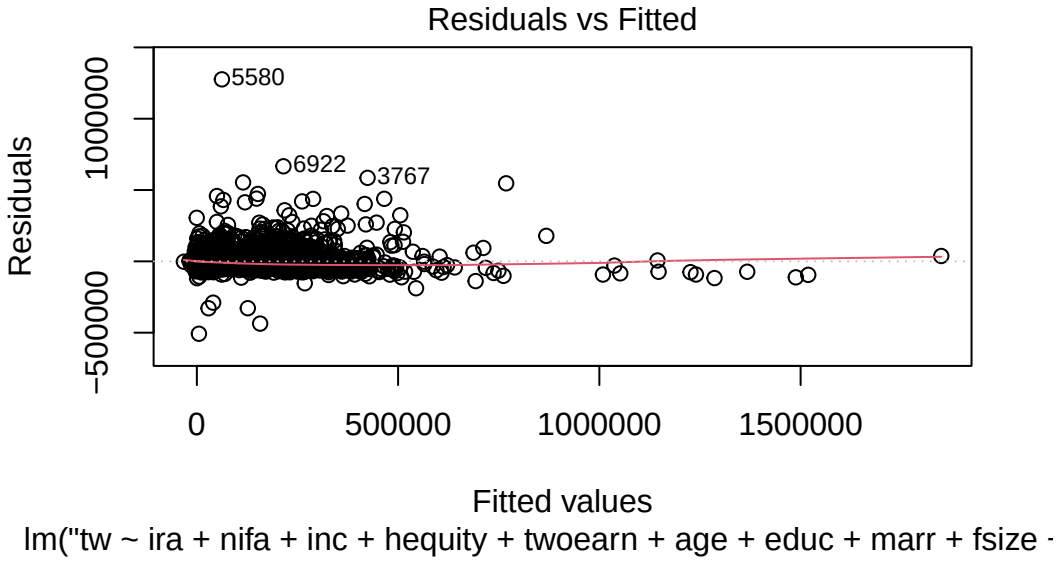
Figure 4. Representation of demograpic variables in relation with the total wealth.

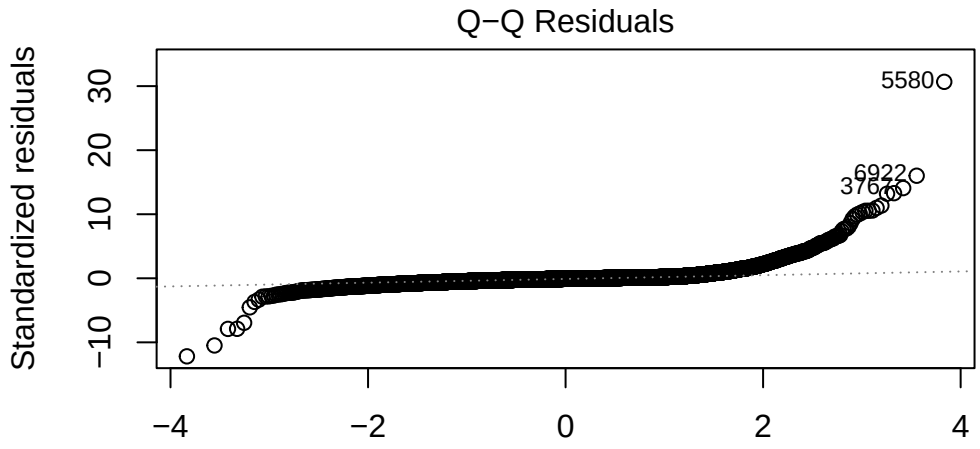
In the case of demographic variables represented in Figure 4, it is interesting to see that total wealth is positively reated to age and education (stronger to education), meaning that an increase in these independent variables leads to an increase of the total wealth. The family size of the household, on the other hand, is negatively related to total wealth; as the number of members in a household increases, total wealth decreases. In this second figure, linear relationships are more evident than in the previous one. In sum, the majority of the independent variables are positively relted to the outcome, but there is the presence of some predictors, like age, which follow the opposite trend. However, the scatterplots shows some relationships that are not prevalently linear. Thus, it is important to take into account non-linearity to create the best model (a more flexible model would be a solution but with the risk of overfitting).

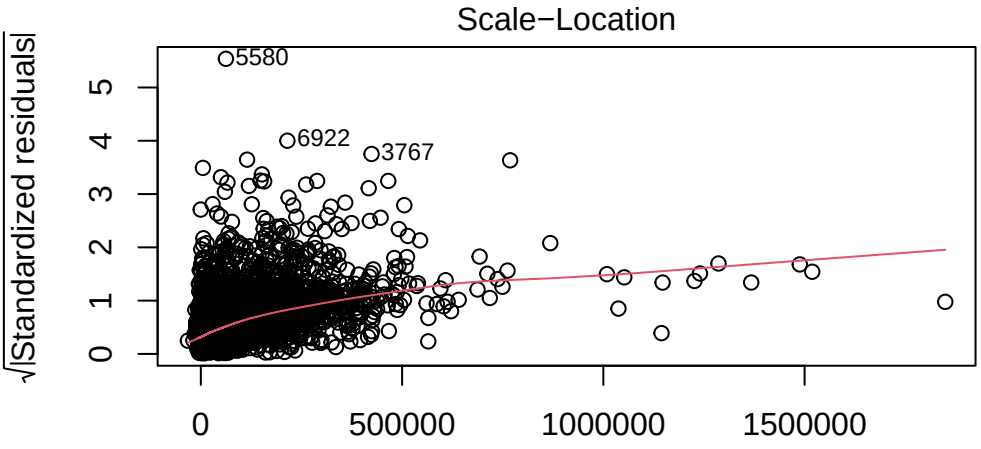
3. Data Analysis

3.1 Linear Regression

Even though the data presents non linear realations, running a linear regression would be useful to understand the dispersion of data and their leverage. Specifically, an analysis of the residual is helpful to understand how to handle outliers based on their leverage or the level of the variance throught the data. For this purpose, the following figures represent the result of a linear regression run using the predictors previously selected.




Q−Q Residuals
Standardized residuals
5580
6922
3767
−10
0
10
20
30
−4
−2
0
2
4
Theoretical Quantiles
lm("tw ~ ira + nifa + inc + hequity + twoearn + age + educ + marr + fsize -


Scale−Location
√|Standardized residuals|
5580
6922
3767
0
1
2
3
4
5
0
500000
1000000
1500000
Fitted values
lm("tw ~ ira + nifa + inc + hequity + twoearn + age + educ + marr + fsize -

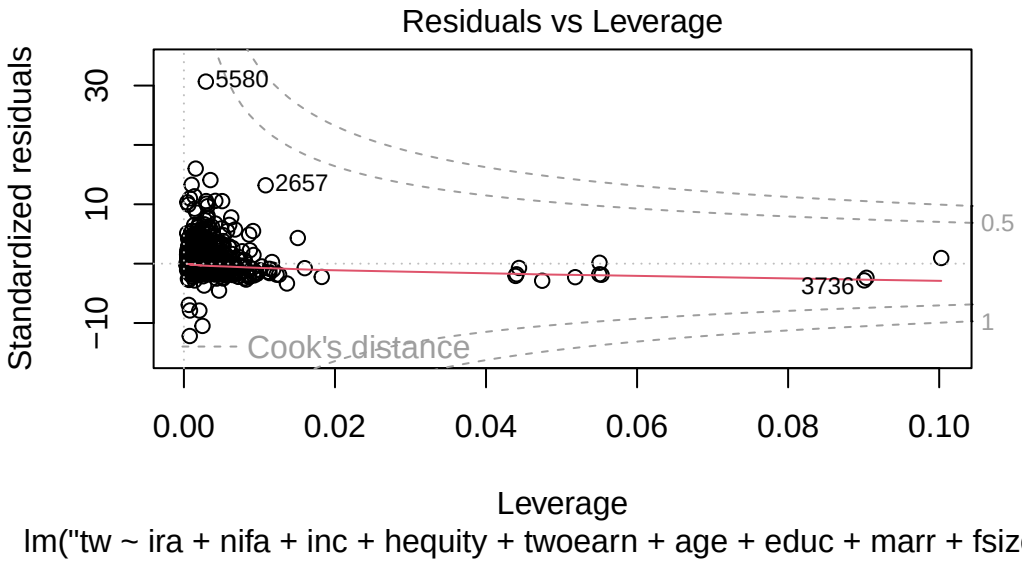


Figure 5. Representation of Residuals after running a Linear Regression.

Figure 5 reports four different plots. The Residual vs Fitted plot shows how much the error variates. In this case, the data is heteroschedastic, meaning that the data do not present the same variance for all the values on the X axis, yet the variance increases when the total wealth is lower. In this plot, three outliers can be spotted. In the following plot, the Q-Q Residuals, the same outliers are in the highest range. Because of them and other observations in the limits, the plot suggests a skewed distribution. Since outliers can skew the disrtribution and mislead the results, understanding the influence that they have on the data is of primamry importance. The Residual vs Leverage plot helps in this visualization. It is important to notice that three outliers are higlighted in the plot; however, they remain in the boundaries depicted in the graph (dotted lines). This means their leverage does not influence the data and the performance of the models (the same models were run with and without outliers and the results did not change, confirming that their leverage is of low impact). For this reason, the outliers are not removed from the dataset as their influence does not affect the results.

4. Model Selection

For the purpose of this paper, different models are run to predict the total wealth of American households. After that, the best performing model is selected. In order to decide the best model, in-sample and out-of-sample evalutaions are considered. The model with the lowest Mean Squared Predicted Error is selected as it would be the best one in generalizing, so predicting on unseen data.

4.1 Ridge Regression

The first model takes into account a Ridge regression. Before running the regression, a cross validation is performed to select the best lamba in a range from -10 to 10. Figure 6 shows how the Mean Squared Predict Error changes with the changing in lambda. As it is possible to see, the MSPE remains stable for the majority of lambdas in this range, but it increases when lambda increases too much, meaning that the restriction is too high and the model is not able to make good predictions anymore.

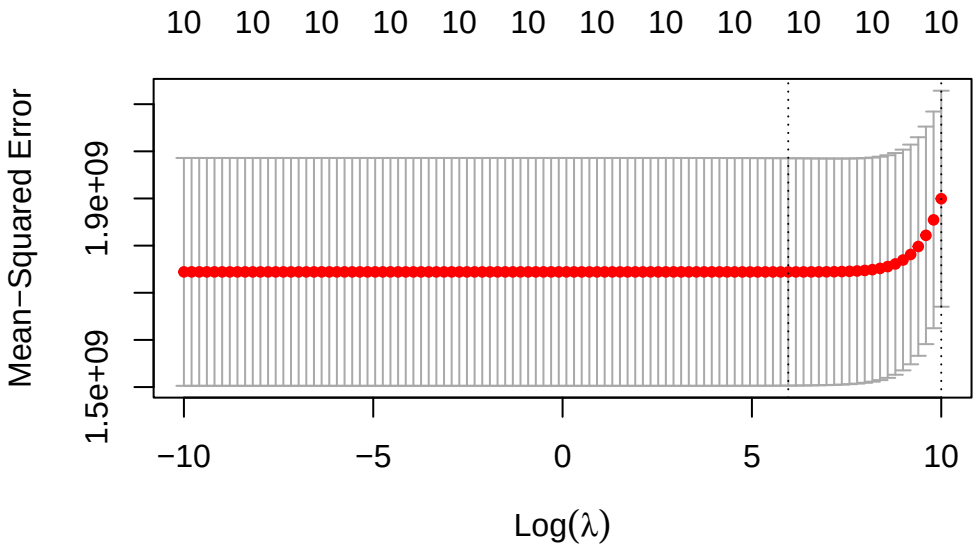


Figure 6. Mean Squared Predicted Error vs Lambda.

The Ridge regression is run with the best lamba obtained from the cross validation. The Mean Squared Error obtained from this model accounts for 1733045432. Even if this value does not have significance by itself in evaluating the model, it will be compared with the MSE of other models to understand the in-sample performances of each.

4.2 Lasso Regression

The same process of the Ridge regression is followed for the Lasso regression, which is the second model considered in this project. The first step is the selection of the lambda thanks to which the model performs best, so has the lowest Mean Squared Error. Figure 7 shows that the level of error is stable for almost all the ranges of lambda, which starts from -10 to 10, but it increases when lambda increases too much, meaning that the restriction is too high for the model to have good performance: the model is underfitting.

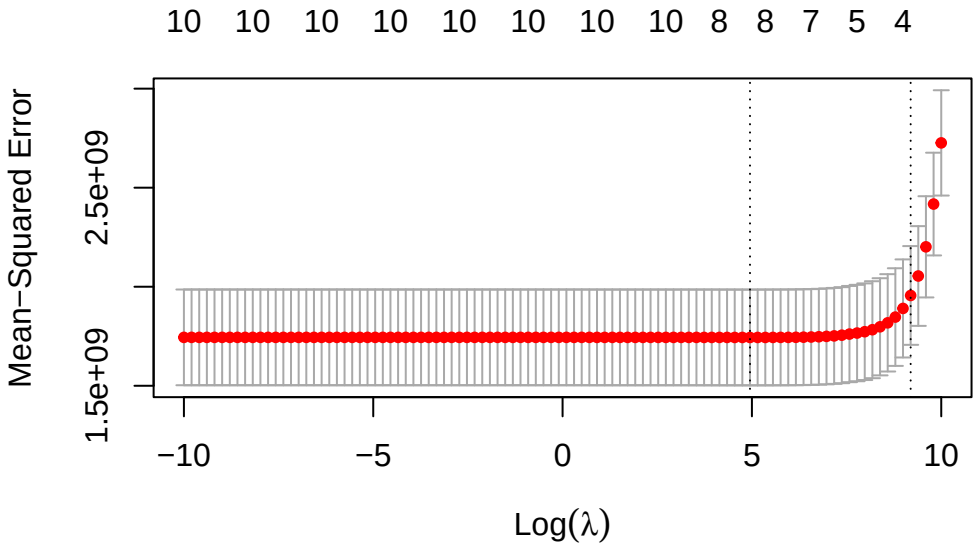


Figure 7. Mean Squared Error vs Lambda.

Even in this case, the Mean Squared Error is calculated, and it accounts for 1733173462, which value will be compared to that of other models to understand the one that is best performing on in-sample evaluation.

4.3 Stepwise Selection

In this paper, a forward stepwise and backward stepwise selections are run. In order to do so, two models are initially created, one uses all the independent variables selected as predictors while another one it has none besides the 1 intercept. These models are then used to add or detract the predictors in order to create a subset and select the model with the best perfromance. If for the stepwise forward selections variables are added, in the backward stepwise selection variables are detracted during the process.

In this case, the Mean Squared Error accounts for 1733248344. Even if these values will be compared to the MSE of the other models, the equivalent values signify that the two processes selected the same model or a really similar one as the best performing one. Therefore, the values obtained from the two different models are the same. This fact can happen when running forwards and backward stepwise selection models on the same data, but it is not always the case.

4.4 Model with Splines

4.4.1 Polynomial Regressions

In the analysis performed with the satterplots, it was evident how the preditors relate to the outcome variable in different ways. If for demographic variables, the visualization (figure 4) quite showed a linear relation, for the financial variables the relation was mainly non-linear. To understand the extent of non-linearity of the demographic variables, polynomial regressions can show if picking non-linear regressions for those selected variables is approprate as they show some curvature that can be grasped by the polynomial. As a consequence, for the demographic variables, a cross validation to select the best polynomial degree is run. If the best degree corresponds to 1, the varibles will be considered as linear predcitors. The first demographic variable for which the curvature is tested is hequity (the difference between mortgage and house value). After running regression models predicting total wealth with hequity as predictor, the graph of Figure 8 represents at which polynomial degree the variable performs best (the degree on the x-axis and the MSPE on the y-axis). As it is evident, one polynomial degree is the optimal choice for this variable, meaning that the relationship is linear. For this reason, a spline for this variable is not used.

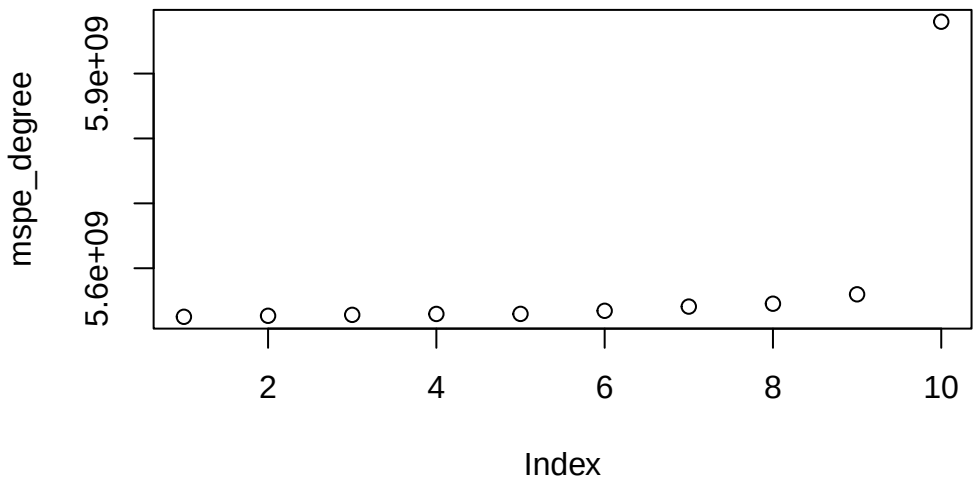


Figure 8. Polynomial Degree for Hequity.

The second variable is nifa. For nifa, the minimum MSPE is obtained with the degree equals to 5. Even if the difference with the 1 degree (so linearity) does not look prominent, at 5 polynomial degree the error is the lowest, so it is important to ackonwledge a non-linear relation between this variable and the outcome. A spline is applied for nifa in order to capture the curvature of this variable.

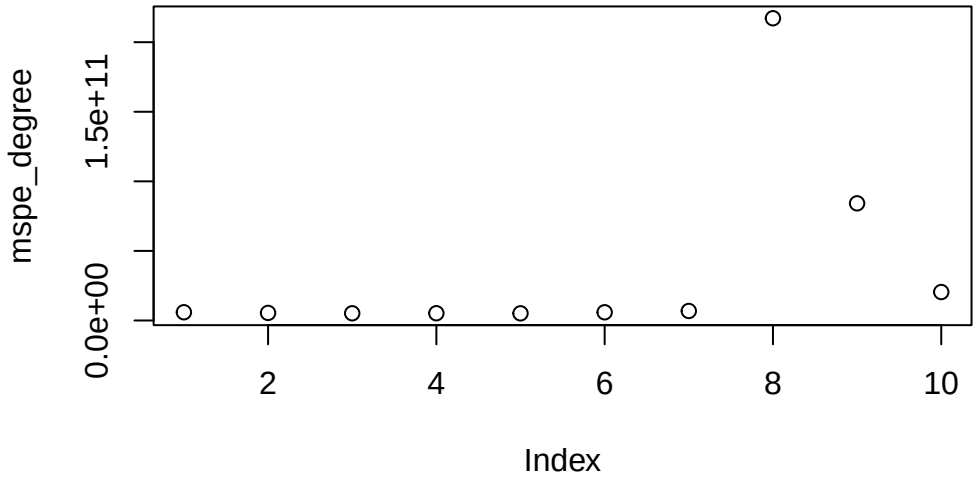


Figure 9. Polynomial Degrees and Mean Squared Predicted Error for NIFA.

Another variable that, from the scatterplots, shows non-linearity is ira. By running regression models with ira predcicting total wealth, the graph (Figure 10) shows that the polynomial degree with the lowest Mean Squared Predicted Error fluctuates; however, the lower residual is obtained with the degree being equal to 4. As it assumes non-linearity, a splines will be included for this variable as well.

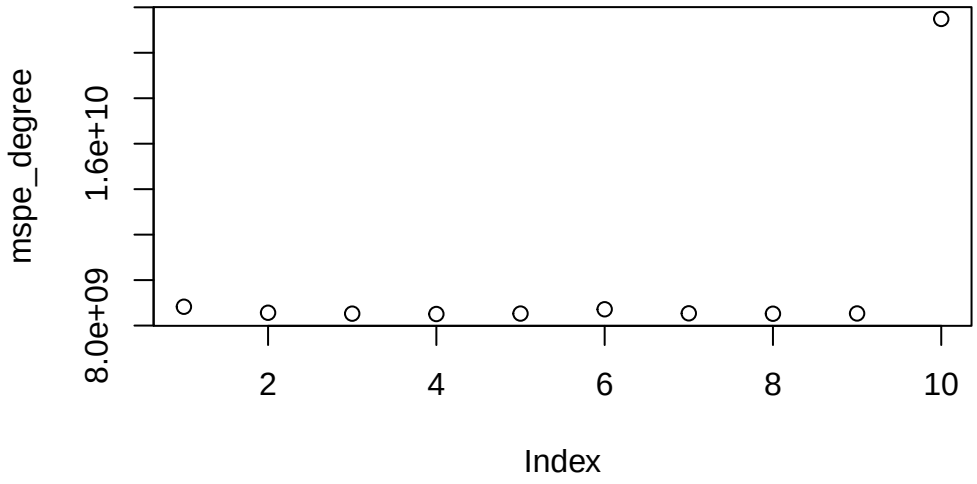


Figure 10. Polynomial Degrees and Mean Squared Predicted Error for IRA.

The last independent variable analysed for its curvature is income. Even if the scatterplots were showing a positive relationship that would be assumed to be linear, checking this factor potentially imporves the model. Indeed, Figure 11 shows that there is not a large difference between the Mean Squared Predicted Errors that have polynomial degrees on a range between one and seven. However, the lowest MSPE is achieved with the polynomial degree at four, implying non lineraity, so a spline is employed for this predictor.

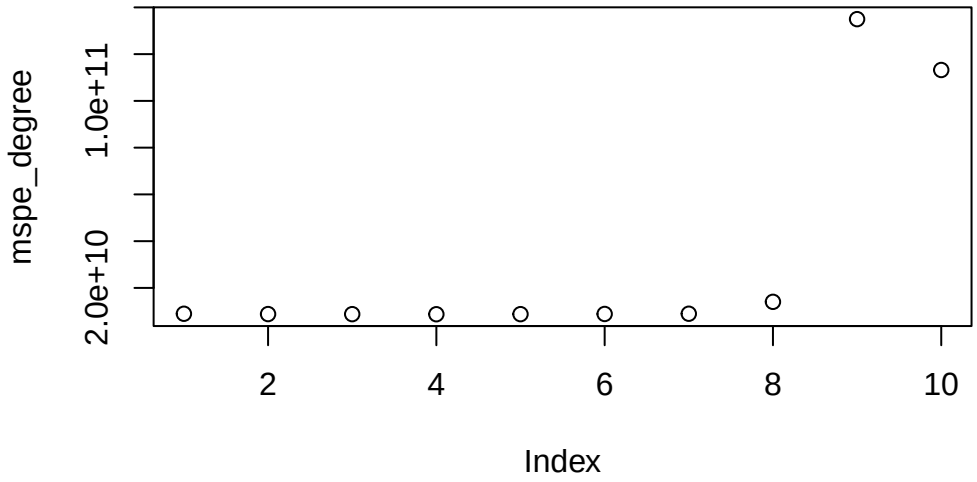


Figure 11. Polynomial Degrees and Mean Squared Predicted Error for Income.

4.4.2 Model with Splines

Running the regressions for the demographic variables made clear that some predictors are not linearly related to the total wealth, but underlying non-linear patterns are present. In this sense, a linear regression with splines is able to consider non-linearity for selected predictors while keeping the others with the assumption of linearity. The following model applies splines for NIFA, IRA, and Income, running a linear regression for the other variables selected also for the previous models. The degrees of freedom are kept low for all the variables the splines are applied to as the polynomial regressions show that these predictors perform better with a low number as ploynomial degrees. Indeed, by increasing the degrees of freedom of the splines brings to an increase in flexibility. As a flexible model would be able to detect different patterns that are not included in a less flexible one, the first can incorprate a lot of noise in the model. This means that the model is performing well in seen data and not in unseen ones, meaning that it is overfitting and not able to generalise since it is relying too much on the data and not on the underlying patterns. The Mean Squared Error (1706430182) is compared to the one obtained from other models to understand the in-sample performance of each.

5. Models Evaluation

5.1 In-sample evalutaion

After running the five different models that include the same variables but employ different approaches, the Mean Squared Error is a good parameter to compare their performances. The model with the lowest MSE is better performing on in-sample evaluation since it produces the lowest residual. As it is possible to see in table 3, the model with splines is the best performing one as it accounts for the lowest residual. Table 4 reports the same residual, but instead of the MSE it shows the Squared Root of the Mean Squared Error; the error in dollars for each model is reported. However, a good performing model needs to be able to generalize on unseen data, so an out-of-sample evaluation is required to pick the best model.

	"Table 3"	mse_lasso	mse_ridge	mse_forward_mod	mse_backward_mod	mse_splines
1	Table 3	1733173462	1733045432	1733248344	1733248344	1706430182

	"Table 4"	mse_lasso	mse_ridge	mse_forward_mod	mse_backward_mod	mse_splines
1	Table 4	41631	41630	41632	41632	41309

5.2 Out-of-Sample Evaluation

A good perfroming model is one that can predict the output with a low residual. In order to evaluate this performance, the model needs to be tested on unseen data; cross-validation is the approach to accomplish this aim. In this paper, a 10-fold cross-validation is performed on each model previously run. In other words, the data is randomly split in 10 different folds of the same size, while 9 of them will be used for training the model, the 10th is employed to test the model. This process is repeated for each fold, so that each of the 10 folds is once used for testing and the other for training. The Mean Squared Predicted Error is used as evaluation parameter; once again, the model with the lowest residual is the best performing one. Running a cross-validation of this kind is essential to understand the real performance of the models; for example, achieving good results on in-sample evaluation but really scarce ones in out-of-sample evaluation might signify that the model is overfitting. Thus, it is not able to get good predictions as it is strctly linked to the data incorporating also the noise; it is not good in generalizing on unseen data, which is the purpose of the model itself. In this case, Table 5 shows the Square Root values of the Mean Squared Predicted Error, so the amount of the residual expressed in dollars. The values reported are really similar among them. Specifically, the forward stepwise selection and the backward stepwise selectionmodels have the same MSPE as the Lasso regression. The model employing the Ridge regression is the one with the lowest performance while the model with splines achieves the lowest residual, meaning it is the best performing one. The increase in residual from the in-sample to the out-of-sample evaluation for the model incluind splines might show that the model is slightly overfitting. This fact might be due to the increased flexibility of some variables, which makes the model incorporate

random noise. Neverthless, the linear regression with splines is the best performing one among all the models as it generates the lowest residual, so it is more generalizable.

	"Table 5"	step back	step forw	lasso	ridge	splines
1	Table 5	1742632335	1742632335	1742667518	1768261731	1731859479

6. Conclusion

The aim of this project was to predict the total wealth of American households using data from the 1991 Survey of Income and Program Participation. The first step was establishing which independent variables to include in the models based on the relations among them, in order to avoid multicollinearity, and to the independent variable, so to understand the presence of underlying patterns differning from linearity. Linear regressions and visualizations helped in this aim. After understanding the presence of non-linearity for some predictors, the model selection started with a Ridge and Lasso regressions, for both of which cross validation was used to pick the best lambda. Other two models included forward stepwise and backward stepwise selection, for which a full and a null model have been created to add or drop variables according to the process. The two processes ended up selecting a really similar if not the same model, which had the same performances both in in-sample and out-of-sample evaluations. Because of some predictors showing linearity while others showing a different relation to the outcome, a linear regression including three splines for variables selected through polynomial regressions was run. This choice was the best performing one for both the evaluations. However, the increase error in the out-of-sample evaluation might signify a slightly overfit happening in this model. Neverthless, this remains the best performing one.

Reference

For this project, I employed slides from lectures and discussion sessions of the ECON 178 course.